Asymptotically Optimal Methods for Sequential Change-Point Detection

Thesis by

Yajun Mei

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2003

(Defended May 28, 2003)

Acknowledgements

I would like to express my deepest gratitude to my advisor, Gary Lorden, for introducing me to the wonderful world of sequential analysis, and for his constant support and guidance during my graduate study and thesis preparation. It has been a great pleasure to work under his supervision.

I would also like to thank Emmanuel Candes, Robert Sherman and David Wales for serving on my thesis committee. Special thanks go to Robert Sherman for his lectures that provided me with statistical knowledge and for being my referee. Many thanks to Moshe Pollak for useful discussions, suggestions and sharing his insightful ideas.

I am grateful to Venugopal V. Veeravalli for bringing the problem of decentralized quickest change detection to my attention and Alexander G. Tartakovsky for fruitful discussions.

I have been fortunate to work in the pleasant atmosphere of the Department of Mathematics at the California Institute of Technology. I would like to thank all the members and staff of the department for their support. Jay Bartroff is especially thanked for his proofreading portions of this thesis.

Last but not least at all, I want to thank my wife, Fang Yan, for her unconditional love, support and patience. I dedicate this thesis to her.

Abstract

This thesis studies sequential change-point detection problems in different contexts. Our main results are as follows:

- We present a new formulation of the problem of detecting a change of the parameter value in a one-parameter exponential family. Asymptotically optimal procedures are obtained.

- We propose a new and useful definition of "asymptotically optimal to first-order" procedures in change-point problems when both the pre-change distribution and the post-change distribution involve unknown parameters. In a general setting, we define such procedures and prove that they are asymptotically optimal.

- We develop asymptotic theory for sequential hypothesis testing and change-point problems in decentralized decision systems and prove the asymptotic optimality of our proposed procedures under certain conditions.

- We show that a published proof that the so-called modified Shiryayev-Roberts procedure is exactly optimal is incorrect. We also clarify the issues involved by both mathematical arguments and a simulation study. The correctness of the theorem remains in doubt.

- We construct a simple counterexample to a conjecture of Pollak that states that certain procedures based on likelihood ratios are asymptotically optimal in change-point problems even for dependent observations.

Contents

Chapter 1

Introduction

Sequential *change-point* detection problems have many important applications, including industrial quality control, reliability, fault detection, clinical trials, finance, signal detection, surveillance and security systems. Extensive research has been done in this field during the last few decades. For recent reviews, we refer readers to Basseville and Nikiforov [1] and Lai [10], and the references therein.

In change-point problems, one observes a sequence of independent observations $X_1, X_2, \ldots$ from some process. Initially, the process is "in control," i.e., the X's have some distribution f. At some unknown time ν, the process may go "out of control" and the X's have another distribution g. The problem is to detect the occurrence of the change as soon as possible while keeping false alarms as infrequent as possible.

When both the pre-change distribution f and the post-change distribution g are completely specified, the problem is well understood and has been solved under a variety of criteria. Some popular procedures are Shewhart's "control chart," Moving Average control charts, Page's CUSUM procedure, and the Shiryayev-Roberts procedure. The first asymptotic theory, using a minimax approach, was provided in Lorden [16].

In practice, the assumption of known pre-change distribution f and post-change distribution g is too restrictive. Motivated by applications in statistical quality control, the standard formulation of a more flexible model assumes that the pre-change distribution f is given and the post-change distribution g involves unknown parameters.

However, as shown by several examples in Chapter 3, there are many situations in practice in which the pre-change distribution intrinsically involves unknown parameters. Chapter 3 proposes a different formulation for such change-point problems and provides asymptotically optimal procedures.

As an example, consider the problem of detecting shifts in the mean μ of independent normal observations with known variance. Suppose $\mu_0 < \mu_1$. The standard formulation is to specify a required frequency of false alarms when $\mu = \mu_0$ before a change occurs, and to minimize detection delay if μ changes to some value larger than μ_1, see, e.g., Lai [10], Lorden [16], Pollak [25], Siegmund and Venkatraman [34]. In Chapter 3, we assume that μ is partially specified before the change, say $\mu \leq \mu_0$. Our formulation specifies a required detection delay if μ changes to μ_1 and seeks to minimize the frequency of false alarms for all possible μ less than μ_0.

It is natural to combine the standard formulation with our formulation by considering change-point problems when both the pre-change distribution and the post-change distribution involve unknown parameters. Ideally we want to optimize all possible false alarm rates and all possible detection delays. Unfortunately this cannot be done, and there is no attractive definition of optimality in the literature for this problem. In Chapter 3, we propose a useful definition of "asymptotically optimal to first-order" procedures, thereby generalizing Lorden's asymptotic theory, and prove the optimality of our proposed procedures.

Recently change-point problems have been applied in so-called "decentralized" or "distributed" decision systems, which have many important applications, including multi-sensor data fusion, mobile and wireless communication, surveillance systems, and economic theory of teams (see Blum, Kassam, and Poor [2], and Veeravalli [38], and the many references therein). Figure 1.1 illustrates the general setting of decentralized decision systems. In such a system, at time n, each of a set of sensors S_l receives an observation X_n^l , and then sends a message to a central processor, called the *fusion center*, which makes a final decision when observations are stopped. In order to reduce the communication costs, it is required that the sensor messages belong to a *finite alphabet*. This limitation is dictated in practice by the need for data

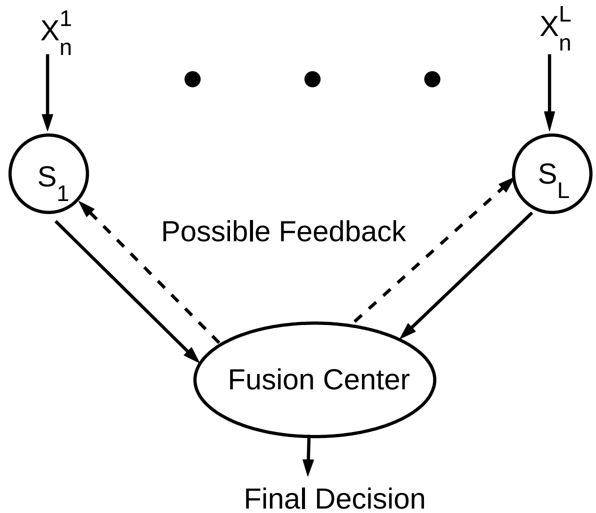

Figure 1.1: General setting for decentralized decision systems

compression and limitations of channel bandwidth. In *decentralized change-point problems*, it is assumed that at some unknown time, the distributions of the sensor observations X_n^l change abruptly and simultaneously at all sensors. The goal is to detect the change as soon as possible over all possible protocols for generating sensor messages and over all possible decision rules at the fusion center, under a restriction on the frequency of false alarms.

Even if it is assumed that the observations are independent in time as well as from sensor to sensor and the pre-change distributions and post-change distributions are completely specified, little research has been done on optimality theory under a minimax criterion. Veeravalli, Basar, and Poor [40] and Veeravalli [38] pointed out that there are five possible information structures depending on how the local information and possible "feedback" are used at the sensors. Crow and Schwartz [6] and Tartakovsky and Veeravalli [36] have studied the simplest information structure using a minimax approach, but both have restrictions on the class of sensor message protocols. Chapter 4 develops optimality theory for all five information structures in decentralized change-point problems and offers asymptotically optimal procedures which are easy to implement. For that purpose, we develop an asymptotic theory of sequential hypothesis testing in decentralized decision systems, which is of interest on its own.

The problem of detecting a change in distribution for dependent observations is an important topic in the literature, for example, see Lai [11]. It is natural to extend Page's CUSUM procedure or the Shiryayev-Roberts procedure, both of which are based on likelihood ratios, to dependent observations by simply replacing the probability densities with the corresponding conditional densities. Pollak conjectured ([24], [44]) that such procedures are asymptotically optimal under general conditions. In Chapter 5, we disprove this conjecture by constructing a counterexample where the pre-change distribution is a so-called "mixture distribution" and the post-change distribution is fully specified.

In the change-point literature, as in sequential analysis more generally, theorems establishing exact optimality of statistical procedures are quite rare. Moustakides [21] and Ritov [27] showed that for the simplest problem where both pre-change distribution f and post-change distribution g are fully specified, Page's CUSUM procedure is exactly optimal in the following sense: It minimizes the "worst case" detection delay defined in Chapter 2, subject to a specified frequency of false alarms. Besides Page's CUSUM procedure and its generalizations, the most commonly used and studied approach to define change-point procedures is that of Shiryayev [30] and Roberts [28]. Yakir ([43], 1997) published a proof that when both f and g are fully specified, a modification of the Shiryayev-Roberts procedure is exactly optimal with respect to a slightly different measure of quickness of detection. In Chapter 5 we show that Yakir's proof is wrong. It is still an open problem whether the modified Shiryayev-Roberts procedure is in fact optimal, although its asymptotic optimality was proved in Pollak [24].

Chapter 2

Preliminaries

This chapter covers certain preliminaries to the thesis results. We briefly discuss sequential statistical hypothesis testing and then we discuss the simplest change-point problems. Here we put emphasis on what we will need in later chapters. For a more general setting and background, we refer interested readers to Basseville and Nikiforov [1], Govindarajulu [8], and Siegmund [33].

2.1 Sequential Hypotheses Testing

For the usual statistical tests the sample size is fixed before the data are taken, but for a sequential test the total sample size depends on the data and is thus a random variable. We are interested in sequential tests because they are economical in the sense that we may reach a decision earlier via a sequential test than via a fixed sample size test. The criteria for the choice of sequential tests are the operating characteristic (OC) and the average sample number (ASN) functions which will be defined in the following.

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables observed sequentially, i.e., one at a time, and let p be their density function. Suppose we want to test the null hypothesis $H_0 : p = f_\theta$ for some $\theta \in \Theta$ against the alternative hypothesis $H_1 : p = g_\lambda$ for some $\lambda \in \Lambda$. Here $\{f_\theta, \theta \in \Theta\}$ and $\{g_\lambda, \lambda \in \Lambda\}$ are two disjoint subsets of probability densities. Denote by $\mathbf{P}_\theta$ and $\mathbf{P}_\lambda$ respectively the distribution when $X_1, X_2, \ldots$ are independent and identically distributed with density

f_θ and g_λ. Let $\mathbf{E}_\theta$ and $\mathbf{E}_\lambda$ be the corresponding expectations.

The operating characteristic (OC) function of a sequential test is determined by

(i) the probability of type I error, $\mathbf{P}_\theta$(test rejects H_0); and

(ii) the probability of type II error, $\mathbf{P}_\lambda$(test accepts H_0).

It is desirable that these error probabilities should be as small as possible. A typical requirement for a sequential test is that $\mathbf{P}_\theta(\text{Reject } H_0) \le \alpha$ for all possible θ and $\mathbf{P}_\lambda(\text{Accept } H_0) \le \beta$ for all possible λ, where α and β are given error probability bounds.

As noted earlier, the number of observations, say N, required by a sequential test is random, and so the distribution of N must be considered. It is usually characterized by the expected value of N when f_θ or g_λ is the true underlying density. Thus, we can define the average sample number (ASN) functions as

(i) $\mathbf{E}_\theta N$; and

(ii) $\mathbf{E}_\lambda N$.

Another desirable property of a sequential test is to have small ASN functions, e.g., small subject to the error probability bounds α and β.

Note that N can also be thought of as the time at which the test decides to stop taking observations since we observe X's one at a time. Since the decision to stop at time N is only based on the first N observations, N is called a stopping time:

Definition 2.1. (Stopping time). *An integer-valued (or possibly ∞) random variable N is said to be a stopping time with respect to the sequence $X_1, X_2, \dots$ if for every $1 \le n < \infty$, the event $\{N = n\}$ depends only on $X_1, X_2, \dots, X_n$.*

Let us first consider the simplest case. Suppose $X_1, X_2, \dots$ are independent and identically distributed with probability density function p, and we are interested in testing a simple hypothesis $H_0 : p = f_\theta$ against a simple alternative hypothesis $H_1 : p = g_\lambda$, where both θ and λ are given.

The best procedure, in a certain sense, is the following sequential probability ratio test (SPRT), which was developed by Abraham Wald during World War II. Choose two constants A and B such that $0 < A < B < \infty$, and define the likelihood ratio

$$L_n = \prod_{i=1}^{n} \frac{g_\lambda(X_i)}{f_\theta(X_i)}.$$

Define a stopping time $N =$ first $n \geq 1$ such that $L_n \notin (A, B)$. In other words,

$$N = \inf\{n \geq 1 : L_n \notin (A, B)\}. \tag{2.1}$$

The SPRT will stop sampling at time N and

$$\begin{aligned} \text{reject } H_0 \quad & \text{if } L_N \geq B, \\ \text{accept } H_0 \quad & \text{if } L_N \leq A. \end{aligned}$$

Theorem 2.2. (Wald and Wolfowitz [41], 1948). *Among all tests (sequential or not) for which*

$$\mathbf{P}_\theta(\textit{ Reject } H_0) \leq \alpha, \quad \textit{and} \quad \mathbf{P}_\lambda(\textit{ Accept } H_0) \leq \beta$$

and for which $\mathbf{E}_\theta N$ *and* $\mathbf{E}_\lambda N$ *are finite, the sequential probability ratio test (SPRT) with error probability* α *and* β *minimizes both* $\mathbf{E}_\theta N$ *and* $\mathbf{E}_\lambda N$ *simultaneously.*

Remark: Lorden [19] showed that the condition that the expected sample sizes are finite is not needed.

In order to determine the OC and ASN functions of the SPRT, Wald developed the following two important propositions which will also be useful in later chapters.

Proposition 2.3. (Wald's Equation). *Let* $X_1, X_2, \ldots$ *be independent and identically distributed with mean* $\mu = \mathbf{E}X_1$*. For any stopping time* N *with* $\mathbf{E}N < \infty$,

$$\mathbf{E}(\sum_{i=1}^{N} X_i) = \mu \; \mathbf{E}N. \tag{2.2}$$

Proposition 2.4. (Wald's Likelihood Ratio Identity). *Suppose* $X_1, X_2, \dots$ *are independent and identically distributed with density* f *and* g *under* $\mathbf{P}_f$ *and* $\mathbf{P}_g$ *respectively. Define the likelihood ratio sequence*

$$L_n = \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)}.$$

For any stopping time N

$$\mathbf{P}_f(N < \infty) = \mathbf{E}_g(L_N^{-1}; N < \infty). \tag{2.3}$$

In the above we consider SPRTs to test a simple null hypothesis against a simple alternative hypothesis. For the problem of testing a composite hypothesis against a composite alternative, we refer interested readers to Govindarajulu [8], Kiefer and Sacks [9], and Siegmund [33].

2.2 Change-Point Problems

Suppose there is a process that produces a sequence of independent observations $X_1, X_2, \dots$. Initially the process is "in control" and the true distribution of the X's is f_θ for some $\theta \in \Theta$. At some unknown time ν, the process goes "out of control" in the sense that the distribution of $X_\nu, X_{\nu+1}, \dots$ is g_λ for some $\lambda \in \Lambda$. It is desired to raise an alarm as soon as the process is out of control so that we can take appropriate action. This is known as a change-point problem, or quickest change detection problem.

Denote by $\mathbf{P}^{\nu}_{\theta,\lambda}$ the probability measure under which $X_1, \dots, X_{\nu-1}$ are distributed according to a "pre-change" distribution f_θ for some $\theta \in \Theta$ and $X_\nu, X_{\nu+1}, \dots$ are distributed according to a "post-change" distribution g_λ for some $\lambda \in \Lambda$. We shall also use $\mathbf{P}_\theta$ to denote the probability measure under which $X_1, X_2, \dots$ are independent and identically distributed with density f_θ (corresponding to $\nu = \infty$). By analogy with hypothesis testing terminology, we will refer to Θ (Λ) as a "simple" pre-change (post-change) hypothesis if it contains a single point and as a "composite" pre-change (post-change) hypotheses if it contains more than one point.

In change-point problems, a procedure for detecting that a change has occurred is defined as a stopping time N. The interpretation of N is that, when $N = n$, we stop at n and declare that a change has occurred somewhere in the first n observations. We want to find a stopping time N which will stop as soon as possible after a change occurs but will continue sampling as long as possible if no change occurs. Thus, the performance of a stopping time N is evaluated by two criteria: the *detection delay*, and the *frequency of false alarms*, which we next define.

The detection delay can be evaluated by the following "worst case" detection delay defined in Lorden [16],

$$\overline{\mathbf{E}}_\lambda N = \sup_{\nu \geq 1} \Big(\operatorname{ess\,sup} \mathbf{E}^{\nu}_{\theta,\lambda} \big[(N - \nu + 1)^+ | X_1, \dots, X_{\nu-1} \big] \Big). \tag{2.4}$$

Note that the definition of $\overline{\mathbf{E}}_\lambda N$ does not depend upon the pre-change distribution f_θ by virtue of the essential superum, which takes the "worst possible X's before the change." The detection delay is also called the short *Average Run Length (ARL).*

It turns out that if $\overline{\mathbf{E}}_\lambda N$ is finite, then $\mathbf{P}_\theta(N < \infty) = 1$, which means that we will have a false alarm with probability 1 even when there are no changes. The false alarm rate is usually measured by $1/\mathbf{E}_\theta N$, where $\mathbf{E}_\theta N$ is often called the long *Average Run Length (ARL).* Imagining repeated application of such procedures, practitioners refer to the *frequency of false alarms* as $1/\mathbf{E}_\theta N$ and the *mean time between false alarms* as $\mathbf{E}_\theta N$.

A good procedure N should have $\mathbf{E}_\theta N$ large for all $\theta \in \Theta$ while keeping $\overline{\mathbf{E}}_\lambda N$ small for all $\lambda \in \Lambda$. In general, however, it is harder to find any sort of optimal procedures when Θ (or Λ) is composite because there are no procedures which simultaneously optimize over Θ (or Λ). The problem of finding an optimal procedure is solvable in the simplest case where both Θ and Λ are simple, i.e., the pre-change distribution f_θ and the post-change distribution g_λ are given. The (exactly) optimal procedures are given by Page's CUSUM method, which uses the stopping times

$$T_a = \inf \big\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log \frac{g_\lambda(X_i)}{f_\theta(X_i)} \geq a \big\}. \tag{2.5}$$

Theorem 2.5. (Moustakides [21], 1986). *For any $a > 0$, Page's CUSUM procedure T_a minimizes the worst-case detection delay $\overline{\mathbf{E}}_\lambda N$ among all stopping times N satisfying $\mathbf{E}_\theta N \geq \mathbf{E}_\theta T_a$.*

Unfortunately, in hypothesis testing and change-point problems, exact optimality theorems such as Theorems 2.2 and 2.5 are limited to the simplest case where both Θ and Λ are simple. In more practical problems where Θ or Λ is composite, no procedure can exactly and simultaneously optimize over Θ and Λ, and only "asymptotic optimality" theorems have been established. In the asymptotic optimality approach, we typically first construct an (asymptotic) lower bound as the ASN functions goes to ∞ in hypothesis testing problems or as the mean time between false alarms $\mathbf{E}_\theta N$ goes to ∞ in change-point problems. Then we show that a given class of statistical procedures attains the lower bound asymptotically. In this thesis, we establish asymptotic optimality theorems in various contexts.

Chapter 3

Composite Pre-Change Hypotheses

In this chapter we study change-point problems when unknown parameters are present in the pre-change distribution. First, we study the case where all distributions are from a one-parameter exponential family of densities and the post-change distribution is completely specified. Then we consider the general case where both the pre-change distribution and the post-change distribution involve unknown parameters.

3.1 Simple Post-Change Hypotheses

Assume that the probability density of the data is indexed by θ and written as $f_\theta(x)$. In parametric change-point problems, standard formulations assume that θ is equal to some known value θ_0 before a change occurs. When the true θ is unknown before a change occurs, it is typical to assume that a training sample is available so that one can use the method of "point estimation" to obtain a value θ_0. However, it is well known that the performance of such procedures is very sensitive to the error in estimating θ. (See Stoumbos, Reynolds, Ryan and Woodall [35]). Thus we need to study change-point problems for composite pre-change hypotheses, which allow a range of "acceptable" values of θ.

There are many practical situations where the need to take action in response to a change in a parameter θ is definable by a fixed threshold value. For example, suppose we are interested in monitoring water quality and that a contaminant A is considered unacceptable if it reaches or exceeds level b. Another example occurs when one is monitoring or tracking a weak signal in a noisy environment. If the signal disappears, one wants to detect the disappearance as quickly as possible. Parameters θ associated with the signal, e.g. its strength, are described by a composite hypothesis before it disappears, but by a simple hypothesis (strength equal to zero) afterwards.

In this section we consider the problem of detecting a change in distribution from f_θ to f_λ where $\theta \in \Theta$, an interval on the real line, and λ is a given value outside the interval Θ. The problem is to find a stopping time N such that the mean time between false alarms, $\mathbf{E}_\theta N$, is as large as possible for all $\theta \in \Theta$ subject to the constraint

$$\overline{\mathbf{E}}_\lambda N \leq \gamma, \tag{3.1}$$

where $\gamma > 0$ is a given constant and $\lambda \notin \Theta$.

One cannot maximize $\mathbf{E}_\theta N$ for all $\theta \in \Theta$ subject to (3.1) since the maximum for each θ is uniquely attained by Page's CUSUM procedure for detecting a change in distribution from f_θ to f_λ. A natural idea is to maximize $\mathbf{E}_\theta N$ asymptotically for all $\theta \in \Theta$. Lorden [16] showed that for each pair (θ, λ)

$$\overline{\mathbf{E}}_\lambda N \geq (1 + o(1)) \frac{\log \mathbf{E}_\theta N}{I(\lambda, \theta)}, \tag{3.2}$$

as $\mathbf{E}_\theta N \to \infty$. This suggests defining the asymptotic efficiency of a family $\{N(a)\}$ as

$$e(\theta, \lambda) = \liminf_{a \to \infty} \frac{\log \mathbf{E}_\theta N(a)}{I(\lambda, \theta) \overline{\mathbf{E}}_\lambda N(a)}, \tag{3.3}$$

where $\{N(a)\}$ is required to satisfy $\mathbf{E}_\theta N(a) \to \infty$ as $a \to \infty$. Then $e(\theta, \lambda) \leq 1$ for all families, so we can define

Definition 3.1. *$\{N(a)\}$ is* asymptotically efficient *at (θ, λ) if $e(\theta, \lambda) = 1$.*

The problem in this section is to find a family of stopping times that is asymptotically efficient at (θ, λ) for all $\theta \in \Theta$.

In this section, it will be assumed that $\{f_\xi\}_{\xi \in \Omega}$ are the densities of a one-parameter exponential family with natural parameter space Ω with respect to a sigma-finite measure F. Denote

$$f_\xi(x) = \exp(\xi x - b(\xi)), \quad -\infty < x < \infty, \quad \xi \in \Omega, \tag{3.4}$$

and assume $\Omega = (\underline{\xi}, \bar{\xi})$ is the natural parameter space of ξ. Let $\mathbf{P}_\xi, \mathbf{E}_\xi$ denote the probability measure and expectation, respectively, when $X_1, X_2, \ldots$ are independent and identically distributed with density f_ξ. Differentiating the identity $\int e^{\xi x - b(\xi)} dF(x) = 1$ gives us

$$\mathbf{E}_\xi X_i = b'(\xi) \quad \text{and} \quad \mathrm{Var}_\xi(X_i) = b''(\xi) \geq 0.$$

In particular, b is strictly convex (unless F is degenerate at a point).

Before studying change-point problems in Section 3.2.2, we first consider the corresponding open-ended hypothesis testing problems in Section 3.2.1, since the basic arguments are clearer for hypothesis testing problems and are readily extendable to change-point problems.

3.1.1 Open-Ended Hypothesis Testing

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with probability density f_ξ of the form (3.4) on the natural parameter space $\Omega = (\underline{\xi}, \bar{\xi})$. Suppose we are interested in testing the null hypothesis

$$H_0 : \xi \in \Theta = [\theta_0, \theta_1]$$

against the alternative hypothesis

$$H_1 : \xi \in \Lambda = \{\lambda\},$$

where $\underline{\xi} < \theta_0 < \theta_1 < \lambda < \bar{\xi}$.

Motivated by applications to change-point problems, we consider the following "open-ended" hypothesis testing problems. Assume that if H_0 is true, sampling costs nothing and our preferred action is just to observe $X_1, X_2, \ldots$ without stopping. On the other hand, if H_1 is true, each observation costs a fixed amount and we want to stop sampling as soon as possible and reject the null hypothesis H_0.

Since there is only one terminal decision, a statistical procedure for an open-ended hypothesis testing problem is a stopping time N. The null hypothesis H_0 is rejected if and only if $N < \infty$. A good procedure N should keep the error probabilities $\mathbf{P}_\theta(N < \infty)$ small for all $\theta \in \Theta$ while keeping $\mathbf{E}_\lambda N$ small.

The problem in this subsection is to find a stopping time N such that $\mathbf{P}_\theta(N < \infty)$ will be as small as possible for all $\theta \in \Theta = [\theta_0, \theta_1]$ subject to the constraint

$$\mathbf{E}_\lambda N \leq \gamma, \tag{3.5}$$

where $\gamma > 0$ is a given constant.

Unfortunately we cannot minimize all error probabilities simultaneously. A natural approach is to employ an asymptotic formulation. By Proposition 3.2 below, for any stopping time N,

$$\frac{|\log \mathbf{P}_\theta(N < \infty)|}{I(\lambda, \theta)} \leq \gamma \tag{3.6}$$

for all $\theta \in \Theta = [\theta_0, \theta_1]$. This suggests choosing a family of stopping times which asymptotically maximizes

$$\frac{|\log \mathbf{P}_\theta(N < \infty)|}{I(\lambda, \theta)} \tag{3.7}$$

for all $\theta \in \Theta$ subject to (3.5), as $\gamma \to \infty$.

Proposition 3.2. *For any stopping time N, relation (3.5) implies (3.6).*

Proof.

$$\begin{aligned} I(\lambda,\theta)\ \mathbf{E}_\lambda N &= \mathbf{E}_\lambda\Big(\sum_{i=1}^{N}\log\frac{f_\lambda(X_i)}{f_\theta(X_i)}\Big) && \text{by Proposition 2.3} \\ &= \mathbf{E}_\lambda\Big(-\log\prod_{i=1}^{N}\frac{f_\theta(X_i)}{f_\lambda(X_i)}\Big) \\ &\geq -\log\mathbf{E}_\lambda\Big(\prod_{i=1}^{N}\frac{f_\theta(X_i)}{f_\lambda(X_i)}\Big) && \text{by Jensen's inequality} \\ &= -\log\mathbf{P}_\theta(N<\infty) && \text{by Proposition 2.4} \end{aligned}$$

Relation (3.6) follows at once from the fact that $\mathbf{P}_\theta(N<\infty)\leq 1$. □

Next, we propose a class of stopping times which is asymptotically optimal. Our proposed open-ended tests of $H_0:\theta_0\leq\xi\leq\theta_1$ versus $H_1:\xi=\lambda\ (>\theta_1)$ are defined by the stopping times

$$M(a)=\inf\Big\{n\geq 1:\inf_{\theta_0\leq\theta\leq\theta_1}\big(\sum_{i=1}^{n}\log\frac{f_\lambda(X_i)}{f_\theta(X_i)}-I(\lambda,\theta)a\big)>0\Big\}. \tag{3.8}$$

Theorem 3.3. *Suppose $\{N(a)\}$ are stopping times such that $\mathbf{E}_\lambda N(a)\leq\mathbf{E}_\lambda M(a)$. For all $\theta_0\leq\theta\leq\theta_1$ as $a\to\infty$*

$$\frac{|\log\mathbf{P}_\theta(N(a)<\infty)|}{I(\lambda,\theta)}\leq a+(C+o(1))\sqrt{a}, \tag{3.9}$$

and

$$\frac{|\log\mathbf{P}_\theta(M(a)<\infty)|}{I(\lambda,\theta)}\geq a, \tag{3.10}$$

where

$$C=\Big(\frac{\lambda-\theta_1}{I(\lambda,\theta_1)}-\frac{\lambda-\theta_0}{I(\lambda,\theta_0)}\Big)\cdot\sqrt{\frac{b''(\lambda)}{2\pi}}. \tag{3.11}$$

The proof is based on the following lemmas.

Lemma 3.4. *For $\theta \in [\theta_0, \theta_1]$, define*

$$M_\theta(a) = \inf\Big\{n \geq 1 : \sum_{i=1}^{n} \log \frac{f_\lambda(X_i)}{f_\theta(X_i)} - I(\lambda, \theta)a > 0\Big\}. \qquad (3.12)$$

Then

$$\mathbf{P}_\theta(M(a) < \infty) \leq \mathbf{P}_\theta(M_\theta(a) < \infty) \leq \exp(-I(\lambda, \theta)a), \qquad (3.13)$$

and hence (3.10) holds.

Proof. The first inequality in (3.13) follows at once from the fact that $M(a) \geq M_\theta(a)$ for all $\theta \in [\theta_0, \theta_1]$. By Wald's likelihood ratio identity,

$$\mathbf{P}_\theta(N < \infty) = \mathbf{E}_\lambda(\exp(-l_N); N < \infty)$$

where $l_n = \sum_{i=1}^n \log(f_\lambda(X_i)/f_\theta(X_i))$. The second inequality in (3.13) follows from the fact that $l_{M_\theta(a)} \geq I(\lambda, \theta)a$. □

We now derive approximations for $\mathbf{E}_\lambda M(a)$. Observe that the log-likelihood ratios

$$\sum_{i=1}^{n} \log \frac{f_\lambda(X_i)}{f_\theta(X_i)} = (\lambda - \theta)S_n - n(b(\lambda) - b(\theta)),$$

where $S_n = X_1 + \cdots + X_n$. Thus the Kullback-Leibler information numbers can be written

$$I(\lambda, \theta) = \mathbf{E}_\lambda \log(f_\lambda(X)/f_\theta(X)) = (\lambda - \theta)b'(\lambda) - (b(\lambda) - b(\theta)),$$

while

$$\mathrm{Var}_\lambda\big(\log(f_\lambda(X)/f_\theta(X))\big) = (\lambda - \theta)^2 b''(\lambda).$$

Hence, for any $\theta \in [\theta_0, \theta_1]$,

$$\sum_{i=1}^{n} \log \frac{f_\lambda(X_i)}{f_\theta(X_i)} - I(\lambda, \theta)a = (\lambda - \theta)\Big[S_n - b'(\lambda)a - (n - a)\frac{b(\lambda) - b(\theta)}{\lambda - \theta}\Big].$$

Define

$$\phi(\theta) = \frac{b(\lambda) - b(\theta)}{\lambda - \theta}. \tag{3.14}$$

Then for $\theta \in [\theta_0, \theta_1]$, the stopping time $M_\theta(a)$ defined in (3.12) can be written as

$$M_\theta(a) = \inf\Big\{n \geq 1 : S_n \geq b'(\lambda)a + (n-a)\phi(\theta)\Big\}, \tag{3.15}$$

since $\lambda > \theta_1 \geq \theta$. Similarly, the stopping time $M(a)$ can be written as

$$M(a) = \inf\Big\{n \geq 1 : S_n \geq b'(\lambda)a + \sup_{\theta_0 \leq \theta \leq \theta_1} \big[(n-a)\phi(\theta)\big]\Big\}. \tag{3.16}$$

Since $b(\theta)$ is convex, $\phi(\theta)$ is a strictly increasing function of θ and $b'(\lambda) > \phi(\theta_1)$. Hence the supremum in (3.16) is attained at $\theta = \theta_0$ if $n \leq a$, and at $\theta = \theta_1$ if $n > a$, so that

$$\{M(a) = m\} = \{M(a) = M_{\theta_0}(a) = m\} \quad \text{for all} \quad m \leq a. \tag{3.17}$$

For simplicity, we omit a and θ, writing $M = M(a)$ and $M_i = M_{\theta_i}(a)$ for $i = 0, 1$.

Lemma 3.5. *For $a > 0$,*

$$\mathbf{E}_\lambda M(a) \geq a + \frac{\phi(\theta_1) - \phi(\theta_0)}{b'(\lambda) - \phi(\theta_1)} \mathbf{E}_\lambda(a - M_0; M_0 \leq a). \tag{3.18}$$

Proof. Since $\mathbf{E}_\lambda X_i = b'(\lambda)$, by Wald's equation, $b'(\lambda)\mathbf{E}_\lambda M = \mathbf{E}_\lambda S_M$. By (3.16), for all $\theta \in [\theta_0, \theta_1]$, $S_M - b'(\lambda)a \geq (M-a)\phi(\theta)$. In particular, it holds for $\theta = \theta_0$ or θ_1. Thus

$$\begin{aligned}
b'(\lambda)\mathbf{E}_\lambda M &= \mathbf{E}_\lambda S_M = b'(\lambda)a + \mathbf{E}_\lambda(S_M - b'(\lambda)a; M \leq a) + \mathbf{E}_\lambda(S_M - b'(\lambda)a; M > a) \\
&\geq b'(\lambda)a + \mathbf{E}_\lambda\Big((M-a)\phi(\theta_0); M \leq a\Big) + \mathbf{E}_\lambda\Big((M-a)\phi(\theta_1); M > a\Big) \\
&= b'(\lambda)a + \phi(\theta_0)\mathbf{E}_\lambda(M-a; M \leq a) + \phi(\theta_1)\mathbf{E}_\lambda(M-a; M > a)
\end{aligned}$$

Plugging in $\mathbf{E}_\lambda(M-a; M>a) = \mathbf{E}_\lambda M - a - \mathbf{E}_\lambda(M-a; M\le a)$, we get

$$\mathbf{E}_\lambda M \ge a + \frac{\phi(\theta_1)-\phi(\theta_0)}{b'(\lambda)-\phi(\theta_1)}\mathbf{E}_\lambda(a-M; M\le a). \tag{3.19}$$

Relation (3.18) follows at once from (3.17) and (3.19). □

Lemma 3.6. *For $a>0$,*

$$\mathbf{E}_\lambda M < a + \frac{\phi(\theta_1)-\phi(\theta_0)}{b'(\lambda)-\phi(\theta_1)}\mathbf{E}_\lambda(a-M_0; M_0\le a) + C_1, \tag{3.20}$$

where C_1 depends on θ_0, θ_1 and λ but not on a.

Proof. Observe that

$$\mathbf{E}_\lambda M = a - \mathbf{E}_\lambda(a-M; M\le a) + \mathbf{E}_\lambda(M-a; M>a),$$

and $\mathbf{E}_\lambda(M-a; M\le a) = \mathbf{E}_\lambda(M_0-a; M_0\le a)$. Thus it suffices to show that

$$\mathbf{E}_\lambda(M-a; M>a) < \frac{b'(\lambda)-\phi(\theta_0)}{b'(\lambda)-\phi(\theta_1)}\mathbf{E}_\lambda(a-M_0; M_0\le a) + C_1. \tag{3.21}$$

To prove this inequality, define a stopping time

$$N_i(a) = \inf\Big\{n\ge 1 : \sum_{i=1}^{n}(X_i-\phi(\theta_i)) \ge a\Big\},$$

for $i=0,1$. Assume a is an integer. (For general a, using $[a]$, the largest integer $\le a$, permits one to carry through the following argument with minor modifications). By (3.17), we have

$$\begin{aligned}\mathbf{E}_\lambda\Big(M-a\Big|M>a\Big) &= \int_{-\infty}^{0}\mathbf{E}_\lambda\Big(M-a\Big|S_a-b'(\lambda)a=x, M_0>a\Big)\\ &\qquad \mathbf{P}_\lambda\Big(S_a-b'(\lambda)a\in dx\Big|M_0>a\Big).\end{aligned}$$

Conditioned on the event $\{S_a - b'(\lambda)a = x, M_0 > a\}$,

$$\begin{aligned} M - a &= \inf\Big\{m \geq 1 : X_{a+1} + \cdots + X_{a+m} + S_a \geq b'(\lambda)a + m\phi(\theta_1)\Big\} \\ &= \inf\Big\{m \geq 1 : \sum_{i=1}^{m}(X_{a+i} - \phi(\theta_1)) \geq b'(\lambda)a - S_a = -x\Big\}, \end{aligned}$$

which is equivalent to $N_1(-x)$ since $X_1, X_2, \ldots$ are independent and identically distributed. Thus

$$\mathbf{E}_\lambda\Big(M - a\Big|M > a\Big) = \int_{-\infty}^{0} \mathbf{E}_\lambda N_1(-x)\ \mathbf{P}_\lambda\Big(S_a - b'(\lambda)a \in dx\Big|M_0 > a\Big). \tag{3.22}$$

Similarly,

$$\mathbf{E}_\lambda\Big(M_0 - a\Big|M_0 > a\Big) = \int_{-\infty}^{0} \mathbf{E}_\lambda N_0(-x)\ \mathbf{P}_\lambda\Big(S_a - b'(\lambda)a \in dx\Big|M_0 > a\Big). \tag{3.23}$$

Now for $i = 0, 1$ and any $a > 0$, define

$$R_i(a) = \sum_{i=1}^{N_i(a)}(X_i - \phi(\theta_i)) - a.$$

Then by Theorem 1 in Lorden [15]

$$\sup_{a\geq 0} \mathbf{E}_\lambda R_i(a) \leq \mathbf{E}_\lambda(X_i - \phi(\theta_i))^2/(b'(\lambda) - \phi(\theta_i)) < \infty.$$

By Wald's equation, $(b'(\lambda) - \phi(\theta_i))\mathbf{E}_\lambda N_i(a) = a + \mathbf{E}_\lambda R_i(a)$, so that

$$\sup_{a\geq 0} \mathbf{E}_\lambda\Big(N_i(a) - \frac{a}{b'(\lambda) - \phi(\theta_i)}\Big) < \infty,$$

for $i = 0, 1$. Hence, there exists a constant C_2 such that for all $a > 0$,

$$\mathbf{E}_\lambda N_1(a) \leq \frac{b'(\lambda) - \phi(\theta_0)}{b'(\lambda) - \phi(\theta_1)}\mathbf{E}_\lambda N_0(a) + C_2.$$

Plugging into (3.22), and comparing it with (3.23), we have

$$\mathbf{E}_\lambda\Big(M-a\Big|M>a\Big)<\frac{b'(\lambda)-\phi(\theta_0)}{b'(\lambda)-\phi(\theta_1)}\mathbf{E}_\lambda\Big(M_0-a\Big|M_0>a\Big)+C_2.$$

Since $\{M>a\}=\{M_0>a\}$, we have

$$\mathbf{E}_\lambda(M-a;M>a)<\frac{b'(\lambda)-\phi(\theta_0)}{b'(\lambda)-\phi(\theta_1)}\mathbf{E}_\lambda(M_0-a;M_0>a)+C_2\mathbf{P}_\lambda(M_0>a). \quad (3.24)$$

Again, using Theorem 1 in Lorden [15] and Wald's equation, we have

$$\sup_{a\geq 0}(\mathbf{E}_\lambda M_0-a)<C_3$$

for some constant C_3. Thus

$$\begin{aligned}\mathbf{E}_\lambda(M_0-a;M_0>a) &= \mathbf{E}_\lambda M_0-a+\mathbf{E}_\lambda(a-M_0;M_0\leq a)\\ &\leq C_3+\mathbf{E}_\lambda(a-M_0;M_0\leq a).\end{aligned}$$

Relation (3.21) follows at once from (3.24) and the fact that $\mathbf{P}_\lambda(M_0>a)\leq 1$. Hence, the lemma holds. □

Lemma 3.7. *Suppose $Y_1,Y_2,\ldots$ are independent and identically distributed with mean $\mu>0$ and finite variance σ^2. Define*

$$N_a=\inf\Big\{n\geq 1:\sum_{i=1}^{n}Y_i\geq a\Big\}.$$

Then as $a\to\infty$.

$$\mathbf{E}(\frac{a}{\mu}-N_a;N_a\leq\frac{a}{\mu})=\sqrt{a}\Big(\frac{\sigma}{\sqrt{2\pi\mu^3}}+o(1)\Big).$$

Proof. It is well known that as $a\to\infty$,

$$\mathbf{E}N_a=\frac{a}{\mu}+O(1)\quad\text{and}\quad \mathrm{Var}(N_a)=(1+o(1))\frac{a\sigma^2}{\mu^3}, \quad (3.25)$$

and that

$$\hat{N}_a = \frac{N_a - a/\mu}{\sqrt{a\sigma^2/\mu^3}}$$

is asymptotically standard normal. (See page 372 in Feller [7], and equation (5) in Siegmund [32]).

The asymptotic normality of $\hat{N}_a$ suggests that, letting Z denote a standard normal random variable,

$$\begin{aligned} \mathbf{E}\Big(\hat{N}_a; \hat{N}_a \le 0\Big) &= \mathbf{E}\big(Z; Z \le 0\big) + o(1) \\ &= -\frac{1}{\sqrt{2\pi}} + o(1), \end{aligned} \tag{3.26}$$

and thus

$$\mathbf{E}\Big(\frac{a}{\mu} - N_a; N_a \le \frac{a}{\mu}\Big) = -\sqrt{a\sigma^2/\mu^3}\ \mathbf{E}\Big(\hat{N}_a; \hat{N}_a \le 0\Big) = \sqrt{a\sigma^2/\mu^3}\Big(\frac{1}{\sqrt{2\pi}} + o(1)\Big),$$

which would prove the lemma.

To prove (3.26) note that by (3.25),

$$\mathbf{E}\hat{N}_a = O(\frac{1}{\sqrt{a}}) \quad \text{and} \quad \mathrm{Var}(\hat{N}_a) = 1 + o(1),$$

and hence $\mathbf{E}\hat{N}_a^2 = (\mathbf{E}\hat{N}_a)^2 + \mathrm{Var}(\hat{N}_a) = 1 + o(1)$, as $a \to \infty$. Pick $\eta > 0$ and note that

$$0 \ge \mathbf{E}\big(\hat{N}_a; \hat{N}_a \le -\eta\big) \ge -\mathbf{E}\big(\frac{\hat{N}_a^2}{\eta}; \hat{N}_a \le -\eta\big) \ge -\frac{1}{\eta}\mathbf{E}\big(\hat{N}_a^2\big) = -\frac{1 + o(1)}{\eta}.$$

Since $\mathbf{E}\big(\hat{N}_a; \hat{N}_a \le 0\big) = \mathbf{E}\big(\hat{N}_a; -\eta \le \hat{N}_a \le 0\big) + \mathbf{E}\big(\hat{N}_a; \hat{N}_a \le -\eta\big)$, we have

$$0 \ge \mathbf{E}\big(\hat{N}_a; \hat{N}_a \le 0\big) - \mathbf{E}\big(\hat{N}_a; -\eta \le \hat{N}_a \le 0\big) \ge -\frac{1 + o(1)}{\eta}.$$

Now the bounded convergence theorem implies that

$$\mathbf{E}\big(\hat{N}_a; -\eta \le \hat{N}_a \le 0\big) \to \mathbf{E}\big(Z; -\eta \le Z \le 0\big), \quad \text{as} \ \ a \to \infty.$$

Thus,

$$\begin{aligned} 0 &\geq \limsup_{a\to\infty} \mathbf{E}\big(\hat{N}_a; \hat{N}_a \leq 0\big) - \mathbf{E}\big(Z; -\eta \leq Z \leq 0\big) \\ &\geq \liminf_{a\to\infty} \mathbf{E}\big(\hat{N}_a; \hat{N}_a \leq 0\big) - \mathbf{E}\big(Z; -\eta \leq Z \leq 0\big) \geq -\frac{1+o(1)}{\eta}. \end{aligned}$$

Relation (3.26) follows by letting $\eta \to \infty$. □

Proof of Theorem 3.3: Relation (3.10) is proved in Lemma 3.4, so it suffices to prove (3.9). By (3.12), $M_0 = M_{\theta_0}(a)$ can be written as

$$M_0 = \inf\Big\{n \geq 1 : \sum_{i=1}^{n} \frac{1}{I(\lambda, \theta_0)} \log \frac{f_\lambda(X_i)}{f_{\theta_0}(X_i)} > a\Big\}.$$

Note that

$$\mathbf{E}_\lambda\Big(\frac{1}{I(\lambda,\theta_0)} \log \frac{f_\lambda(X)}{f_{\theta_0}(X)}\Big) = 1,$$

and

$$\mathrm{Var}_\lambda\Big(\frac{1}{I(\lambda,\theta_0)} \log \frac{f_\lambda(X)}{f_{\theta_0}(X)}\Big) = \frac{b''(\lambda)}{(b'(\lambda) - \phi(\theta_0))^2},$$

which we denote by σ_0^2. Thus by Lemma 3.7,

$$\mathbf{E}_\lambda\Big(a - M_0; M_0 \leq a\Big) = \sqrt{a}\Big(\frac{\sigma_0}{\sqrt{2\pi}} + o(1)\Big). \tag{3.27}$$

Using Lemma 3.5 and Lemma 3.6, we have

$$\mathbf{E}_\lambda M = a + \frac{\phi(\theta_1) - \phi(\theta_0)}{b'(\lambda) - \phi(\theta_1)} \mathbf{E}_\lambda(a - M_0; M_0 \leq a) + O(1).$$

By (3.27) and the definition of $\phi(\theta)$ in (3.14), we have

$$\mathbf{E}_\lambda M = a + (C + o(1))\sqrt{a}, \tag{3.28}$$

as $a \to \infty$, where C is defined in (3.11).

It follows that for all $\theta \in [\theta_0, \theta_1]$,

$$\frac{|\log \mathbf{P}_\theta(N(a) < \infty)|}{I(\lambda, \theta)} \leq \mathbf{E}_\lambda N(a) \leq \mathbf{E}_\lambda M(a) = a + (C + o(1))\sqrt{a},$$

The first inequality is from Proposition 3.2. The second one is the condition of the theorem. Thus (3.9) holds, and the theorem is proved. □

3.1.2 Change-Point Problems

Now let us consider the problem of detecting a change in distribution from f_θ for some $\theta \in \Theta = [\theta_0, \theta_1]$ to f_λ. As described earlier, we seek a family of stopping times that is asymptotically efficient at (θ, λ) for all $\theta \in \Theta$.

A method to find such a family is suggested by the following result, which indicates the relationship between open-ended hypothesis testing and change-point problems.

Theorem 3.8. (Lorden [16]). *Let N be a stopping time with respect to $X_1, X_2, \ldots$ such that $\mathbf{P}_\theta(N < \infty) \leq \alpha$. For $k = 1, 2, \ldots,$ let N_k denote the stopping time obtained by applying N to $X_k, X_{k+1}, \ldots$ and define*

$$N^* = \min_{k \geq 1}(N_k + k - 1),$$

Then N^* is a stopping time with

$$\mathbf{E}_\theta N^* \geq \frac{1}{\alpha},$$

and for any λ,

$$\overline{\mathbf{E}}_\lambda N^* \leq \mathbf{E}_\lambda N.$$

Let $M(a)$ be the stopping time defined in (3.8), and let $M_k(a)$ be the stopping time obtained by applying $M(a)$ to the observations $X_k, X_{k+1}, \ldots$. Define a new stopping time by $M^*(a) = \min_{k\geq 1}\big(M_k(a)+k-1\big)$. In other words,

$$M^*(a) = \inf\Big\{n \geq 1 : \max_{1\leq k\leq n} \inf_{\theta_0\leq\theta\leq\theta_1} \big(\sum_{i=k}^{n} \log\frac{f_\lambda(X_i)}{f_\theta(X_i)} - I(\lambda,\theta)a\big) > 0\Big\}. \tag{3.29}$$

The next theorem shows that the family $\{M^*(a)\}$ is asymptotically efficient at (θ,λ) for all $\theta\in\Theta$.

Theorem 3.9. *For any $a>0$ and $\theta_0\leq\theta\leq\theta_1$,*

$$\mathbf{E}_\theta M^*(a) \geq \exp(I(\lambda,\theta)a), \tag{3.30}$$

and as $a\to\infty$

$$\overline{\mathbf{E}}_\lambda M^*(a) \leq a + (C+o(1))\sqrt{a}, \tag{3.31}$$

where C is as defined in (3.11). Moreover, if $\{N(a)\}$ is a family of stopping times such that (3.30) holds for some θ, then

$$\overline{\mathbf{E}}_\lambda N(a) \geq a + O(1), \quad \text{as} \quad a\to\infty. \tag{3.32}$$

Proof. By Theorem 3.8 and Lemma 3.4, for all $\theta\in[\theta_0,\theta_1]$, we have

$$\mathbf{E}_\theta M^*(a) \geq \frac{1}{\mathbf{P}_\theta(M(a)<\infty)} \geq \exp(I(\lambda,\theta)a),$$

which establishes (3.30). Theorem 3.8 and equation (3.28) imply that

$$\overline{\mathbf{E}}_\lambda M^*(a) \leq \mathbf{E}_\lambda(M(a)) = a + (C+o(1))\sqrt{a}.$$

Relation (3.32) follows from the following proposition, which improves Lorden's lower bound in (3.2). □

Proposition 3.10. *Given θ and $\lambda\neq\theta$, there exists an $M=M(\theta,\lambda)>0$ such that* *for any stopping time N*

$$\log \mathbf{E}_\theta N \le I(\lambda,\theta)\overline{\mathbf{E}}_\lambda N + M. \tag{3.33}$$

Proof. By equation (2.53) on page 26 of Siegmund [33], there exist C_1 and C_2 such that for Page's CUSUM procedure T_a defined in (2.5),

$$\mathbf{E}_\theta T_a \le C_1 e^a, \quad \text{and} \quad I(\lambda,\theta)\overline{\mathbf{E}}_\lambda T_a \ge a - C_2,$$

for all $a > 0$. Let $a = \log \mathbf{E}_\theta N - \log C_1$. Then

$$\mathbf{E}_\theta N = C_1 e^a \ge \mathbf{E}_\theta T_a.$$

The optimality property of T_a (Moustakides [21]) implies that

$$I(\lambda,\theta)\log \overline{\mathbf{E}}_\lambda N \quad \ge \quad I(\lambda,\theta)\log \overline{\mathbf{E}}_\lambda T_a \ge a - C_2 = \log \mathbf{E}_\theta N - \log C_1 - C_2.$$

□

The following corollary follows at once from Theorem 3.9.

Corollary 3.11. *Suppose $\{N(a)\}$ is a family of stopping times such that*

$$\overline{\mathbf{E}}_\lambda N(a) \le \overline{\mathbf{E}}_\lambda M^*(a).$$

Then for all $\theta_0 \le \theta \le \theta_1$ and $a > 0$

$$\frac{\log \mathbf{E}_\theta M^*(a)}{I(\lambda,\theta)} \ge a,$$

and as $a \to \infty$,

$$\frac{\log \mathbf{E}_\theta N(a)}{I(\lambda,\theta)} \le a + (C + o(1))\sqrt{a},$$

where C is as defined in (3.11).

For computer implementation, by (3.17),

$$\begin{aligned} M^*(a) &= \inf\Big\{n \geq 1 : \max_{n-b\leq k\leq n} \sum_{i=k}^{n} \log \frac{f_\lambda(X_i)}{f_{\theta_0}(X_i)} \geq I(\lambda,\theta_0)a, \text{ or} \\ &\quad \max_{1\leq k\leq n-b-1} \sum_{i=k}^{n} \log \frac{f_\lambda(X_i)}{f_{\theta_1}(X_i)} \geq I(\lambda,\theta_1)a\Big\}, \end{aligned}$$

where $b = [a]$. Let $W_k = \max\{W_{k-1}, 0\} + \log(f_\lambda(X_k)/f_{\theta_1}(X_k))$ and $W_0 = 0$. Then $M^*(a)$ can be written in the following convenient form

$$\begin{aligned} M^*(a) &= \inf\Big\{n \geq 1 : \max_{n-b\leq k\leq n} \sum_{i=k}^{n} \log \frac{f_\lambda(X_i)}{f_{\theta_0}(X_i)} \geq I(\lambda,\theta_0)a, \text{ or} \\ &\quad W_{n-b-1} + \sum_{i=n-b}^{n} \log \frac{f_\lambda(X_i)}{f_{\theta_1}(X_i)} \geq I(\lambda,\theta_1)a\Big\}. \end{aligned} \tag{3.34}$$

Since W_k can be calculated recursively, this form reduces the memory requirements at every stage n from the full data set $\{X_1, \ldots, X_n\}$ to the data set of size $b+2$, i.e., $\{X_{n-b-1}, X_{n-b}, \ldots, X_n\}$. It is easy to see that this form involves only $O(a)$ computations at every stage n.

3.1.3 Extension to Half-Open Interval

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with probability density f_ξ of the form (3.4) and suppose we are interested in testing the null hypothesis

$$H_0 : \xi \in \Theta = (\underline{\xi}, \theta_1]$$

against the alternative hypothesis

$$H_1 : \xi \in \Lambda = \{\lambda\},$$

where $\theta_1 < \lambda$. Recall that $\Omega = (\underline{\xi}, \bar{\xi})$ is the natural parameter space of ξ. Assume

$$\lim_{\theta\to\underline{\xi}} \mathbf{E}_\theta X = -\infty. \tag{3.35}$$

This condition is equivalent to $\lim_{\theta\to\underline{\xi}} b'(\theta) = -\infty$ since $b'(\theta) = \mathbf{E}_\theta X$. Many distributions satisfy this condition. For example, (3.35) holds for the normal distributions since $\mathbf{E}_\theta X = \theta$ and $\underline{\xi} = -\infty$. It also holds for the negative exponential density since $b(\theta) = -\log\theta,\ \underline{\xi} = 0$ and $\mathbf{E}_\theta X = b'(\theta) = -1/\theta$.

As in (3.8), our proposed open-ended test $M(a)$ of $H_0 : \xi \in \Theta = (\underline{\xi}, \theta_1]$ against $H_1 : \xi = \lambda$ is defined by

$$\hat{M}(a) = \inf\Big\{n \geq 1 : \inf_{\underline{\xi}<\theta\leq\theta_1} \big(\sum_{i=1}^{n} \log\frac{f_\lambda(X_i)}{f_\theta(X_i)} - I(\lambda,\theta)a\big) \geq 0\Big\}.$$

As in (3.16), $\hat{M}(a)$ can be written as

$$\hat{M}(a) = \inf\Big\{n \geq 1 : \sum_{i=1}^{n} X_i \geq b'(\lambda)a + \sup_{\underline{\xi}<\theta\leq\theta_1} \Big[(n-a)\phi(\theta)\Big]\Big\}, \tag{3.36}$$

where $\phi(\theta)$ is defined in (3.14). By L'Hôptial's rule and the condition in (3.35),

$$\lim_{\theta\to\underline{\xi}} \phi(\theta) = \lim_{\theta\to\underline{\xi}} \frac{b(\lambda)-b(\theta)}{\lambda-\theta} = \lim_{\theta\to\underline{\xi}} b'(\theta) = \lim_{\theta\to\underline{\xi}} \mathbf{E}_\theta X = -\infty.$$

Thus for any $n < a$, $\sum_{i=1}^{n} X_i$ is finite but $\sup_{\underline{\xi}<\theta\leq\theta_1}\big[(n-a)\phi(\theta)\big] = \infty$. So $\hat{M}(a)$ will never stop at time $n < a$. Recall that $\phi(\theta)$ is an increasing function of θ, hence the supremum in (3.36) is attained at $\theta = \theta_1$ if $n \geq a$. Therefore,

$$\begin{aligned}\hat{M}(a) &= \inf\big\{n \geq a : \sum_{i=1}^{n} X_i \geq b'(\lambda)a + (n-a)\phi(\theta_1)\big\} \\ &= \inf\big\{n \geq a : \sum_{i=1}^{n} \log\frac{f_\lambda(X_i)}{f_{\theta_1}(X_i)} \geq I(\lambda,\theta_1)a\big\}.\end{aligned} \tag{3.37}$$

Using arguments similar to the proofs of (3.28) and Lemma 3.4, we have

Lemma 3.12. *For $a > 0$,*

$$\mathbf{P}_\theta(\hat{M}(a) < \infty) \leq \exp(-I(\lambda,\theta)a), \tag{3.38}$$

for any $\theta \in (\underline{\xi}, \theta_1]$. *Moreover, as* $a \to \infty$,

$$\mathbf{E}_\lambda \hat{M}(a) = a + (\hat{C} + o(1))\sqrt{a}, \tag{3.39}$$

and

$$\hat{C} = \frac{\lambda - \theta_1}{I(\lambda, \theta_1)} \cdot \sqrt{\frac{b''(\lambda)}{2\pi}}, \tag{3.40}$$

For the problem of detecting a change in distribution from some f_θ with $\theta \in \Theta = (\underline{\xi}, \theta_1]$ to f_λ, define $\hat{M}^*(a)$ from $\hat{M}(a)$ as before, so that

$$\hat{M}^*(a) = \inf \Big\{ n \geq a : \max_{1 \leq k \leq n-a} \sum_{i=k}^{n} \log \frac{f_\lambda(X_i)}{f_{\theta_1}(X_i)} \geq I(\lambda, \theta_1) a \Big\}. \tag{3.41}$$

By Theorem 3.8 and Lemma 3.12, we have

Lemma 3.13. *For* $a > 0$ *and* $\theta \in (\underline{\xi}, \theta_1]$,

$$\mathbf{E}_\theta \hat{M}^*(a) \geq \exp(I(\lambda, \theta) a), \tag{3.42}$$

and as $a \to \infty$

$$\overline{\mathbf{E}}_\lambda \hat{M}^*(a) \leq a + (\hat{C} + o(1))\sqrt{a}, \tag{3.43}$$

where $\hat{C}$ *is defined in (3.40).*

Thus the analogs of Theorems 3.3 and 3.9 hold.

3.1.4 Numerical Examples

In this subsection we describe the results of a Monte Carlo experiment designed to check the insights obtained from the asymptotic theory of Sections 3.1.2 and 3.1.3.

The simulations considered the problem of detecting a change in distribution from f_θ to g_λ, where $f_\theta = N(\theta, 1)$ with $\theta \in \Theta = [-1, -0.5]$, and $g_\lambda = N(\lambda, 1)$ with

Table 3.1: $\mathbf{E}_\theta N$ for different procedures
(The best possible values are obtained from optimal envelope of Page's CUSUM procedures)

θ	best possible	$M^*(a)$ ($a = 18.50$)	$T(-0.5, a)$ ($a = 2.92$)	$T(-1.0, a)$ ($a = 9.88$)
-0.5	233 ± 7	206 ± 6	233 ± 7	125 ± 3
-0.6	523 ± 15	501 ± 15	518 ± 15	297 ± 8
-0.7	$1,384 \pm 43$	$1,324 \pm 43$	$1,227 \pm 37$	938 ± 29
-0.8	$5,157 \pm 165$	$4,688 \pm 148$	$3,580 \pm 113$	$4,148 \pm 129$
-0.9	$22,942 \pm 699$	$19,217 \pm 606$	$10,613 \pm 343$	$21,617 \pm 658$
-1.0	$118,223 \pm 3,711$	$83,619 \pm 2,566$	$31,641 \pm 1,036$	$118,223 \pm 3,711$

$\lambda \in \Lambda = \{0\}$. In this case our procedure $M^*(a)$ reduces to the form

$$\begin{aligned} M^*(a) &= \inf\Big\{n \geq 1 : \max_{1 \leq k \leq n} \inf_{\theta \in [-1,-0.5]} \Big[\sum_{i=k}^{n} (X_i - \frac{\theta}{2}) + \frac{\theta}{2} a\Big] > 0\Big\} \\ &= \inf\Big\{n \geq 1 : \max_{n-[a] \leq k \leq n} \sum_{i=k}^{n} (X_i + \frac{1}{2}) \geq \frac{1}{2} a, \quad \text{or} \\ & \qquad \max_{1 \leq k \leq n-[a]-1} \sum_{i=k}^{n} (X_i + \frac{1}{4}) \geq \frac{1}{4} a\Big\}. \end{aligned}$$

Table 3.1 compares our procedure $M^*(a)$ and two versions of Page's CUSUM procedure $T(\theta_0, a)$ over a range of θ values. Here

$$\begin{aligned} T(\theta_0, a) &= \inf\{n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log \frac{g_\lambda(X_i)}{f_{\theta_0}(X_i)} \geq a\} \\ &= \inf\{n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} (-\theta_0)[X_i - \frac{\theta_0}{2}] \geq a\}. \end{aligned}$$

The threshold value a for Page's CUSUM procedure $T(\theta_0, a)$ and our procedure $M^*(a)$ is determined from the criterion $\overline{\mathbf{E}}_\lambda N \approx 20$. A 10,000-repetition Monte Carlo simulation was then performed to determine the appropriate values of a to yield the desired detection delay to within the range of sampling error. Moreover, with the thresholds used, the detection delay $\overline{\mathbf{E}}_\lambda N$ is close enough to 20 so that the difference is negligible, i.e., correcting the threshold to get exactly 20 (if we knew how to do that) would change $\mathbf{E}_\theta N$ by an amount that would make little difference in light of the simulation errors $\mathbf{E}_\theta N$ already has.

Table 3.1 also reports the best possible $\mathbf{E}_\theta N$ at each of the values of θ subject to $\overline{\mathbf{E}}_\lambda N \approx 20$. Note that they are obtained from an optimal envelope of Page's CUSUM procedures, and therefore not possible in practice. Each result in Table 3.1 is based on 1000 simulations and is recorded as the Monte Carlo estimate $\pm$ standard error.

Table 3.1 shows that $M^*(a)$ performs well over a broad range of θ, which is consistent with the asymptotic theory of $M^*(a)$ developed in Sections 3.1.2 and 3.1.3 showing that $M^*(a)$ attains (up to $O(\sqrt{a})$) the asymptotic upper bounds for $\log \mathbf{E}_\theta N$ in Corollary 3.11 as $a \to \infty$.

3.2 Composite Post-Change Hypotheses

Let Θ and Λ be two compact disjoint subsets of some Euclidean space. Let $\{f_\theta; \theta \in \Theta\}$ and $\{g_\lambda; \lambda \in \Lambda\}$ be two sets of densities, absolutely continuous with respect to the same non-degenerate σ-finite measure. In this section, we are interested in detecting a change in distribution from f_θ for some $\theta \in \Theta$ to g_λ for some $\lambda \in \Lambda$. Here we no longer assume the densities belong to exponential families, and we assume that both Θ and Λ are composite. We require that Θ is compact.

Ideally we would like a stopping time N which minimizes the detection delay $\overline{\mathbf{E}}_\lambda N$ for all $\lambda \in \Lambda$ and maximizes $\mathbf{E}_\theta N$ for all $\theta \in \Theta$, i.e., seek a family $\{N(a)\}$ which is asymptotically efficient for all $(\theta, \lambda) \in \Theta \times \Lambda$. However, in general such a family does not exist. For example, for $\Lambda = \{\lambda_1, \lambda_2\}$, it is easy to see from (3.3) that there exists a family that is asymptotically efficient at both (θ, λ_1) and (θ, λ_2) for all $\theta \in \Theta$ only if

$$\frac{I(\lambda_2, \theta)}{I(\lambda_1, \theta)} \quad \text{is constant in } \theta \in \Theta.$$

This fails in general when Θ is composite. For example, if f_θ and g_λ belong to a one-parameter exponential family and Θ is an interval, a simple argument shows that

$I(\lambda_2, \theta)/I(\lambda_1, \theta)$ is a constant if and only if $\lambda_1 = \lambda_2$.

It is natural to consider the following definition:

Definition 3.14. *A family of stopping times $\{N(a)\}$ is asymptotically optimal to first-order if*

(i) for each $\theta \in \Theta$, there exists at least one $\lambda_\theta \in \Lambda$ such that the family is asymptotically efficient at (θ, λ_θ); and

(ii) for each $\lambda \in \Lambda$, there exists at least one $\theta_\lambda \in \Theta$ such that the family is asymptotically efficient at $(\theta_\lambda, \lambda)$.

It turns out that such asymptotically optimal procedures can be obtained from the problem of finding a stopping time N such that $\mathbf{E}_\theta N$ is as large as possible for all $\theta \in \Theta$, subject to a constraint of the form

$$\sup_{\lambda \in \Lambda} \left(q_0(\lambda) \overline{\mathbf{E}}_\lambda N \right) \leq a, \tag{3.44}$$

where $a > 0$ is given. Here $q_0(\lambda) > 0$ can be thought of as the cost per observation of delay if the post-change observations have distribution g_λ. Our Theorem 3.15 uses a weaker, asymptotic version of (3.44).

Our proposed procedures $T^*(a)$ and $T_1^*(a)$ are defined as follows. First, define

$$p(\theta) = \inf_{\lambda \in \Lambda} \frac{I(\lambda, \theta)}{q_0(\lambda)}. \tag{3.45}$$

Next, let η be an a priori distribution fully supported on Λ. Define an open-ended test $T(a)$ by

$$T(a) = \inf \left\{ n \geq 1 : \inf_{\theta \in \Theta} \left[\frac{1}{p(\theta)} \log \frac{\int_\Lambda [g_\lambda(X_1) \cdots g_\lambda(X_n)] \, \eta(d\lambda)}{f_\theta(X_1) \cdots f_\theta(X_n)} \right] \geq a \right\}, \tag{3.46}$$

Finally, define

$$T^*(a) = \min_{k \geq 1} (T_k(a) + k - 1),$$

where $T_k(a)$ is obtained by applying $T(a)$ to $X_k, X_{k+1}, \ldots$. Equivalently,

$$T^*(a) = \inf\Big\{n \geq 1 : \max_{1\leq k\leq n} \inf_{\theta\in\Theta} \Big[\frac{1}{p(\theta)} \log \frac{\int_\Lambda [g_\lambda(X_k)\cdots g_\lambda(X_n)]\,\eta(d\lambda)}{f_\theta(X_k)\cdots f_\theta(X_n)}\Big] \geq a\Big\}. \quad (3.47)$$

We also define a slightly different procedure

$$T_1^*(a) = \inf\Big\{n \geq 1 : \inf_{\theta\in\Theta} \max_{1\leq k\leq n} \Big[\frac{1}{p(\theta)} \log \frac{\int_\Lambda [g_\lambda(X_k)\cdots g_\lambda(X_n)]\,\eta(d\lambda)}{f_\theta(X_k)\cdots f_\theta(X_n)}\Big] \geq a\Big\}. \quad (3.48)$$

The next theorem and its corollaries, whose proofs are given in Section 3.2.1, show that $T^*(a)$ and $T_1^*(a)$ are asymptotically optimal to first-order.

Theorem 3.15. *Assume that A1 - A3 below hold.*

(i) If Λ is compact and $\{N(a)\}$ is a family of stopping times such that (3.44) holds. Then for all $\theta \in \Theta$

$$\log \mathbf{E}_\theta N(a) \leq (1 + o(1))p(\theta)a \quad (3.49)$$

as $a \to \infty$.

(ii) For all $a > 0$ and $\theta \in \Theta$,

$$\mathbf{E}_\theta T^*(a) \geq \exp(p(\theta)a). \quad (3.50)$$

If Θ is compact, then

$$\overline{\mathbf{E}}_\lambda T^*(a) \leq (1 + o(1))\frac{a}{q(\lambda)} \quad \text{as } a \to \infty \quad (3.51)$$

for all $\lambda \in \Lambda$, where

$$q(\lambda) = \inf_{\theta\in\Theta} \frac{I(\lambda,\theta)}{p(\theta)}, \quad (3.52)$$

and $q(\lambda) \geq q_0(\lambda)$.

(iii) Relations (3.50) and (3.51) still hold if $T^*(a)$ is replaced by $T_1^*(a)$.

Corollary 3.16. *Assume that A1-A3 hold and Θ and Λ are compact. Then both $T^*(a)$ and $T_1^*(a)$ are asymptotically optimal to first-order.*

Corollary 3.17. *Under the assumptions of Corollary 3.16, if $\{N(a)\}$ is a family of procedures such that*

$$\limsup_{a\to\infty} \frac{\overline{\mathbf{E}}_\lambda N(a)}{\overline{\mathbf{E}}_\lambda T^*(a)} \le 1 \quad \textit{for all } \lambda \in \Lambda,$$

then

$$\limsup_{a\to\infty} \frac{\log \mathbf{E}_\theta N(a)}{\log \mathbf{E}_\theta T^*(a)} \le 1 \quad \textit{for all } \theta \in \Theta.$$

Similarly, if

$$\liminf_{a\to\infty} \frac{\log \mathbf{E}_\theta N(a)}{\log \mathbf{E}_\theta T^*(a)} \ge 1 \quad \textit{for all } \theta \in \Theta,$$

then

$$\liminf_{a\to\infty} \frac{\overline{\mathbf{E}}_\lambda N(a)}{\overline{\mathbf{E}}_\lambda T^*(a)} \ge 1 \quad \textit{for all } \lambda \in \Lambda.$$

The same assertions are true if $T^*(a)$ is replaced by $T_1^*(a)$.

Remark: Corollary 3.17 shows that our procedures $T^*(a)$ and $T_1^*(a)$ are also asymptotically optimal in the following sense: If a family of procedures $\{N(a)\}$ performs asymptotically as well as our procedures (or better) uniformly over Θ, then our procedures perform asymptotically as well as $\{N(a)\}$ (or better) uniformly over Λ, and the same is true if the roles of Θ and Λ are reversed.

Throughout this section, we impose the following assumptions on the densities f_θ and g_λ as well as on $p(\theta)$.

A1. $p(\theta)$ is continuous and $p_0 = \inf_{\theta\in\Theta} p(\theta) > 0$.

A2. The Kullback-Leibler information numbers $I(\lambda,\theta) = \mathbf{E}_\lambda \log(g_\lambda(X)/f_\theta(X))$ are finite. Furthermore,

- **(a)** $I_0 = \inf_\lambda \inf_\theta I(\lambda,\theta) > 0$
- **(b)** $I(\lambda,\theta)$ and $I(\lambda) = \inf_\theta I(\lambda,\theta)$ are both continuous in λ.

A3. For all θ, λ

- **(a)** $\mathbf{E}_\lambda[\log(g_\lambda(X)/f_\theta(X))]^2 < \infty$
- **(b)** $\lim_{\rho\to 0} \mathbf{E}_\lambda[\log \sup_{|\theta'-\theta|\le\rho} \ f_{\theta'}(X) - \log f_\theta(X)]^2 = 0$
- **(c)** $\lim_{\lambda'\to\lambda} \mathbf{E}_\lambda[\log g_{\lambda'}(X) - \log g_\lambda(X)]^2 = 0.$

Assumptions A2 and A3 are part of the assumptions 2 and 3 in Kiefer and Sacks [9]. Assumption A2 guarantees that Θ and Λ are "separated."

3.2.1 Asymptotic Optimality

First we establish the lower bound (3.50) on the mean times between false alarms for our procedures $T^*(a)$ and $T_1^*(a)$.

Lemma 3.18. *For all $a > 0$ and $\theta \in \Theta$, $T^*(a)$ and $T_1^*(a)$ satisfy (3.50).*

Proof. Define

$$t(\theta, a) = \inf\left\{n \ge 1 : \frac{1}{p(\theta)} \log \frac{\int_\Lambda [g_\lambda(X_1)\cdots g_\lambda(X_n)]\,\eta(d\lambda)}{f_\theta(X_1)\cdots f_\theta(X_n)} \ge a\right\}$$

and

$$t^*(\theta, a) = \min_{k\ge 1}(t_k(\theta, a) + k - 1), \tag{3.53}$$

where $t_k(\theta, a)$ is obtained by applying $t(\theta, a)$ to $X_k, X_{k+1}, \ldots$. Then it is clear that $T^*(a) \ge T_1^*(a) \ge t^*(\theta, a)$, and hence

$$\mathbf{E}_\theta T^*(a) \ge \mathbf{E}_\theta T_1^*(a) \ge \mathbf{E}_\theta\Big[t^*(\theta, a)\Big]$$

Using Theorem 3.8, we have

$$\mathbf{E}_\theta\Big[t^*(\theta, a)\Big] \ge \frac{1}{\mathbf{P}_\theta(t(\theta, a) < \infty)},$$

and it suffices to show that

$$\mathbf{P}_\theta(T(a) < \infty) \le \exp(-p(\theta)a).$$

Define

$$L_n = \frac{\int_\Lambda [g_\lambda(X_1)\cdots g_\lambda(X_n)]\,\eta(d\lambda)}{f_\theta(X_1)\cdots f_\theta(X_n)}.$$

For any stopping time N,

$$\begin{aligned}
\mathbf{P}_\theta(N < \infty) &= \sum_1^\infty \mathbf{P}_\theta\{N = n\} \\
&= \sum_1^\infty \int_{\{N=n\}} \Big[f_\theta(X_1)\cdots f_\theta(X_n)\Big] d\xi_1\cdots d\xi_n \\
&= \sum_1^\infty \int_{\{N=n\}} \Big[L_n^{-1}\int_\Lambda [g_\lambda(X_1)\cdots g_\lambda(X_n)]\,\eta(d\lambda)\Big] d\xi_1\cdots d\xi_n \\
&= \int_\Lambda \Big(\sum_1^\infty \int_{\{N=n\}} \Big[L_n^{-1}\, g_\lambda(X_1)\cdots g_\lambda(X_n)\Big] d\xi_1\cdots d\xi_n\Big)\,\eta(d\lambda) \\
&= \int_\Lambda \mathbf{E}_\lambda(L_N^{-1}; N < \infty)\,\eta(d\lambda).
\end{aligned}$$

By definition, $L_{t(\theta,a)} \ge \exp(p(\theta)a)$. Thus,

$$\begin{aligned}
\mathbf{P}_\theta(t(\theta,a) < \infty) &\le \int_\Lambda \mathbf{E}_\lambda\big(\exp(-p(\theta)a);\; t(\theta,a) < \infty\big)\,\eta(d\lambda) \\
&= \exp(-p(\theta)a)\int_\Lambda \mathbf{P}_\lambda\big(t(\theta,a) < \infty\big)\,\eta(d\lambda) \\
&\le \exp(-p(\theta)a),
\end{aligned}$$

which proves the lemma. □

Next we derive an upper bound on the detection delay for our procedures.

Lemma 3.19. *Suppose that A1-A3 hold and Θ is compact. Then (3.51) holds for all $\lambda \in \Lambda$.*

Proof. By definition, $\overline{\mathbf{E}}_\lambda T_a^*(a) \le \overline{\mathbf{E}}_\lambda T^*(a) \le \mathbf{E}_\lambda T(a)$, so it suffices to show that

$$\mathbf{E}_\lambda T(a) \le (1+o(1))\frac{a}{q(\lambda)}$$

for any $\lambda \in \Lambda$, where $q(\lambda)$ is defined in (3.52). We'll use the method in Kiefer and Sacks [9] to prove this inequality. Fix $\lambda_0 \in \Lambda$. Let $\{\epsilon_a; a > 0\}$ be a set of positive numbers with $\epsilon_a \to 0$ as $a \to \infty$. By assumption A1 - A3 and the compactness of Θ, there exist a finite covering $\{U_i, 1 \le i \le k_a\}$ of Θ (with $\theta_i \in U_i$) and positive numbers δ_a such that for all $\lambda \in V_a = \{\lambda \mid |\lambda - \lambda_0| < \delta_a\}$, and $i = 1, \cdots, k_a$,

$$\mathbf{E}_{\lambda_0}\big[\log g_\lambda(X) - \log \sup_{\theta \in U_i} f_\theta(X)\big] \ge I(\lambda_0, \theta_i) - \epsilon_a, \tag{3.54}$$

and

$$\sup_{\theta \in U_i} |p(\theta) - p(\theta_i)| < \epsilon_a.$$

Let $N_1(a)$ be the smallest n such that

$$\log \int_{V_a} \Big[g_\lambda(X_1)\cdots g_\lambda(X_n)\Big]\eta(d\lambda) \ge \sup_{\theta\in\Theta}\Big[p(\theta)a + \sum_{j=1}^n \log f_\theta(X_j)\Big]. \tag{3.55}$$

Clearly $N_1(a) \ge T(a)$. By Jensen's inequality, the left-hand side of (3.55) is greater than or equal to

$$\begin{aligned} &\int_{V_a} \log\Big[g_\lambda(X_1)\cdots g_\lambda(X_n)\Big]\frac{\eta(d\lambda)}{\eta(V_a)} + \log \eta(V_a) \\ =\ & \sum_{j=1}^n \int_{V_a} \log g_\lambda(X_j)\frac{\eta(d\lambda)}{\eta(V_a)} - |\log \eta(V_a)| \end{aligned} \tag{3.56}$$

since $\eta(V_a) \le 1$. Since $\{U_i\}$ covers Θ, the right-hand side of (3.55) is less than or equal to

$$\sup_{1\le i\le k_a} \sup_{\theta\in U_i} \Big[p(\theta)a + \sum_{j=1}^n \log f_\theta(X_j)\Big]$$

$$\leq \sup_{1\leq i\leq k_a} \Big[(p(\theta_i)+\epsilon_a)a + \sup_{\theta\in U_i}\sum_{j=1}^n \log f_\theta(X_j)\Big]$$
$$\leq \sup_{1\leq i\leq k_a} \Big[(p(\theta_i)+\epsilon_a)a + \sum_{j=1}^n \log \sup_{\theta\in U_i} f_\theta(X_j)\Big]. \tag{3.57}$$

For $j = 1, 2, \cdots$, put

$$Y_j = \int_{V_a} \log g_\lambda(X_j)\frac{\eta(d\lambda)}{\eta(V_a)}, \qquad \text{and} \quad Z_j^i \;=\; \log \sup_{\theta\in U_i} f_\theta(X_j) \;\text{ for }\; i = 1, \cdots, k_a.$$

Let $N_2(a)$ be the smallest n such that

$$\sum_{j=1}^n Y_j - \max_{1\leq i\leq k_a}\Big[\sum_{j=1}^n Z_j^i + \big(p(\theta_i)+\epsilon_a\big)a\Big] \geq |\log \eta(V_a)|,$$

or, equivalently, the smallest n such that for all $1 \leq i \leq k_a$,

$$\sum_{j=1}^n \frac{Y_j - Z_j^i}{p(\theta_i)} \geq a\Big[1 + \frac{\epsilon}{p(\theta_i)} + \frac{|\log \eta(V_a)|}{ap(\theta_i)}\Big].$$

Using (3.56) and (3.57), it is clear that $N_2(a) \geq N_1(a)$. Recall that $p_0 = \inf_{\theta\in\Theta} p(\theta)$, and define

$$\tau_a = \frac{\epsilon_a}{p_0} + \frac{|\log \eta(V_a)|}{ap_0}.$$

Let $N_3(a)$ be the smallest n such that

$$\min_{1\leq i\leq k_a}\sum_{j=1}^n \frac{Y_j - Z_j^i}{p(\theta_i)} \geq a(1+\tau_a),$$

or equivalently

$$\sum_{j=1}^n \Big[\frac{Y_j - Z_j^1}{p(\theta_1)} - \epsilon_a\Big] + \min_{1\leq i\leq k_a}\sum_{j=1}^n \Big[\frac{Y_j - Z_j^i}{p(\theta_i)} - \frac{Y_j - Z_j^1}{p(\theta_1)} + \epsilon_a\Big] \geq a(1+\tau_a).$$

Clearly $N_3(a) \geq N_2(a)$. From (3.54), we have

$$\mathbf{E}_{\lambda_0}\Big[\frac{Y_j - Z_j^i}{p(\theta_i)} - \epsilon_a\Big] \geq \frac{I(\lambda_0,\theta_i)}{p(\theta_i)} - \epsilon_a(1+\frac{1}{p_0}) \quad \text{for } i = 1, \ldots, k_a. \tag{3.58}$$

Suppose that $\{U_i\}$ are indexed so that the minimum (over i) of the left-hand side of (3.58) occurs when $i = 1$. For $n = 1, 2, \ldots,$ define

$$S_n = \sum_{j=1}^{n} \Big[\frac{Y_j - Z_j^1}{p(\theta_1)} - \epsilon_a \Big],$$

and

$$B_n^i = \sum_{j=1}^{n} \Big[\frac{Y_j - Z_j^1}{p(\theta_1)} - \frac{Y_j - Z_j^i}{p(\theta_i)} + \epsilon_a \Big] \quad \text{for } i = 1, \ldots, k_a.$$

Let $N^*(a)$ be the smallest n such that, simultaneously,

$$S_n > a(1 + \tau_a), \qquad \text{and} \qquad \min_{1 \le i \le k_a} B_n^i \ge 0.$$

Clearly, $N^*(a) \ge N_3(a)$. Now it suffices to show that as $a \to \infty$

$$\mathbf{E}_{\lambda_0} N^*(a) \le (1 + o(1)) \frac{a}{q(\lambda_0)}. \tag{3.59}$$

The proof of (3.59) relies mainly on two facts. First, the *last* time $\min_{1 \le i \le k_a} B_n^i < 0$ has finite expectation under $\mathbf{P}_{\lambda_0}$ because the summands in B_n^i have positive ($\ge \epsilon_a$) mean and finite variance. Second, by (3.52) and (3.58),

$$\mathbf{E}_{\lambda_0} \Big[\frac{Y_j - Z_j^1}{p(\theta_1)} - \epsilon_a \Big] \ge q(\lambda_0) - \epsilon_a (1 + \frac{1}{p_0}).$$

The detail of the proof is omitted as it is identical to the proof of Lemma 2 in Kiefer and Sacks [9]. □

Proof of Theorem 3.15: To prove (i), use Lorden's lower bound (see (3.2)) to obtain

$$\log \mathbf{E}_\theta N(a) \le \inf_{\lambda \in \Lambda} \Big((1 + o(1)) I(\lambda, \theta) \overline{\mathbf{E}}_\lambda N(a) \Big) \le \inf_{\lambda \in \Lambda} \Big((1 + o(1)) I(\lambda, \theta) \frac{a}{q_0(\lambda)} \Big),$$

so that (3.49) follows from the compactness of Λ.

To prove (ii) and (iii), use Lemmas 3.18 and 3.19, so that it suffices to show that $q(\lambda) \geq q_0(\lambda)$ for all $\lambda \in \Lambda$. Fix $\lambda_0 \in \Lambda$. By (3.45), we have $p(\theta) \leq I(\lambda_0, \theta)/q_0(\lambda_0)$ and hence $I(\lambda_0, \theta)/p(\theta) \geq q_0(\lambda_0)$ for any $\theta \in \Theta$. Thus $q(\lambda_0) \geq q_0(\lambda_0)$ by the definition in (3.52). This complete the proof of the theorem. □

Proof of Corollaries 3.16 and 3.17: First note that

$$p(\theta) = \inf_{\lambda \in \Lambda} \frac{I(\lambda, \theta)}{q(\lambda)}, \tag{3.60}$$

where $p(\theta)$ and $q(\lambda)$ are defined in (3.45) and (3.52), respectively.

To prove (3.60), fix $\theta_0 \in \Theta$. On the one hand, recall that $q(\lambda) \geq q_0(\lambda)$ for all $\lambda \in \Lambda$, so that

$$\inf_{\lambda \in \Lambda} \frac{I(\lambda, \theta_0)}{q(\lambda)} \leq \inf_{\lambda \in \Lambda} \frac{I(\lambda, \theta_0)}{q_0(\lambda)} = p(\theta_0),$$

by the definition of $p(\theta)$ in (3.45).

On the other hand, (3.52) implies that $q(\lambda) \leq I(\lambda, \theta_0)/p(\theta_0)$ for all $\lambda \in \Lambda$, so $p(\theta_0) \leq I(\lambda, \theta_0)/q(\lambda)$ for all $\lambda \in \Lambda$. Thus

$$p(\theta_0) \leq \inf_{\lambda \in \Lambda} \frac{I(\lambda, \theta_0)}{q_0(\lambda)},$$

and hence (3.60) holds.

To prove Corollary 3.16, note that the asymptotic efficiency of $T^*(a)$ and $T_1^*(a)$ at (θ, λ) is

$$e(\theta, \lambda) = \frac{p(\theta) q(\lambda)}{I(\lambda, \theta)},$$

and so they are asymptotically optimal to first-order by virtue of the compactness of Θ and Λ and relations (3.52) and (3.60).

Applying Lorden's lower bound, we can prove Corollary 3.17 in the same way as part (i) of Theorem 3.15. □

Remark: Instead of $T(a)$ in (3.46), we can also define the following stopping time in open-ended hypothesis testing problems:

$$\hat{T}(a) = \inf\left\{n \geq 1 : \inf_{\theta\in\Theta}\left[\frac{1}{p(\theta)}\log\frac{\sup_\lambda[g_\lambda(X_1)\cdots g_\lambda(X_n)]}{f_\theta(X_1)\cdots f_\theta(X_n)}\right] \geq a\right\}, \tag{3.61}$$

and then use it to construct the corresponding procedures in change-point problems. When f_θ and g_λ are from the same exponential family, we can obtain an upper bound on $\mathbf{P}_\theta(\hat{T}(a) < \infty)$ by equation (13) on page 636 in Lorden [17], and so we get a lower bound on the mean time between false alarms. The upper bound on detection delay follows from Lemma 3.19 and the fact that $\hat{T}(a) \leq T(a)$. These procedures are therefore also asymptotically optimal to first-order if f_θ and g_λ belong to exponential families.

3.2.2 $p(\theta)$ and $q(\lambda)$

Here are some examples of choices of $p(\theta)$ and $q(\lambda)$ and the corresponding procedures provided by our theorem.

Example 1: If there exists I_0 such that for all $\theta \in \Theta$, $\inf_{\lambda\in\Lambda} I(\lambda,\theta) = I_0$, then $q_0(\lambda) \equiv I_0$ yields

$$p(\theta) = 1, \quad \text{and} \quad q(\lambda) = \inf_{\theta\in\Theta} I(\lambda,\theta).$$

This is even true for composite Θ and Λ. In particular, if Θ is simple, say $\{\theta_0\}$, then the considerations of Section 3.2 reduce to the standard formulation where the pre-change distribution is completely specified. Moreover, Pollak [23] proved that $T(a)$, defined in (3.46), has a *second-order* optimality property in the context of open-ended hypothesis testing if f_θ and g_λ belong to exponential families.

Example 2: If there exists I_0 such that for all $\lambda \in \Lambda$, $\inf_{\theta\in\Theta} I(\lambda,\theta) = I_0$, then $q_0(\lambda) \equiv 1$ yields

$$p(\theta) = \inf_{\lambda\in\Lambda} I(\lambda,\theta), \quad \text{and} \quad q(\lambda) = 1,$$

even for composite Θ and Λ. In particular, if Λ is simple, say $\{\lambda\}$, then the considerations of Section 3.2 reduce to those of the problem in Section 3.1.

Example 3: Suppose f_θ and g_λ are exponentially distributed with unknown means $1/\theta$ and $1/\lambda$, respectively. Assume $\Theta = \{\theta : \theta \in [\theta_0, \theta_1]\}$, and $\Lambda = \{\lambda : \lambda \in [\lambda_0, \lambda_1]\}$, where $\theta_0 < \theta_1 < \lambda_0 < \lambda_1$. Then the pairs $(p(\theta), q(\lambda))$ defined in (3.45) and (3.52) are not unique. For example, the following two pairs are non-equivalent:

$$\begin{cases} p_1(\theta) = I(\lambda_0, \theta) \\ q_1(\lambda) = I(\lambda, \theta_0)/I(\lambda_0, \theta_0) \end{cases} \quad \text{and} \quad \begin{cases} p_2(\theta) = I(\lambda_1, \theta)I(\lambda_0, \theta_1)/I(\lambda_1, \theta_1) \\ q_2(\lambda) = I(\lambda, \theta_1)/I(\lambda_0, \theta_1) \end{cases}$$

Suppose $t_1^*(a)$ and $t_2^*(a)$ are the procedures defined by (3.47) for the pairs $(p_1(\theta), q_1(\lambda))$ and $(p_2(\theta), q_2(\lambda))$, respectively. Even though both $t_1^*(a)$ and $t_2^*(a)$ are asymptotically optimal to first-order, it is easy to see that $t_1^*(a)$ has larger mean times between false alarms uniformly over Θ while $t_2^*(a)$ has smaller detection delays uniformly over Λ.

In our theorems, we require that Θ and Λ are compact. If they are not compact, then our procedures *may* or *may not* be asymptotically optimal. However, we can still sometimes apply our ideas in these situations, as shown in the following example.

Example 4: Suppose we want to detect a change from negative to positive in the mean of independent normally distributed random variables with variance 1.

In the context of open-ended hypothesis testing, we want to test

$$H_0 : \theta \in \Theta = (-\infty, 0] \quad \text{against} \quad H_1 : \lambda \in \Lambda = (0, \infty).$$

Let us examine the procedures $\hat{T}(a)$ defined in (3.61) for different choices of $q_0(\lambda)$.

(I) $q_0(\lambda) = 1$ leads to

$$p(\theta) = \frac{\theta^2}{2}, \quad \text{and} \quad q(\lambda) = 1,$$

and

$$\hat{t}_0(a) = \inf\{n \geq a : S_n \geq 0\}, \qquad \text{where} \quad S_n = \sum_{i=1}^{n} X_i.$$

Hence we use the following stopping time to detect a change in mean from negative to positive:

$$\hat{t}_0^*(a) = \inf\Big\{n \geq a : \max_{0\leq k\leq n-a} \big(S_n - S_k\big) \geq 0\Big\}.$$

Note that the maximum is taken over $0 \leq k \leq n - a$.

(II) Assume $q_0(\lambda) = \lambda^{1/\beta}$ with $\beta \geq 1/2$. Then we have

$$p(\theta) = k_\beta|\theta|^{2-(1/\beta)} \quad \text{and} \quad q(\lambda) = \lambda^{1/\beta}, \quad \text{with} \quad k_\beta = 2\beta^2(2\beta-1)^{(1/\beta)-2},$$

(we assume $0^0 = 1$) and thus

$$\hat{t}_\beta(a) = \inf\Big\{n \geq 1 : S_n \geq a^\beta n^{1-\beta}\Big\}.$$

This suggests using a stopping time of the form

$$\hat{t}_\beta^*(a) = \inf\Big\{n \geq 1 : \max_{0\leq k\leq n} \big[(S_n - S_k)(n-k)^{\beta-1}\big] \geq a^\beta\Big\}$$

to detect a change in mean from negative to positive. Observe that for $\beta = 1$, $\hat{t}_\beta(a)$ is just the one-sided SPRT and $\hat{t}_\beta^*(a)$ is just a special form of Page's CUSUM procedures. For $\beta = 1/2$, $\hat{t}_\beta(a)$ and $\hat{t}_\beta^*(a)$ have also been studied extensively in the literature, since they are based on the generalized likelihood ratio. Different motivation to obtain these two procedures can be found for $\hat{t}_\beta(a)$ in Chapter IV of Siegmund [33], which is from the viewpoint of the repeated significant test, and for $\hat{t}_\beta^*(a)$ in Siegmund and Venkatraman [34], which is from the viewpoint of the generalized likelihood ratio. For $\hat{t}_\beta(a)$ with $0 < \beta \leq 1$, see Chow, Hsiung and Lai [5] and equation (9.2) on page 188 in Siegmund [33].

Though one cannot use our theorems directly to analyze the properties of $\hat{t}_0^*(a)$ and $\hat{t}_\beta^*(a)$, I conjecture that they are indeed asymptotically optimal to first-order. I believe that a proof can be based on an asymptotic expression for the mean time between false alarms and the detection delay. I plan to study the asymptotic behavior of these procedures in future research.

3.2.3 Numerical Examples

In this subsection we report some simulation studies comparing the performance of our procedures in Section 3.2.1 with a commonly used procedure in the literature.

The simulations considered the problem of detecting a change in distribution from f_θ to g_λ, where f_θ and g_λ are exponentially distributed with unknown means $1/\theta$ and $1/\lambda$, respectively, and $\theta \in \Theta = [0.8, 1]$ and $\lambda \in \Lambda = [2, 3]$.

By (3.45), $q_0(\lambda) \equiv 1$ leads to $p(\theta) = I(2, \theta)$ where $I(\lambda, \theta) = \frac{\theta}{\lambda} - 1 - \log \frac{\theta}{\lambda}$, and so our procedure based on (3.61) is defined by

$$\hat{T}^*(a) = \inf \Big\{ n \geq 1 : \max_{1 \leq k \leq n} \inf_{0.8 \leq \theta \leq 1} \sup_{2 \leq \lambda \leq 3} \frac{1}{p(\theta)} \sum_{i=k}^{n} \Big(\log \frac{\lambda}{\theta} - (\lambda - \theta) X_i \Big) \geq a \Big\}.$$

A commonly used procedure in the literature is so-called *generalized likelihood ratio procedure*, see Lorden [16], Siegmund and Venkatraman [34]. The procedure requires specification of the nominal value θ_0 (of the parameter of the pre-change distribution), and it is defined by the stopping time

$$\begin{aligned} \tau(\theta_0, a) &= \inf\{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{\lambda \in \Lambda} \sum_{i=k}^{n} \log \frac{g_\lambda(X_i)}{f_{\theta_0}(X_i)} \geq a \} \\ &= \inf\{ n \geq 1 : \max_{1 \leq k \leq n} \sup_{2 \leq \lambda \leq 3} \sum_{i=k}^{n} \big(\log \frac{\lambda}{\theta_0} - (\lambda - \theta_0) X_i \big) \geq a \}. \end{aligned}$$

Note that $\tau(\theta_0, a)$ can be thought of as our procedure $\hat{T}^*(a)$ whose Θ contains the single point θ_0. The choice of θ_0 can be made by considering the pre-change distribution which is *closest* to the post-change distributions because it is always more difficult to detect a smaller change. For our example, $\theta_0 = 1$.

An effective method to implement $\tau(\theta_0, a)$ numerically can be found in Lorden [16]. Similarly, we can perform $\hat{T}^*(a)$ as follows. Compute V_n and W_n recursively by $V_n = \max(V_{n-1} + \log \frac{2}{0.8} - (2 - 0.8) X_n, 0)$ and $W_n = \max(W_{n-1} + \log 2 - X_n, 0)$. Stop whenever $W_n \geq p(0.8) a / (2 - 0.8)$. In addition, whenever $V_n = 0$ one can begin a new cycle, discarding all previous observations and starting fresh on the incoming observations. Moreover, each time a new cycle begins compute at each stage $n = 1, 2, \ldots,$

$$Q_k^{(n)} = X_n + \cdots + X_{n-k+1}, \quad k = 1, \ldots, n.$$

If $p(1)a/(3-1) \leq W_n \leq p(0.8)a/1.2$, we will also stop at the first n such that $Q_k^{(n)} < c_k$ for some k, where

$$c_k = \inf_{0.8 \leq \theta \leq 1} \sup_{2 \leq \lambda \leq 3} \Big[k \frac{\log \lambda - \log \theta}{\lambda - \theta} - \frac{p(\theta) a}{\lambda - \theta} \Big].$$

Table 3.2 provides a comparison of the performances for our procedure $\hat{T}^*(a)$ with $\tau(\theta_0, a)$. The threshold a for each of these two procedures is determined from the criterion $\mathbf{E}_{\theta=1} N(a) \approx 600$. The results in Table 3.2 are based on 1000 simulations for $\mathbf{E}_\theta N$ and 10000 simulations for $\mathbf{E}_\lambda N$. Note that for these two procedures, the detection delay $\overline{\mathbf{E}}_\lambda N = \mathbf{E}_\lambda N$. Table 3.2 shows that at a small additional cost of detection delay, $\hat{T}^*(a)$ can significantly improve the mean times between false alarms compared to $\tau(1, a)$. This is consistent with the asymptotic theory in Sections 3.2.1.

Table 3.2: Comparison of two procedures in change-point problems with composite pre-change and composite post-change hypotheses

		$\hat{T}^*(a)$	$\tau(1, a)$
a		22.50	5.02
$\mathbf{E}_\theta N$	$\theta = 1$	601 ± 18	606 ± 19
	$\theta = 0.9$	$1,448 \pm 43$	$1,207 \pm 36$
	$\theta = 0.8$	$3,772 \pm 116$	$2,749 \pm 90$
$\mathbf{E}_\lambda N$	$\lambda = 2$	21.41 ± 0.10	21.92 ± 0.11
	$\lambda = 2.2$	18.09 ± 0.07	18.18 ± 0.09
	$\lambda = 2.5$	15.08 ± 0.05	14.76 ± 0.06
	$\lambda = 2.7$	13.75 ± 0.04	13.22 ± 0.05
	$\lambda = 3$	12.29 ± 0.04	11.62 ± 0.04

Chapter 4

Sequential Decentralized Decision Systems

In this chapter we discuss sequential hypothesis testing and change-point problems in the context of decentralized decision systems. See Figure 1.1 for the general setting. Suppose there are L sensors in a system. At time n, an observation X_n^l is made at each sensor S_l. Based on the information available at S_l at time n, a message U_n^l is chosen from a finite list or *alphabet* and is sent to a fusion center. The fusion center uses the stream of messages from the sensors as inputs to a sequential hypothesis test or change-point detection procedure. Without loss of generality, we assume that U_n^l takes a value in $\{0, 1, \ldots, D^l - 1\}$.

In Veeravalli [38] and Veeravalli, Basar, and Poor [40], the authors considered five different cases, depending on how "feedback" and "local information" is used at the sensors.

Case **A**) *System with Neither Feedback from the Fusion Center nor Local Memory:*

$$U_n^l = \phi_n^l(X_n^l).$$

Case **B**) *System with no Feedback and Full Local Memory:*

$$U_n^l = \phi_n^l(X_{[1,n]}^l), \quad \text{where} \quad X_{[1,n]}^l = (X_1^l, \ldots, X_n^l).$$

Case **C**) *System with no Feedback and Local Memory Restricted to Past Decisions:*

$$U_n^l = \phi_n^l(X_n^l, U_{[1,n-1]}^L), \quad \text{where} \quad U_{[1,n-1]}^L = (U_1^l, \ldots, U_{n-1}^l).$$

Case **D**) *System with Full Feedback and Full Local Memory:*

$$U_n^l = \phi_n^l(X_{[1,n]}^l; U_{[1,n-1]}^l, \ldots, U_{[1,n-1]}^L).$$

Case **E**) *System with Full Feedback, but Local Memory Restricted to Past Decisions:*

$$U_n^l = \phi_n^l(X_n^l; U_{[1,n-1]}^l, \ldots, U_{[1,n-1]}^L).$$

This chapter develops an asymptotic theory of decentralized change-point problems, giving in all cases procedures that are asymptotically optimal and easy to implement. For decentralized change-point problems, we assume that at time ν the change affects all sensor simultaneously. Like previous authors, we consider only the case of simple hypotheses, both pre-change and post-change. The problem is to minimize detection delay in cases **A** - **E** over all possible sensor message functions and fusion center decision procedures, subject to a constraint on the mean time between false alarms. Crow and Schwartz [6] and Tartakovsky and Veeravalli [36] studied this problem for case **A**, but both imposed restrictions on the sensors' message functions. Our theory involves no such restrictions.

Section 4.1 provides an asymptotic theory of decentralized sequential hypothesis testing problems. We develop bounds on the expected sample sizes of hypothesis tests and find asymptotically optimal procedures. In Section 4.2 we provide asymptotic theory and asymptotically optimal procedures in change-point problems. Section 4.3 establishes a sufficient condition for our theorems to be applied. Section 4.4 gives simulation results for examples in both open-ended hypothesis testing problems and change-point problems.

Throughout this chapter, we make the following assumptions, which are standard:

(A1) The sensor observations are independent over time as well as from sensor to

sensor.

(A2) The densities of the sensor observations are either $f^1, \dots, f^L$ or $g^1, \dots, g^L$, where the f's and g's are given. For each l, the Kullback-Leibler information number

$$I(g^l, f^l) = \int \log\Big(\frac{g^l(x)}{f^l(x)}\Big) g^l(x) dx \tag{4.1}$$

is finite and positive, and

$$\int \Big(\log \frac{g^l(x)}{f^l(x)}\Big)^2 g^l(x) dx < \infty. \tag{4.2}$$

We now introduce some notations. Let D be a positive integer. Consider a random variable Y whose density function is either f or g with respect to a measure μ, and assume that the Kullback-Leibler information number $I(g, f)$ is finite. For a (*deterministic* or *random*) measurable function ϕ from the range of Y to a finite alphabet of size D, say $\{0, 1, \dots, D-1\}$, denote by f_ϕ and g_ϕ respectively the density of $\phi(Y)$ when the density of Y is f or g. Let

$$Z_\phi = \log \frac{g_\phi(\phi(Y))}{f_\phi(\phi(Y))},$$

and define

$$I_D(g, f) = \sup_\phi \mathbf{E}_g Z_\phi \tag{4.3}$$

and

$$V_D(g, f) = \sup_\phi \mathbf{E}_g (Z_\phi)^2. \tag{4.4}$$

It is well known (Tsitsiklis [37]) that $I_D(g, f) \le I(g, f)$, i.e., that reduction of the datat from Y to $\phi(Y)$ cannot increase the information. Tsitsiklis [37] showed that the supremum $I_D(g, f)$ is achieved by a Monotone Likelihood Ratio Quantizer (MLRQ)

φ of the form

$$\varphi(Y) = d \quad \text{if and only if} \quad \lambda_d \leq \frac{g(Y)}{f(Y)} < \lambda_{d+1}, \tag{4.5}$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{D-1} \leq \lambda_D = \infty$ are constants. These optimal MLRQ's are not easy calculated, but we follow the standard practice in the literature of developing procedures that assume sensor messages are constructed optimally in the sensor. Some of our theorems assume that $V_D(g, f) < \infty$. A sufficient condition for finiteness of $V_2(g, f)$ is given in Section 4.3.

Using these notations, define the information numbers

$$I_{\mathbf{D}} = \sum_{l=1}^{L} I_{D^l}(g^l, f^l), \tag{4.6}$$

where $\mathbf{D} = (D^1, D^2, \cdots, D^L)$, and

$$I_{tot} = \sum_{l=1}^{L} I(g^l, f^l). \tag{4.7}$$

These two information numbers are key to our theorems.

4.1 Sequential Hypothesis Testing

In this section we consider the problem of decentralized sequential hypothesis testing. As stated in *(A1)* and *(A2)*, there are two possible probability measures, $\mathbf{P}_0$ and $\mathbf{P}_1$. Under $\mathbf{P}_0$, the observations at sensor S_l, $X_1^l, X_2^l, \cdots$, are independent and identically distributed with density function f^l, and under $\mathbf{P}_1$, they have density g^l. The problem is to test the simple null hypothesis

$$H_0 : \quad \mathbf{P}_0 \text{ is true}$$

against the simple alternative hypothesis

$$H_1 : \quad \mathbf{P}_1 \text{ is true.}$$

A sequential procedure consists of a rule to determine the sensor messages, a stopping time τ used by the fusion center and a final decision rule that chooses $\mathbf{P}_0$ or $\mathbf{P}_1$ based on the information up to time τ at the fusion center. We first establish information bounds and then find asymptotically optimal procedures for open-ended hypothesis testing problems in all cases and for two-decision hypothesis testing problems in cases **B** and **D**.

4.1.1 Information Bounds for Open-Ended Tests

Motivated by applications to change-point problems, we first consider "open-ended," or "one-sided" tests of the null hypothesis H_0, i.e., tests defined by a stopping time τ that stop sampling only to reject H_0 and must satisfy

$$\mathbf{P}_0(\tau < \infty) \leq \alpha. \tag{4.8}$$

Lemma 4.1. *For any stopping time τ satisfying (4.8),*

(i) In cases **A**, **C** *and* **E**, *if $V_{D^l}(g^l, f^l)$, defined in (4.4), is finite for all $1 \leq l \leq L$, then we have*

$$\mathbf{E}_1\tau \geq (1 + o(1))\frac{|\log \alpha|}{I_{\mathbf{D}}}, \tag{4.9}$$

where $I_{\mathbf{D}}$ is defined in (4.6).

(ii) In cases **B** *and* **D**, *we have*

$$\mathbf{E}_1\tau \geq \frac{|\log \alpha|}{I_{tot}}, \tag{4.10}$$

where I_{tot} is defined in (4.7).

Proof. Using Wald's equation and Jensen's inequality, it is straightforward to prove (4.10). We next prove (4.9) in cases **A, C** and **E**. Note that we can write

$$U_n^l = \psi_n^l(X_n^l),$$

where ψ_n^l are allowed to depend on $U^1_{[1,n-1]}, \dots, U^L_{[1,n-1]}$. Denote by $f^l_{\psi,n}$ and $g^l_{\psi,n}$ respectively the conditional density induced on U_n^l given $U^1_{[1,n-1]}, \dots, U^L_{[1,n-1]}$ when the density of X_n^l is f^l or g^l. Denote by Z_n^l the conditional log-likelihood ratio function of U_n^l, $\log\left(g^l_{\psi,n}(U_n^l)/f^l_{\psi,n}(U_n^l)\right)$.

Since $X_n^1, \dots, X_n^L$ are independent, so are $U_n^1, \dots, U_n^L$ given $U^1_{[1,n-1]}, \dots, U^L_{[1,n-1]}$. Thus in the fusion center the conditional log-likelihood ratio of $(U_n^1, \dots, U_n^L)$ given $U^1_{[1,n-1]}, \dots, U^L_{[1,n-1]}$ is

$$\mathbf{Z}_n = \sum_{l=1}^{L} Z_n^l.$$

By Theorem 1 (or Theorem 3) in Lai [12], to prove (4.9) it suffices to show that for any $\delta > 0$,

$$\limsup_{n\to\infty} P_1\Big\{ \max_{t\le n} \sum_{k=1}^{t} \mathbf{Z}_k \ge I_{\mathbf{D}}(1+\delta)n \Big\} = 0. \tag{4.11}$$

Since

$$\mathbf{E}_1 \mathbf{Z}_k = \sum_{l=1}^{L} \mathbf{E}_1 Z_k^l \le \sum_{l=1}^{L} I_{D^l}(g^l, f^l) = I_{\mathbf{D}},$$

the left-hand side of (4.11) is less than or equal to

$$\begin{aligned} &\limsup_{n\to\infty} \mathbf{P}_1\Big\{ \max_{t\le n} \sum_{k=1}^{t}\sum_{l=1}^{L} \left(Z_k^l - \mathbf{E}_1 Z_k^l\right) \ge I_{\mathbf{D}}\delta n\Big\} \\ \le &\sum_{l=1}^{L} \limsup_{n\to\infty} \mathbf{P}_1\Big\{ \max_{t\le n} \sum_{k=1}^{t} \left(Z_k^l - \mathbf{E}_1 Z_k^l\right) \ge \delta_1 n\Big\}, \end{aligned}$$

where $\delta_1 = I_{\mathbf{D}}\delta/L$.

Note that $\sum_{k=1}^{n}\left(Z_k^l - \mathbf{E}_1 Z_k^l\right)$ is a martingale. Doob's submartingale inequality

(Theorem 14.6 of Williams [42]) implies that

$$\mathbf{P}_1\Big\{\max_{t\le n}\sum_{k=1}^{t}\big(Z_k^l - E_1 Z_k^l\big)\ge \delta_1 n\Big\}\le \frac{\sum_{k=1}^n \mathbf{E}_1(Z_k^l)^2}{\delta_1^2 n^2}.$$

Note that for any k, $\mathbf{E}_1(Z_k^l)^2 \le V_{D^l}(g^l, f^l)$ by definition, and hence

$$\mathbf{P}_1\Big\{\max_{t\le n}\sum_{k=1}^{t}\Big(Z_k^l - E_1 Z_k^l\Big)\ge \delta_1 n\Big\}\le \frac{V_{D^l}(g^l, f^l)}{\delta_1^2 n},$$

which implies (4.11) since $V_{D^l}(g^l, f^l)$ is finite. Relation (4.9) follows. □

4.1.2 Asymptotically Optimal Open-Ended Tests

We now propose asymptotically optimal procedures satisfying (4.8) for all cases in open-ended hypothesis testing problems.

In cases **A, C** and **E**, we propose the following open-ended tests, denoted by $M(a)$:

Each sensor uses the optimal monotone likelihood ratio quantizer φ^l. Namely,

$$U_n^l = \varphi^l(X_n^l) = d \quad \text{if and only if } \lambda_d^l \le \frac{g^l(X_n^l)}{f^l(X_n^l)} < \lambda_{d+1}^l,$$

where $0 = \lambda_0^l \le \lambda_1^l \le \cdots \le \lambda_{D-1}^l \le \lambda_D^l = \infty$ are optimally chosen in the sense that the Kullback-Leibler information number $I(g_\varphi^l, f_\varphi^l)$ achieves the supremum $I_{D^l}(g^l, f^l)$. Here f_φ^l and g_φ^l are the densities induced on U_n^l when the observations X_n^l are distributed as f^l and g^l, respectively.

Based on the independent, identically distributed observations $\mathbf{U}_n = (U_n^1, \ldots, U_n^L)$, the fusion center then uses the one-sided sequential probability ratio test (SPRT) with log-likelihood ratio boundary a, i.e., our stopping time $M(a)$ is given by

$$M(a) = \inf\Big\{n \ge 1 : \sum_{k=1}^{n}\Big(\sum_{l=1}^{L}\log\frac{g_\varphi^l(U_k^l)}{f_\varphi^l(U_k^l)}\Big)\ge a\Big\}. \tag{4.12}$$

Lemma 4.2. *For $\alpha < 1$ let $a = |\log \alpha|$, then $M(a)$ satisfies (4.8) and*

$$\mathbf{E}_1 M(a) = \frac{a}{I_{\mathbf{D}}} + O(1), \tag{4.13}$$

as $a \to \infty$, where $I_{\mathbf{D}}$ is defined in (4.6).

Remark: This lemma and part (i) of Lemma 4.1 establish the asymptotic optimality of our procedures $M(a)$ in open-ended hypothesis testing problems in cases **A**, **C** and **E**.

Proof. Applying standard asymptotic theory for one-sided SPRTs to the sensor messages $\mathbf{U}_n = (U_n^1, \ldots, U_n^L)$, the proof of (4.13) is straightforward. □

In cases **B** and **D**, our proposed procedure $T(a)$ in the open-ended testing problem is as follows:

For each sensor S_l, one considers whether or not the log-likelihood ratio of f^l versus g^l exceeds the boundary $\pi_l a$, where

$$\pi_l = \frac{I(g^l, f^l)}{\sum_{l=1}^{L} I(g^l, f^l)} = \frac{I(g^l, f^l)}{I_{tot}}. \tag{4.14}$$

That is, for $l = 1, 2, \ldots, L$ and $n = 1, 2, \ldots,$ letting

$$S_n^l = \sum_{i=1}^{n} \log \frac{g^l(X_i^l)}{f^l(X_i^l)}, \tag{4.15}$$

define the sensor messages

$$U_n^l = 1\{S_n^l \geq \pi_l a\},$$

where $1\{A\}$ is the indicator of the event A. The fusion center will stop and reject $\mathbf{P}_0$ if $U_n^l = 1$ for all $l = 1, 2, \ldots, L$, i.e., our stopping time $T(a)$ is

$$T(a) = \inf \Big\{ n \geq 1 : S_n^l \geq \pi_l a \text{ for all } \ l = 1, 2, \ldots, L \Big\}. \tag{4.16}$$

Lemma 4.3. *For $\alpha < 1$ let $a = |\log \alpha|$, then $T(a)$ satisfies (4.8) and*

$$\mathbf{E}_1 T(a) = \frac{a}{I_{tot}} + O(\sqrt{a}), \tag{4.17}$$

as $a \to \infty$, where I_{tot} is defined in (4.7).

Remark: This lemma and part (ii) of Lemma 4.1 establish the asymptotic optimality of our procedures $T(a)$ in open-ended hypothesis testing problems in cases **B** and **D**. Moreover, it is interesting to note that our policy $T(a)$ only uses binary sensor messages, but it is asymptotically optimal for general $D^l (\geq 2)$.

Proof. For any stopping time τ, using Wald's likelihood ratio identity, we have

$$\mathbf{P}_0(\tau < \infty) = \mathbf{E}_1 \exp\big(- \sum_{l=1}^{L} S_{\tau}^l ; \tau < \infty \big).$$

Since $S_{T(a)}^l \geq \pi_l a$ for all l and $\sum_{l=1}^{L} \pi_l = 1$, we have

$$\begin{aligned} \mathbf{P}_0(T(a) < \infty) &\leq \mathbf{E}_1 \exp\big(- \sum_{l=1}^{L} \pi_l a ; T(a) < \infty \big) \\ &= \mathbf{E}_1 (e^{-a}; T(a) < \infty) \\ &\leq e^{-a} = \alpha. \end{aligned}$$

Thus $T(a)$ satisfies (4.8).

To prove (4.17), for $1 \leq l \leq L$, let

$$T_l = \inf \Big\{ n \geq 1 : S_n^l \geq \pi_l a \Big\},$$

and

$$\tau_l(T_l) = \sup \Big\{ n \geq 1 : \sum_{i=T_l+1}^{T_l+n} \log \frac{g^l(X_i^l)}{f^l(X_i^l)} \leq 0 \Big\}.$$

For simplicity, denote $\tau_l = \tau_l(0)$. It is well known (e.g., Theorem D in Kiefer and

Sacks [9]) that for any $1 \le l \le L$,

$$\mathbf{E}_1 \tau_l < \infty \tag{4.18}$$

since $\log \big(g^l(X)/f^l(X)\big)$ has positive mean and finite variance under $\mathbf{P}_1$ by assumption *(A2)*.

By definition of T_l and $\tau_l(T_l)$, we have

$$T(a) \ge \max_{1 \le l \le L} T_l,$$

and

$$T(a) \le \max_{1 \le l \le L} \Big(T_l + \tau_l(T_l)\Big) \le \max_{1 \le l \le L} T_l + \sum_{l=1}^{L} \tau_l(T_l).$$

Now since $X_1^l, X_2^l, \ldots$ are independent and identically distributed under $\mathbf{P}_1$, we have $\mathbf{E}_1 \tau_1(T_l) = \mathbf{E}_1 \tau_1$, and thus

$$\mathbf{E}_1 \max_{1 \le l \le L} T_l \le \mathbf{E}_1 T(a) \le \mathbf{E}_1 \max_{1 \le l \le L} T_l + \sum_{l=1}^{L} \mathbf{E}_1 \tau_l. \tag{4.19}$$

Now T_l can be written in the form

$$T_l = \inf \Big\{ n \ge 1 : \sum_{i=1}^{n} \Big(\frac{1}{I(g^l, f^l)} \log \frac{g^l(X_i^l)}{f^l(X_i^l)} \Big) \ge \frac{a}{I_{tot}} \Big\},$$

and hence Lemma 3.7 leads to

$$\mathbf{E}_1 \big(T_l - \frac{a}{I_{tot}}\big)^+ = O(\sqrt{a}).$$

Thus

$$\begin{aligned} \mathbf{E}_1 \max_{1 \le l \le L} T_l &\le \frac{a}{I_{tot}} + \sum_{l=1}^{L} \mathbf{E}_1 \Big(T_l - \frac{a}{I_{tot}}\Big)^+ \\ &= \frac{a}{I_{tot}} + O(\sqrt{a}). \end{aligned}$$

Relation (4.17) follows at once from (4.18) and (4.19). □

Remark: A heuristic argument indicates that, as $a \to \infty$,

$$\mathbf{E}_1 T(a) = \mathbf{E}_1 \max_{1 \le l \le L} T_l + O(1) = \frac{a}{I_{tot}} + (C + o(1))\sqrt{a}, \tag{4.20}$$

where

$$C = \mathbf{E} \max_{1 \le l \le L} \big(\frac{\sigma_l}{I(g^l, f^l)} Z_l \big) / \sqrt{I_{tot}}$$

and $Z_1, \ldots, Z_L$ are independent standard normal random variables.

4.1.3 Two-Decision Hypothesis Testing

Suppose that we want to decide which of $\mathbf{P}_0$ and $\mathbf{P}_1$ is true. That is, we want to find two-decision sequential tests such that

$$\mathbf{P}_0(\text{Accept } H_1) \le \alpha, \quad \mathbf{P}_1(\text{Accept } H_0) \le \beta. \tag{4.21}$$

Applying the lower bounds in Lemma 4.1, it is easy to find asymptotic lower bounds for the expected sample sizes τ.

Theorem 4.4. *If τ is the sample size of a test satisfying (4.21), then as $\alpha + \beta \to 0$,*

(i) In cases **A**, **C** *and* **E**, *if $V_{D^l}(g^l, f^l)$, defined in (4.4), is finite for all $1 \le l \le L$, then we have*

$$\mathbf{E}_1 \tau \ge (1 - o(1)) \frac{|\log \alpha|}{I_{\mathbf{D}}}, \tag{4.22}$$

where $I_{\mathbf{D}}$ is defined in (4.6). Moreover, if $V_{D^l}(f^l, g^l)$ is also finite for all l, then we have

$$\mathbf{E}_0 \tau \ge (1 - o(1)) \frac{|\log \beta|}{J_{\mathbf{D}}}, \tag{4.23}$$

where $J_{\mathbf{D}} = \sum_{l=1}^{L} I_{D^l}(f^l, g^l)$.

(ii) In cases **B** *and* **D**,

$$\mathbf{E}_1\tau \geq (1-o(1))\frac{|\log\alpha|}{I_{tot}}, \quad \mathbf{E}_0\tau \geq (1-o(1))\frac{|\log\beta|}{J_{tot}}, \tag{4.24}$$

where I_{tot} *is defined in (4.7), and* $J_{tot} = \sum_{l=1}^{L} I(f^l, g^l)$.

Proof. Wald's inequalities (Theorem 2.39 of Siegmund [33]) imply

$$\begin{aligned} I_{tot}\mathbf{E}_1\tau &\geq (1-\beta)\log\frac{1-\beta}{\alpha} + \beta\log\frac{\beta}{1-\alpha}, \\ J_{tot}\mathbf{E}_0\tau &\geq (1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}. \end{aligned} \tag{4.25}$$

Relation (4.24) in part (ii) of the theorem follows at once from these inequalities and the fact that $(1-x)\log(1-x)+x\log x$ attains its minimum value $-\log 2$ when $x=\frac{1}{2}$.

To prove (4.22) in part (i) of the theorem, define a new stopping time τ' by

$$\tau' = \begin{cases} \tau & \text{if } \tau \text{ chooses } \mathbf{P}_1 \\ \tau + M_1 & \text{if } \tau \text{ chooses } \mathbf{P}_0 \end{cases},$$

where M_1 is the stopping time obtained by applying $M(a)$, defined in (4.12) with $a = |\log\alpha|$, to all sensors observations from time τ on, i.e., $X^l_{\tau+1}, X^l_{\tau+2}, \ldots$ for $1 \leq l \leq L$. It is obvious that τ' is also a procedure in cases **A, C** and **E**.

Note that M_1 is independent of τ, and by Lemma 4.2,

$$\mathbf{P}_0\big(M_1 < \infty\big) \leq \alpha, \quad \text{and} \quad \mathbf{E}_1 M_1 = \frac{|\log\alpha|}{I_{\mathbf{D}}} + O(1).$$

Thus

$$\begin{aligned} \mathbf{P}_0\Big(\tau' < \infty\Big) &\leq \mathbf{P}_0\Big(\tau \text{ chooses } \mathbf{P}_1\Big) + \mathbf{P}_0\Big(M_1 < \infty;\ \tau \text{ chooses } \mathbf{P}_0\Big) \\ &\leq \alpha + \mathbf{P}_0\big(M_1 < \infty\big) \leq 2\alpha, \end{aligned}$$

and

$$\begin{aligned}\mathbf{E}_1\tau' &\le \mathbf{E}_1\tau + \mathbf{E}_1\big(M_1;\ \tau \text{ chooses } \mathbf{P}_0\big) \le \mathbf{E}_1\tau + \beta\ \mathbf{E}_1 M_1 \\ &= \mathbf{E}_1\tau + \beta\Big(\frac{|\log\alpha|}{I_{\mathbf{D}}} + O(1)\Big).\end{aligned}$$

Now using the results in Lemma 4.2 for τ', we have

$$\mathbf{E}_1\tau' \ge (1+o(1))\frac{|\log(2\alpha)|}{I_{\mathbf{D}}},$$

and thus

$$\mathbf{E}_1\tau \ \ge\ (1+o(1))\frac{|\log(2\alpha)|}{I_{\mathbf{D}}} - \beta\left(\frac{|\log\alpha|}{I_{\mathbf{D}}} + O(1)\right) = (1-o(1))\frac{|\log\alpha|}{I_{\mathbf{D}}},$$

as $\alpha+\beta\to 0$. Therefore (4.22) holds. The proof of (4.23) is identical. □

It is easy to see that the asymptotic lower bounds in Theorem 4.4 are sharp. For example, the following test achieves the asymptotic lower bound in (4.22) in cases **A, C** and **E**.

Each sensor uses the optimal MLRQ φ^l so that the Kullback-Leibler information number $I(g^l_\varphi, f^l_\varphi)$ achieves the supremum $I_{D^l}(g^l, f^l)$. Based on the independent observations $\mathbf{U}_n = (U^1_n, \ldots, U^L_n)$, the fusion center uses the sequential probability ratio test (SPRT) with log-likelihood ratio boundaries $a < 0 < b$, so that the corresponding error probabilities are at most α and β. In other words, the stopping time of our test is

$$\hat{M}(a,b) = \inf\Big\{n \ge 1 : \sum_{k=1}^{n}\hat{\mathbf{L}}_k \notin (a,b)\Big\}, \quad \text{where} \quad \hat{\mathbf{L}}_k = \sum_{l=1}^{L}\log\frac{g^l_\varphi(U^l_k)}{f^l_\varphi(U^l_k)}. \qquad (4.26)$$

We stop sampling at time $\hat{M}(a,b)$, and if $\hat{M}(a,b) < \infty$

$$\begin{cases}\text{decide } \ H_0 \ \text{ if } \sum_{k=1}^{n}\hat{\mathbf{L}}_k \le a, \\ \text{decide } \ H_1 \ \text{ if } \sum_{k=1}^{n}\hat{\mathbf{L}}_k \ge b.\end{cases}$$

Note that $\hat{\mathbf{L}}_1, \hat{\mathbf{L}}_2, \ldots$ are independent and identically distributed with

$$\mathbf{E}_1 \hat{\mathbf{L}}_k = \sum_{l=1}^{L} I_{D^l}(g^l, f^l) = I_{tot}, \quad \text{and} \quad \mathbf{E}_0 \hat{\mathbf{L}}_k \leq J_{tot}.$$

Applying well known properties of the SPRT (Section 2.2 of Siegmund [33]), we can choose a and b so that our test satisfies (4.21) and achieves the asymptotic lower bound in (4.22).

In cases **A, C** and **E**, I have not been able to find procedures to achieve the lower bounds in (4.22) and (4.23) simultaneously. I believe that such a procedure does not exist in case **A** in general, but I have not been able to devise a proof.

In cases **B** and **D** when $D^l \geq 3$ for each l, we propose the following asymptotically optimal procedures, using a combination of open-ended tests.

Define π_l as in (4.14), and

$$\rho_l = \frac{I(f^l, g^l)}{\sum_{l=1}^{L} I(f^l, g^l)} = \frac{I(f^l, g^l)}{J_{tot}}.$$

For each sensor S_l,

$$U_n^l = \begin{cases} 0 & \text{if } S_n^l \leq -\rho_l b, \\ 1 & \text{if } S_n^l \geq \pi_l a, \\ 2 & \text{otherwise} \end{cases}, \tag{4.27}$$

where S_n^l is defined in (4.15). Finally, the fusion center will

$$\begin{cases} \text{stop and decide } H_1 & \text{if } U_n^l = 1 \text{ for all } 1 \leq l \leq L, \\ \text{stop and decide } H_0 & \text{if } U_n^l = 0 \text{ for all } 1 \leq l \leq L, \\ \text{continue sampling} & \text{otherwise.} \end{cases} \tag{4.28}$$

Thus our stopping time is $\hat{T}(a, b) = \min\big(T(a), T'(b)\big)$, where $T(a)$ is given by (4.16) and analogously

$$T'(b) = \inf\Big\{n \geq 1 : S_n^l \leq -\rho_l\, b \text{ for all } \ 1 \leq l \leq L\Big\}. \tag{4.29}$$

We decide H_1 if $\hat{T}(a,b) = T(a)$ and decide H_0 if $\hat{T}(a,b) = T'(b)$.

Theorem 4.5. *Assume $\alpha + \beta < 1$ let $a = |\log \alpha|$ and $b = |\log \beta|$. Then*

$$\begin{aligned} &\mathbf{P}_0\big(\hat{T}(a,b) \ \textit{accepts}\ H_1\big) \leq \alpha, \\ &\mathbf{P}_1\big(\hat{T}(a,b) \ \textit{accepts}\ H_0\big) \leq \beta. \end{aligned} \tag{4.30}$$

As $\alpha + \beta \to 0$, we have

$$\mathbf{E}_1 \hat{T}(a,b) \leq \mathbf{E}_1 T(a) \leq (1 + o(1)) a / I_{tot}. \tag{4.31}$$

Similarly, if (4.2) holds with f and g interchanged, then

$$\mathbf{E}_0 \hat{T}(a,b) \leq \mathbf{E}_0 T'(b) \leq (1 + o(1)) b / J_{tot}, \tag{4.32}$$

as $\alpha + \beta \to 0$.

Remark: Based on the lower bound in (4.24), under the conditions of this theorem $\hat{T}(a,b)$ is asymptotically optimal in cases **B** and **D** in the sense of minimizing both $\mathbf{E}_0 \hat{T}(a,b)$ and $\mathbf{E}_1 \hat{T}(a,b)$ asymptotically subject to (4.21).

Proof. Note that $\mathbf{P}_0\big(T'(b) < \infty\big) = 1$. By the definition of $\hat{T}(a,b)$, we have

$$\mathbf{P}_0\big(\hat{T}(a,b) \text{ accepts } H_1\big) \ = \ \mathbf{P}_0\big(T(a) < T'(b)\big) \leq \mathbf{P}_0\big(T(a) < \infty\big).$$

The first inequality in (4.30) follows at once from the property of $T(a)$ established in Lemma 4.3. The second inequality in (4.30) is proved similarly.

Relation (4.31) follows from the definition of $\hat{T}(a,b)$ and (4.17). The proof of (4.32) is identical. □

4.2 Change-Point Problems

Assume that at some unknown time ν, the density function of the sensor observations $\{X_n^l\}$ changes simultaneously for all $1 \le l \le L$ from f^l to g^l. That is, for each $1 \le l \le L$, the observations at sensor S_l, $X_1^l, X_2^l, \ldots,$ are independent random variables such that $X_1^l, \ldots, X_{\nu-1}^l$ are independent and identically distributed with density f^l and $X_\nu^l, X_{\nu+1}^l, \ldots$ are independent and identically distributed with density g^l. Recall that we also assume that the observations are independent from sensor to sensor. Let $\mathbf{P}_\nu$ and $\mathbf{E}_\nu$ denote the probability measure and expectation, when the change occurs at time ν. Let $\mathbf{P}_\infty$ and $\mathbf{E}_\infty$ denote the probability measure and expectation, when there is no change, i.e., $\nu = \infty$.

As in the classical change-point problems, we want to design the sensors' message function ϕ_n^l and seek a stopping time τ at the fusion center that minimizes the "worst case" detection delay, as defined by Lorden [16],

$$\overline{\mathbf{E}}_1(\tau) = \sup_{\nu \ge 1} \Big(\operatorname{ess\,sup} \mathbf{E}_\nu \big[(\tau - \nu + 1)^+ \big| X^1_{[1,\nu-1]}, \ldots, X^L_{[1,\nu-1]} \big] \Big)$$

subject to

$$\mathbf{E}_\infty \tau \ge \gamma. \tag{4.33}$$

The worst-case detection delay $\overline{\mathbf{E}}_1(\tau)$ can be replaced by the "average" detection delay, proposed by Shiryayev [31] and Pollak [24],

$$D(\tau) = \sup_{\nu \ge 1} \mathbf{E}_\nu (\tau - \nu | \tau \ge \nu).$$

Although the worst-case detection delay $\overline{\mathbf{E}}_1(\tau)$ is always greater than the average detection delay $D(\tau)$, they are asymptotically equivalent. Either one can be used in our theorems.

We will study this problems in two different cases: limited local memory (cases **A**, **C** and **E**) and full local memory (cases **B** and **D**).

4.2.1 Limited Local Memory

In cases **A**, **C** and **E**, the following procedure $M^*(a)$ has been studied in the literature:

Each sensor uses the optimal MLRQ, φ^l, that achieves the supremum $I_{D^l}(g^l, f^l)$, just as in the definition of $M(a)$ in the open-ended hypothesis testing problem. Based on the independent observations $\mathbf{U}_n = (U_n^1, \ldots, U_n^L)$, the fusion center uses Page's CUSUM with log-likelihood ratio boundary a to detect whether or not a change has occurred, i.e., the stopping time is given by

$$M^*(a) = \inf\Big\{n \geq 1 : \max_{1\leq k\leq n} \sum_{i=k}^{n} \Big(\sum_{l=1}^{L} \log \frac{g_\varphi^l(U_i^l)}{f_\varphi^l(U_i^l)} \Big) \geq a \Big\}. \tag{4.34}$$

Crow and Schwartz [6] showed that $M^*(a)$ is optimal in the sense that the MLRQ is optimized at each sensor. Later Tartakovsky and Veeravalli [36] proved the asymptotic optimality property of $M^*(a)$ in case **A** under the restriction that the sensor message functions $\{\phi^1, \ldots, \phi^L\}$ satisfy the following condition: For all $\nu = 1, 2, \ldots$, as n goes to ∞, $n^{-1} \sum_{i=\nu}^{\nu+n} \sum_{l=1}^{L} Z_i^l$ converges in probability under $\mathbf{P}_\nu$ to some positive constant I, where $Z_i^l = \log \big(g_\phi^l(U_i^l)/f_\phi^l(U_i^l)\big)$. Veeravalli [39] conjectured that $M^*(a)$ is also asymptotically optimal in case **E** because it has performance similar to the Bayes solutions in case **E**. The following theorem shows that under a condition on second moments, $M^*(a)$ is asymptotically optimal without any restriction on the sensors' message functions or the fusion center decision rule in cases **A**, **C** and **E**.

Theorem 4.6. *Assume $V_{D^l}(g^l, f^l)$, defined in (4.4), is finite for all $1 \leq l \leq L$. If $\{\tau(\gamma)\}$ is a family of procedures in cases* **A**, **C** *and* **E** *satisfying (4.33), then*

$$\overline{\mathbf{E}}_1 \tau(\gamma) \geq (1 + o(1)) \frac{\log \gamma}{I_{\mathbf{D}}}, \tag{4.35}$$

as $\gamma \to \infty$, where $I_{\mathbf{D}}$ is defined in (4.6). For $\gamma > 1$ let $a = \log \gamma$. Then $M^*(a)$ satisfies (4.33) and

$$\overline{\mathbf{E}}_1 M^*(a) \leq \frac{\log \gamma}{I_{\mathbf{D}}} + O(1),$$

so that it asymptotically minimizes the detection delay $\overline{\mathbf{E}}_1 M^*(a)$ *as* $\gamma \to \infty$ *in cases* **A**, **C** *and* **E**.

Proof. To prove (4.35), we use a result of Lai [11], which generalizes Lorden's asymptotic theory. Lai [11] showed that

$$\overline{\mathbf{E}}_1 \tau \geq (1 + o(1)) \frac{\log \mathbf{E}_\infty \tau}{I}$$

as $\mathbf{E}_\infty \tau \to \infty$, provided that for all $\delta > 0$,

$$\lim_{n \to \infty} \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbf{P}^{(\nu)} \Big\{ \max_{t \leq n} \sum_{i=\nu}^{\nu+t} \log \frac{g\big(X_i | X_1, \cdots, X_{i-1}\big)}{f\big(X_i | X_1, \cdots, X_{i-1}\big)} \geq I(1+\delta)n \,\Big|\, X_1, \cdots, X_{\nu-1} \Big\} = 0.$$

Thus it suffices to verify that this condition holds for the fusion center observations $X_i = (U_i^1, \ldots, U_i^L)$ and $I = I_{\mathbf{D}}$. The remainder of the proof is similar to that of part (i) of Lemma 4.1.

Observe that $M^*(a)$, defined in (4.34), is Page's CUSUM procedure, so that by applying the standard bounds,

$$\mathbf{E}_\infty M^*(a) \geq e^a$$

and

$$\overline{\mathbf{E}}_1 M^*(a) \leq \frac{a}{I_{\mathbf{D}}} + O(1),$$

where $I_{\mathbf{D}}$ is defined in (4.6). Thus $M^*(a)$ achieves the lower bound in (4.35), and therefore it is asymptotically optimal in cases **A**, **C** and **E**.

4.2.2 Full Local Memory

It has been an open problem also to find asymptotically optimal procedures (including both the sensor and fusion center decision rules) in cases **B** and **D**. We propose the following asymptotically optimal procedures $T^*(a)$ in these cases.

In cases **B** and **D**,define the CUSUM statistic

$$W_n^l = \max_{1\le k\le n} \sum_{i=k}^{n} \log \frac{g^l(X_i^l)}{f^l(X_i^l)} \tag{4.36}$$

for each $l = 1, \dots, L$, and $n = 1, 2, \dots$. Each sensor S_l sends a local summary message based on whether or not the CUSUM statistic exceeds a constant threshold:

$$U_n^l = \begin{cases} 1 & \text{if } W_n^l \ge \pi_l a \\ 0 & \text{otherwise} \end{cases}, \tag{4.37}$$

where π_l is defined in (4.14). Then the fusion center will stop and declare a change has occurred if and only if $U_n^l = 1$ for all $1 \le l \le L$.

This stopping time $T^*(a)$ can be written as

$$T^*(a) = \inf\Big\{n \ge 1 : W_n^l \ge \pi_l a \quad \text{for all } 1 \le l \le L\Big\}. \tag{4.38}$$

Theorem 4.7. *As $a \to \infty$,*

$$\overline{\mathbf{E}}_1 T^*(a) \le \frac{a}{I_{tot}} + O(\sqrt{a}), \tag{4.39}$$

where I_{tot} is defined in (4.7). Furthermore, if we assume

$$\int g^l(x)\Big|\log\frac{g^l(x)}{f^l(x)}\Big|^3 dx < \infty, \tag{4.40}$$

for $1 \le l \le L$, then as $a \to \infty$,

$$\mathbf{E}_\infty T^*(a) \ge (1 + o(1))e^a, \tag{4.41}$$

Moreover, if $\{\tau(a)\}$ is a family of procedures in cases **B** *and* **D** *such that (4.41) holds,* *then*

$$\overline{\mathbf{E}}_1 \tau(a) \geq \frac{a}{I_{tot}} + O(1), \quad \textit{as } a \to \infty. \tag{4.42}$$

Remarks:

(1) It is essential that each sensor continue sending the local messages to the fusion center even after the CUSUM statistic for that sensor exceeds the local threshold.

(2) For each sensor, the mean time between false alarms is $\exp(\pi_l a)$. By the renewal property of the CUSUM statistics, the mean time between false alarms for the fusion center is of order $\prod_{l=1}^{L} \exp(\pi_l a) = \exp(a)$ since we continue sending local messages. (See the proof below. As in Siegmund and Venkatraman [34], the key idea is Lemma 4.11).

(3) It is interesting to note that at the sensors, we cannot replace the CUSUM statistics by the Shiryayev-Roberts statistics defined in (5.2): in that case the mean time between false alarms in the fusion center is roughly $\exp((\max_{l=1}^{L} \pi_l) a)$, which is much smaller than $\exp(a)$ as $a \to \infty$.

(4) This theorem shows that the procedure $T^*(a)$ minimizes the detection delay up to $O(\sqrt{a})$ among all procedures in cases **B** and **D** such that (4.41) holds.

Proof. The inequality (4.39) follows from the definition of $T^*(a)$ and $T(a)$ and the property of $T(a)$ in (4.17) of Lemma 4.3.

To prove (4.41), let $A = \exp(a)$ and note that

$$\begin{aligned}
\mathbf{E}_\infty N &= \sum_{n=1}^{\infty} \mathbf{P}_\infty(N \geq n) = \sum_{n=1}^{\infty} \int_n^{n+1} \mathbf{P}_\infty(N \geq n) dx \\
&\geq \sum_{n=1}^{\infty} \int_n^{n+1} \mathbf{P}_\infty(N \geq x) d\, x = \int_1^{\infty} \mathbf{P}_\infty(N \geq x) dx \\
&= A \int_{1/A}^{\infty} \mathbf{P}_\infty(N \geq tA) dt.
\end{aligned}$$

Thus by Lemma 4.11 below and Fatou's Lemma,

$$\begin{aligned}
\liminf_{a\to\infty} \big(\mathbf{E}_\infty T^*(a)/A\big) &\geq \liminf_{a\to\infty} \int_0^\infty \mathbf{P}_\infty(T^*(a) \geq tA) 1\{t \geq \frac{1}{A}\} dt \\
&\geq \int_0^\infty \liminf_{a\to\infty} \Big[\mathbf{P}_\infty(T^*(a) \geq tA) 1\{t \geq \frac{1}{A}\}\Big] dt \\
&= \int_0^\infty \exp(-t) dt = 1,
\end{aligned}$$

and hence (4.41) holds. The asymptotic lower bound in (4.42) follows at once from Proposition 3.10 in Chapter 3. □

To complete the proof, we need to prove the following lemmas.

Lemma 4.8. *Let W_n^l be the CUSUM statistic defined in (4.36). For $l = 1, \ldots, L$, $n = 1, 2, \ldots,$ and $b > 0$,*

$$\mathbf{P}_\infty\Big(W_n^l \geq b\Big) \leq \exp(-b).$$

Proof. Let S_n^l denote the log-likelihood ratio $\sum_{i=1}^n \log\big(g^l(X_i^l)/f^l(X_i^l)\big)$, and define $S_0^l = 0$. Then the CUSUM statistic takes the form

$$W_n^l = \max_{0\leq k<n} \Big(S_n^l - S_k^l\Big). \tag{4.43}$$

Since $(X_1^l, \ldots, X_n^l)$ have the same joint distribution as $(X_n^l, \ldots, X_1^l)$, W_n^l has the same distribution as $\max_{1\leq i\leq n} S_i^l$. Thus,

$$\mathbf{P}_\infty\Big(W_n^l \geq b\Big) = \mathbf{P}_\infty\Big(\max_{1\leq i\leq n} S_i^l \geq b\Big) = \mathbf{P}_\infty(N_l(b) \leq n),$$

where

$$N_l(b) = \inf\Big\{n \geq 1 : S_n^l \geq b\Big\}.$$

Lemma 4.8 follows from the fact that

$$\mathbf{P}_\infty(N_l(b) \leq n) \leq \mathbf{P}_\infty(N_l(b) < \infty) \leq \exp(-b).$$

□

Lemma 4.9. *For $k = 1, 2, \ldots,$*

$$\mathbf{P}_\infty(T^*(a) = k) \le \frac{1}{A}.$$

Proof. Note that, since the observations are independent from sensor to sensor, application of lemma 4.8 yields

$$\begin{aligned}\mathbf{P}_\infty(T^*(a) = k) &\le \mathbf{P}_\infty\Big(W_k^l \ge \pi_l a \quad \text{for } 1 \le l \le L\Big)\\ &= \prod_{l=1}^{L} \mathbf{P}_\infty(W_k^l \ge \pi_l a)\\ &\le \prod_{l=1}^{L} \exp(-\pi_l a) = \exp(-a) = \frac{1}{A}.\end{aligned}$$

□

Using Lemma 4.9, it is easy to derive

Lemma 4.10. *For $m = 1, 2, \ldots,$*

$$\mathbf{P}_\infty(T^*(a) \le m) \le \frac{m}{A}.$$

Lemma 4.11. *For $t > 0$,*

$$\limsup_{a\to\infty} \mathbf{P}_\infty(N \le tA) \le 1 - \exp(-t). \tag{4.44}$$

Proof. For simplicity, we consider only the case $L = 2$. The same idea can be applied to the cases $L = 1$ and $L \ge 3$. Choose $m = m(a)$ such that $m/a^2 \to \infty$, and $\log m/a \to 0$. Note that

$$\begin{aligned}\mathbf{P}_\infty(N \le tA) &= \mathbf{P}_\infty\Big(\max_{0\le k<tA/m}\ \max_{km+1\le j\le(k+1)m}\Big[\min_{1\le l\le 2}\frac{W_j^l}{\pi_l}\Big] > a\Big)\\ &= \mathbf{P}_\infty\Big(\max_k \max_j \Big[\min_{1\le l\le 2}\max_{i_l}\frac{S_j^l - S_{i_l}^l}{\pi_l}\Big] > a\Big),\end{aligned} \tag{4.45}$$

where the maximum is taken over $0 \le k < tA/m,\ km+1 \le j \le (k+1)m$ and $1 \le i_l \le j$ for $l = 1, 2$. For all such k, define

$$\begin{aligned} C_1(k) &= \{i_1 : km+1 \le i_1 \le j \le (k+1)m\}, \quad C_2(k) = \{i_1 : 1 \le i_1 \le km\}, \\ D_1(k) &= \{i_2 : km+1 \le i_2 \le j \le (k+1)m\}, \quad D_2(k) = \{i_2 : 1 \le i_2 \le km\}. \end{aligned}$$

For simplicity, omit k — e.g. write C_1 for $C_1(k)$, and define

$$B_1 = C_1 \cap D_1,\ B_2 = C_2 \cap D_1,\ B_3 = C_1 \cap D_2,\ B_4 = C_2 \cap D_2.$$

For $r = 1, 2, 3, 4$, define

$$Q_r = \mathbf{P}_\infty\Big(\max_k \max_j \Big[\min_{1 \le l \le 2} \max_{B_r} \frac{S_j^l - S_{i_l}^l}{\pi_l} \Big] > a \Big),$$

where the maximum is taken over $0 \le k < tA/m,\ km+1 \le j \le (k+1)m$ and $(i_1, i_2) \in B_r$. Note that the right-hand side of (4.45) is less than $\sum_{r=1}^4 Q_r$, and hence it suffices to show that

$$\limsup_{a \to \infty} \sum_{r=1}^{4} Q_r \le 1 - \exp(-t).$$

It is easy to see that

$$Q_1 = 1 - \prod_k \mathbf{P}_\infty\Big(\max_j \Big[\min_{1 \le l \le 2} \max_{i_l} \frac{S_j^l - S_{i_l}^l}{\pi_l} \Big] \le a \Big),$$

where the product is taken over $0 \le k < tA/m$, and the maximum is taken over $km + 1 \le i_l \le j \le (k+1)m$ for $l = 1, 2$. Thus

$$Q_1 = 1 - \Big(\mathbf{P}_\infty(T^*(a) > m) \Big)^{tA/m}.$$

By Lemma 4.10 we have

$$Q_1 \le 1 - \Big(1 - \frac{m}{A}\Big)^{tA/m}.$$

Note that since $m/A \to 0$ as $a \to \infty$, for given $\delta > 0$, once a is sufficiently large,

$$1 - \frac{m}{A} \geq \exp(-(1+\delta)\frac{m}{A}),$$

and thus $Q_1 \leq 1 - \exp(-(1+\delta)t)$. Letting $\delta \to 0$, we obtain

$$\limsup_{a\to\infty} Q_1 \leq 1 - \exp(-t).$$

To complete the proof of Lemma 4.11, it suffices to show that for all $\epsilon > 0$, Q_2, Q_3 and Q_4 are smaller than ϵ for sufficiently large a. We will prove this fact for Q_2 in Lemma 4.12. The proofs for Q_3 and Q_4 are similar. □

Lemma 4.12. *Under the condition (4.40) of Theorem 4.7, for all* $\epsilon > 0$, *once* a *is sufficiently large,*

$$Q_2 = \mathbf{P}_\infty\Big(\max_k \max_j \Big[\min_{1\leq l \leq 2} \max_{i_l} \frac{S_j^l - S_{i_l}^l}{\pi_l} \Big] > a \Big) \leq \epsilon,$$

where the maximum is taken over $0 \leq k < tA/m,\ km+1 \leq j \leq (k+1)m,\ 1 \leq i_1 \leq km$, *and* $km + 1 \leq i_2 \leq j \leq (k+1)m$.

Proof. Note that $j - i_1 = j - km + km - i_1$ and $\{S_j^1 - S_{i_1}^1\}$ equals the sum of the independent random walks $\{S_j^1 - S_{km}^1\}$ and $\{S_{km}^1 - S_{i_1}^1\}$. Hence,

$$\begin{aligned}
Q_2 &\leq \frac{tA}{m} \sum_{j=1}^{m} \mathbf{P}_\infty\Big(\max_{1\leq i \leq tA} \overline{S}_i^1 + S_j^1 > \pi_1 a \text{ and } W_j^2 > \pi_2 a \Big) \\
&\leq \frac{tA}{m} \sum_{j=1}^{m} \mathbf{P}_\infty\Big(\max_{1\leq i \leq tA} \overline{S}_i^1 + S_j^1 > \pi_1 a \Big) \mathbf{P}_\infty\Big(W_j^2 > \pi_2 a \Big) \\
&\leq \frac{t \exp(\pi_1 a)}{m} \sum_{j=1}^{m} \mathbf{P}_\infty\Big(\max_{1\leq i \leq tA} \overline{S}_i^1 + S_j^1 > \pi_1\ a \Big),
\end{aligned}$$

using Lemma 4.8 for W_j^2.

Now using Wald's likelihood ratio identity,

$$
\begin{aligned}
& \mathbf{P}_\infty\Big(\max_{1\le i\le tA} \overline{S}_i^1 + S_j^1 > \pi_1 a\Big) \\
\le\ & \mathbf{P}_\infty(S_j^1 > \pi_1 a) + \mathbf{P}_\infty\Big(\max_{1\le i\le tA} \overline{S}_i^1 > \pi_1 a - S_j^1 > 0\Big) \\
\le\ & \mathbf{P}_\infty(S_j^1 > \pi_1 a) + \mathbf{E}_\infty\Big(\exp(S_j^1 - \pi_1 a); \pi_1 a - S_j^1 > 0\Big) \\
=\ & \mathbf{E}_\infty \exp\Big(\min(0, S_j^1 - \pi_1 a)\Big) \\
=\ & \mathbf{E}_1 \exp\Big(\min(-S_j^1, -\pi_1 a)\Big).
\end{aligned}
$$

Thus,

$$
Q_2 \ \le\ \frac{t}{m}\sum_{j=1}^{m} \mathbf{E}_1 \exp\Big(\min(\pi_1 a - S_j^1, 0)\Big).
$$

Applying Lemma 4.13 (below) for S_j^1 under $\mathbf{P}_1$, and letting $m_1 = a^2$, we have for sufficiently large a

$$
\sup_{j\ge m_1} \mathbf{E}_1 \exp\Big(\min(\pi_1 a - S_j^1, 0)\Big) \le \epsilon_1.
$$

Therefore,

$$
Q_2 \le \frac{t}{m}\big(m_1\cdot 1 + (m-m_1)\epsilon_1\big) \le t\Big(\frac{m_1}{m} + \epsilon_1\Big).
$$

and the lemma follows, since the right-hand side goes to 0 as a goes to ∞. □

Lemma 4.13. *Suppose $X_1, X_2, \ldots$ are independent and identically distributed with $\mathbf{E}X_i = \mu > 0$, $Var(X_i) = \sigma^2$, and $\mathbf{E}|X_i|^3 = \rho < \infty$. Let $S_n = X_1 + \ldots + X_n$ and $m_1 = b^2$. Then*

$$
\sup_{n\ge m_1} \mathbf{E} \exp\Big(\min(b - S_n, 0)\Big) \to 0,
$$

as $b \to \infty$.

Proof. First we establish

$$
\mathbf{E} \exp\Big(\min(b - S_n, 0)\Big) \le \frac{3\rho}{\sigma^3\sqrt{n}} + \Phi\Big(\frac{b - n\mu}{\sigma\sqrt{n}}\Big)
$$

$$+A\Big(\frac{b-n\mu}{\sigma\sqrt{n}}+\sigma\sqrt{n}\Big)\exp\Big(b+(\frac{\sigma^2}{2}-\mu)n\Big), \tag{4.46}$$

where $\Phi(x)$ is the standard normal distribution and $A(x)=1-\Phi(x)=\Phi(-x)$.

Let $F_n(x)$ denote the distribution function of S_n. Then

$$\Big|F_n(x)-\Phi\Big(\frac{x-n\mu}{\sigma\sqrt{n}}\Big)\Big|\le\frac{3\rho}{\sigma^3\sqrt{n}}$$

for all x by the Berry-Esseen Theorem. Now

$$\begin{aligned}
\mathbf{E}\exp\big(\min(b-S_n,0)\big) &= F_n(b)+\int_b^\infty \exp(b-x)dF_n(x)\\
&= \int_b^\infty F_n(x)\exp(b-x)dx\\
&\le \int_b^\infty \Big(\frac{3\rho}{\sigma^3\sqrt{n}}+\Phi\Big(\frac{x-n\mu}{\sigma\sqrt{n}}\Big)\Big)\exp(-x+b)dx\\
&= \frac{3\rho}{\sigma^3\sqrt{n}}+\Phi\Big(\frac{x-n\mu}{\sigma\sqrt{n}}\Big)+\\
&\quad \int_b^\infty \phi(\frac{x-n\mu}{\sigma\sqrt{n}})\exp(-x+b)\frac{1}{\sigma\sqrt{n}}dx,
\end{aligned}$$

and hence (4.46) holds.

We next bound each term on the right-hand side of (4.46). For $n\ge m_1$, the first two terms are uniformly bounded by

$$\frac{3\rho}{\sigma^3 b}+\Phi\Big(\frac{1-\mu b}{\sigma}\Big),$$

which goes to 0 as $b\to\infty$.

For the third term on the right-hand side of (4.46), we need to consider two cases: (1) $\mu>\sigma^2/2$; and (2) $\mu\le\sigma^2/2$. In case (1), note that $A(x)\le 1$, and so for all $n\ge m_1$, the third term is smaller than

$$\exp\Big(b-(\mu-\frac{\sigma^2}{2})b^2\Big),$$

which goes to 0 as $b \to \infty$. In case (2), note that $A(x) \leq \phi(x)/x$ for all $x > 0$, where $\phi(x)$ is the density function of the standard normal distribution (see page 141 of Williams [42]). Thus the third term is smaller than

$$\frac{\sigma\sqrt{n}}{b+(\sigma^2-\mu)n}\phi\Big(\frac{b-\mu n}{\sigma\sqrt{n}}\Big),$$

which also goes to 0 uniformly for all $n \geq m_1$ as $b \to \infty$.

Therefore, Lemma 4.13 holds. □

4.3 Finiteness of $V_2(g, f)$

In Lemma 4.1 and Theorems 4.4 and 4.7, we assume $V_D(g, f) < \infty$, which is difficult to verify in general. In this section, we give some sufficient conditions to verify it when $D = 2$.

Theorem 4.14. *Suppose $f(y)$ and $g(y)$ are two densities such that*

$$\mathbf{E}_g\Big(\log\frac{g(Y)}{f(Y)}\Big)^2 = \int \log\Big(\frac{g(y)}{f(y)}\Big)^2 g(y)dy < \infty.$$

Define

$$A(t) = \mathbf{P}_f\big(g(Y) > tf(Y)\big), \quad B(t) = \mathbf{P}_g\big(g(Y) > tf(Y)\big).$$

Assume that $A(t)$ and $B(t)$ are continuous functions of t on $(0, \infty)$ and take values 0 and 1 for the same t. Moreover, assume that

$$\limsup_{t\to\infty} \sqrt{B(t)}\ \big|\log A(t)\big| < \infty, \tag{4.47}$$

and

$$\limsup_{t\to 0} \sqrt{1-A(t)}\ \big|\log(1-B(t))\big| < \infty, \tag{4.48}$$

where $\sqrt{0}|\log 0|$ *is interpreted as* 0. *Then* $V_2(g,f) < \infty$.

Proof. Assume that $\phi(Y)$ is a measurable function taking values in $\{0,1\}$. Denote by f_ϕ and g_ϕ respectively the density of $\phi(Y)$ when the density of Y is f or g. Let

$$Z_\phi = \log \frac{g_\phi(\phi(Y))}{f_\phi(\phi(Y))}.$$

Note that when $D = 2$,

$$\mathbf{E}_g(Z_\phi)^2 = \beta_\phi\big(\log \frac{\beta_\phi}{\alpha_\phi}\big)^2 + (1-\beta_\phi)\big(\log \frac{1-\beta_\phi}{1-\alpha_\phi}\big)^2,$$

where $\alpha_\phi = \mathbf{P}_f(\phi(Y)=1)$ and $\beta_\phi = \mathbf{P}_g(\phi(Y)=1)$. Define

$$D(r,s) = r\big(\log \frac{r}{s}\big)^2 + (1-r)\big(\log \frac{1-r}{1-s}\big)^2,$$

for $0 < r, s < 1$ and $D(0,0) = D(1,1) = 0$. To prove $V_2(g,f) < \infty$, it suffices to show that there exists a constant M such that for all ϕ,

$$D(\beta_\phi, \alpha_\phi) < M.$$

If one of α_ϕ and β_ϕ is 0 or 1, it is easy to see that Z_ϕ is 0 with probability 1 under g, and hence $D(\beta_\phi, \alpha_\phi) = 0$. So it suffices to consider the case where $0 < \alpha_\phi, \beta_\phi < 1$. Since $D(b,a) = D(1-b, 1-a)$, assume without loss of generality that $0 < \alpha_\phi \leq \beta_\phi < 1$. (Otherwise consider $1-\phi(Y)$ and use (4.48) instead of (4.47)). Since $1 - B(t)$ is a cumulative distribution function and $B(t)$ is continuous, there exists $t_0 \in (0,\infty)$ such that

$$B(t_0) = \beta_\phi.$$

Now define γ^* by

$$\gamma^* = \begin{cases} 1 & \text{if } g(X) > t_0 f(X); \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbf{P}_f(\gamma^* = 1) = A(t_0)$ and $\mathbf{P}_g(\gamma^* = 1) = B(t_0)$.

The proof of the Neyman-Pearson lemma (page 65 of Lehmann [14]) shows that

$$\int (\gamma^* - \gamma)\big(g(y) - t_0 f(y)\big) d\mu \geq 0,$$

so that

$$\big(B(t) - \beta_\phi\big) - t_0\big(A(t) - \alpha_\phi\big) \geq 0.$$

Since $B(t_0) = \beta_\phi$ by our choice of t_0, we have

$$A(t_0) \leq \alpha_\phi.$$

Note that for fixed r,

$$\frac{\partial D(r,s)}{\partial s} = 2\Big[\frac{1-r}{1-s}\log\frac{1-r}{1-s} - \frac{r}{s}\log\frac{r}{s}\Big],$$

which is positive for all $s \leq r$. Thus $D(r,s)$ is a decreasing function of s in the interval $[0,r]$. In particular,

$$D(\beta_\phi, \alpha_\phi) \leq D(\beta_\phi, A(t_0)) = D\big(B(t_0), A(t_0)\big).$$

Therefore, it suffices to show that there exists a constant M such that for all t,

$$D\big(B(t), A(t)\big) < M.$$

Since $A(t)$ and $B(t)$ are continuous functions of t, it suffices to show that $D\big(B(t), A(t)\big)$ is bounded as t goes to 0 or ∞. It is easy to see that if the likelihood ratio $g(y)/f(y)$ has a positive lower bound $b_0 > 0$, then $D\big(B(t), A(t)\big)$ is 0 if $b < b_0$. So it suffices to consider the case when such a lower bound does not exist.

Now $B(t)$ and $A(t)$ go to 1 as t goes to 0, so

$$\lim_{t\to 0} \sqrt{B(t)}\Big|\log\frac{B(t)}{A(t)}\Big| = 0.$$

By Wald's likelihood ratio identity, we have

$$\begin{aligned} 1 - B(t) &= \mathbf{P}_g\Big(g(Y) \le tf(Y)\Big) = \mathbf{E}_f\Big(\frac{g(Y)}{f(Y)}; g(Y) \le tf(Y)\Big) \\ &\le t\, \mathbf{P}_f\Big(g(Y) \le tf(Y)\Big) = t(1 - A(t)). \end{aligned}$$

Using the fact that $1 - A(t) \le 1$, we know that $\sqrt{1-B(t)}\big|\log\frac{1-B(t)}{1-A(t)}\big|$ is less than

$$\max\Big\{\sqrt{1-B(t)}\big|\log(1-B(t))\big|, \sqrt{1-B(t)}\big|\log t\big|\Big\} \tag{4.49}$$

As $t \to 0$, $B(t) \to 1$, so that the first term in equation (4.49) goes to 0. By Chebyshev's inequality the square of the second term is

$$(\log t)^2 \mathbf{P}_g\Big(\big(-\log\frac{g(Y)}{f(Y)}\big) > |\log t|\Big) \le \mathbf{E}_g\Big(-\log\frac{g(Y)}{f(Y)}\Big)^2,$$

which is finite by the assumption. Hence

$$\limsup_{t\to 0} D\big(B(t), A(t)\big) < \infty$$

Similarly, it is clear that

$$\lim_{t\to\infty} \sqrt{1-B(t)}\Big|\log\frac{1-B(t)}{1-A(t)}\Big| = 0,$$

and

$$\limsup_{t\to\infty} \sqrt{B(t)}\Big|\log\frac{B(t)}{A(t)}\Big| = \limsup_{t\to\infty} \sqrt{B(t)}\big|\log A(t)\big|$$

is finite by the assumption in (4.47). Hence

$$\limsup_{t\to\infty} D\big(B(t), A(t)\big) < \infty,$$

and Theorem 4.14 is proved. □

Corollary 4.15. *Suppose the distribution of the random variable Y belongs to a one-parameter exponential family having the continuous densities*

$$f_\theta(y) = \exp\{\theta y - b(\theta)\}, \quad -\infty < y < \infty,\ \theta \in \Omega$$

with respect to a σ-finite measure μ, where Ω is an open interval on the real line and $b(\theta)$ is twice differentiable with respect to θ. Let $F_\theta(y)$ denote the distribution function of Y. Consider $\theta_0 < \theta_1$ in Ω, and let $f_i = f_{\theta_i}$ and $F_i = F_{\theta_i}$ for $i = 0, 1$. Define $y_0 = \sup\{y : F_0(y) = 0\}$ and $y_1 = \inf\{y : F_1(y) = 1\}$. If

$$\lim_{y \to y_0} \frac{(F_0(y))^{3/2}}{f_0(y)} < \infty, \quad \text{and} \quad \lim_{y \to y_1} \frac{(1 - F_1(y))^{3/2}}{f_1(y)} < \infty,$$

then both $V_2(f_0, f_1)$ and $V_2(f_1, f_0)$ are finite.

Proof. Since $f_1(y)/f_0(y)$ is a monotonically increasing function of y, it suffices for $V_2(f_1, f_0) < \infty$ to show that equations (4.47) and (4.48) hold for $A(t) = 1 - F_0(\log t)$ and $B(t) = 1 - F_1(\log t)$, which is straightforward using L'Hôpital's Rule. The proof is identical for $V_2(f_0, f_1)$. □

Remark: It is easy to check that two normal distributions with the same variance satisfy the conditions in Corollary 4.15, and so do two exponential distributions.

4.4 Numerical Examples

Suppose there are three sensors sending binary messages to the fusion center, i.e., $L = 3$ and $D^l = 2$. Assume that the observations at sensor S_l are independent and identically distributed normal random variables with mean 0 and variance 1 under H_0 and with mean μ_l and variance 1 under H_1.

If $\mu_l > 0$, then in cases **A, C** and **E**, the likelihood ratio at sensor S_l is a monotonically increasing function of the observation, and hence the MLRQ at each sensor

S_l can be written as

$$U_k^l = \begin{cases} 1 & X_k^l \geq \lambda^l; \\ 0 & \text{otherwise.} \end{cases}$$

Thus the Kullback-Leibler information number for U_k^l is

$$I(\lambda^l) = h\big(\Phi(\lambda^l - \mu_l), \Phi(\lambda^l)\big),$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable and $h(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$. Since the function $I(\lambda^l)$ has a unique maximum value over $[0,\infty]$, it is easy to find the optimal λ^l.

Let us consider the case of three nonidentical sensors: $\mu_1 = 0.4, \mu_2 = 0.8$ and $\mu_3 = 1$. The optimal thresholds λ^l are $0.3169, 0.6347$ and 0.7941, respectively, and the corresponding optimal Kullback-Leibler information numbers $I(\lambda^l)$ are $0.0509, 0.2038$ and 0.3186, respectively. Therefore,

$$I_{\mathbf{D}} = 0.5733, \quad \text{and} \quad I_{tot} = \sum_{l=1}^{3} \frac{(\mu_l)^2}{2} = 0.9.$$

Consider the following three optimal procedures: (i) $M(a)$ defined by (4.12) in cases **A, C** and **E**; (ii) $T(a)$ defined by (4.16) in cases **B** and **D**, and (iii) $N(a)$, the optimal one-sided SPRT for the centralized problem, which is defined by

$$N(a) = \inf\{n \geq 1 : \mathbf{S}_n \geq a\}, \quad \text{where} \quad \mathbf{S}_n = \sum_{k=1}^{n}\sum_{l=1}^{L} \log\frac{g^l(X_k^l)}{f^l(X_k^l)}. \tag{4.50}$$

Since $1/I_{\mathbf{D}} = 1.7443$ and $1/I_{tot} = 1.1111$, the asymptotic theory in Section 4.1 indicates that

$$\begin{aligned} \mathbf{E}_1 M &= 1.7443\,|\log\alpha| + O(1); \\ \mathbf{E}_1 T &= 1.1111\,|\log\alpha| + O(\sqrt{|\log\alpha|}); \\ \mathbf{E}_1 N &= 1.1111\,|\log\alpha| + O(1), \end{aligned} \tag{4.51}$$

where α is the error probability.

For each of these three stopping times $\tau(a)$ and $a = 20 * i$, $1 \leq i \leq 100$, we simulated the expected sample size $\mathbf{E}_1\tau(a)$ directly and used importance sampling to simulate the error probability $\alpha = \mathbf{P}_0(\tau(a) < \infty)$. That is, the estimator of α is

$$\hat{\alpha} = \frac{1}{m}\sum_{j=1}^{m} \exp(-\mathbf{S}_{\tau_j}), \tag{4.52}$$

based on m independent realization of $(\tau, \mathbf{S}_\tau)$ under the probability $\mathbf{P}_1$. Fitting functions of α of the same form as (4.51) to the simulation results based on 2500 repetitions, we obtained

$$\begin{aligned} \mathbf{E}_1 M &= 1.7441\ |\log\alpha| + 0.3109; \\ \mathbf{E}_1 T &= 1.1150\ |\log\alpha| + 2.712\sqrt{|\log\alpha|} - 1.109; \\ \mathbf{E}_1 N &= 1.1112\ |\log\alpha| + 0.2217, \end{aligned}$$

which agree well with the asymptotic expressions in (4.51).

Now consider the same examples for change-point problem, i.e., suppose there are three sensors sending binary messages to the fusion center, so that $L = 3$ and $D^l = 2$. Assume that the observations at sensor S_l are independent and identically distributed normal random variables with mean 0 and variance 1 before the change and independent and identically distributed with mean μ_l and variance 1 after the change. Again, assume $\mu_1 = 0.4, \mu_2 = 0.8$ and $\mu_3 = 1$.

We compare the following three optimal procedures: (i) $M^*(a)$, defined by (4.34) in cases **A, C** and **E**; (ii) $T^*(a)$, defined by (4.38) in cases **B** and **D**, and (iii) $N^*(a)$, Page's CUSUM procedure in the centralized problem, which is defined by

$$N^*(a) = \inf\{n \geq 1 : \max_{0 \leq k \leq n-1} \big(\mathbf{S}_n - \mathbf{S}_k\big) \geq a\},$$

where $\mathbf{S}_n$ is defined by (4.50) and $\mathbf{S}_0 = 0$. Now the asymptotic theory in Section 4.2

indicates that

$$\begin{aligned} \overline{E}_1 M^* &= 1.7443 \ \log \gamma + O(1); \\ \overline{E}_1 T^* &= 1.1111 \ \log \gamma + O(\sqrt{\log \gamma}); \\ \overline{E}_1 N^* &= 1.1111 \ \log \gamma + O(1), \end{aligned} \tag{4.53}$$

where γ is the mean time between false alarms.

For these three procedures $\tau(a)$, the renewal property of CUSUM statistics implies that the detection delay $\overline{\mathbf{E}}_1\tau$ (or $D(\tau)$) is just $\mathbf{E}_1\tau$, the expected sample size when the change happens at time $\nu = 1$. It is therefore straightforward to simulate the detection delay. However, it is very difficult to simulate $\mathbf{E}_\infty\tau(a)$ directly if a is large. As pointed out in Lai [10], if τ is a CUSUM procedure,

$$\mathbf{E}_\infty \tau \sim \frac{n}{\mathbf{P}_\infty(\tau \le n)}, \tag{4.54}$$

provided that $\gamma = \mathbf{E}_\infty\tau$ satisfies $n/\log\gamma \to \infty$ and $\log n = o(\log \gamma)$ as $\gamma \to \infty$. It can be shown that expression (4.54) also holds for these three asymptotically optimal procedures. Thus, by importance sampling, a good estimator of $\mathbf{E}_\infty\tau(a)$ for these procedures is

$$n/\Big\{\frac{1}{m}\sum_{j=1}^{m}\big(\exp(-\mathbf{S}_{\tau_j})1(\tau_j \le n)\big)\Big\},$$

based on m independent realization of $(\tau, \mathbf{S}_\tau)$ up to the specified time $n = a^2$ under the probability $\mathbf{P}_1$.

For $a = 20{*}i$, $1 \le i \le 100$, Monte Carlo experiments with 2500 repetitions yielded estimates for the detection delay $\overline{\mathbf{E}}_1\tau$ and the mean time between false alarms for each of these three optimal procedures. Fitting function of α of the same form as (4.53) to the simulation results, we obtained, with $\gamma = \mathbf{E}_\infty \tau(a)$

$$\begin{aligned} \overline{E}_1 M^* &= 1.7417 \ \log\gamma - 7.423; \\ \overline{E}_1 T^* &= 1.1136 \ \log\gamma + 2.6857\sqrt{\log\gamma} - 8.9343; \\ \overline{E}_1 N^* &= 1.1091 \ \log\gamma - 4.4063, \end{aligned}$$

which agree well with the expression (4.53) obtained from the asymptotic theory in Section 4.2.

Chapter 5

Some Results Related to Classical Change-Point Problems

In this chapter, we study classical change-point problems with given pre-change distribution. We briefly introduce the Shiryayev-Roberts procedures. Then we show that the proof of the optimality property of the randomized Shiryayev-Roberts procedure in Yakir [43] is wrong. Finally we construct a counterexample to disprove Pollak's conjecture on change-point problems for dependent observations. In this chapter we use the notations in Pollak [24] and Yakir [43], which are slightly different from those in the previous chapters.

5.1 Shiryayev-Roberts Procedure

Given densities f and g, suppose that we observe a sequence of independent random variables $X_1, X_2, \ldots$ whose density changes at some unknown time from f to g. Let $\mathbf{P}_k$ denote the probability measure (with change time k) when $X_1, \ldots, X_{k-1}$ have density f, while $X_k, X_{k+1}, \ldots$ have density g. Let $\mathbf{P}_\infty$ denote the probability measure when there is no change, i.e., $X_1, X_2, \ldots$ are independent and identically distributed with density f. Then Page's CUSUM procedure in (2.5) can be written as

$$T_A = \inf \big\{ n \geq 1 : \max_{1 \leq k \leq n} \prod_{i=k}^{n} \frac{g(X_i)}{f(X_i)} \geq A \big\}.$$

The following alternative to Page's CUSUM procedure T_A has been suggested by

Shiryayev [30] and Roberts [28]:

$$N_A = \inf \Big\{ n \geq 1 : \sum_{k=1}^{n} \prod_{i=k}^{n} \frac{g(X_i)}{f(X_i)} \geq A \Big\}. \tag{5.1}$$

This so-called Shiryayev-Roberts procedure is a limit of Bayes solutions and its properties have been studied in Pollak [24, 25]. It is well known that the behaviors of these two procedures are similar (see, e.g., Pollak and Siegmund [26]).

Note that if we define

$$R_n = \sum_{k=1}^{n} \prod_{i=k}^{n} \frac{g(X_i)}{f(X_i)}, \tag{5.2}$$

then

$$R_n = (1 + R_{n-1}) \frac{g(X_n)}{f(X_n)},$$

for $n = 1, 2, \ldots$, where $R_0 = 0$.

This inspired Pollak [24] to consider the following randomized Shiryayev-Roberts procedure:

$$\tau(A, \varphi) = \inf\{n \geq 0 : \ R_n^* \geq A\}, \tag{5.3}$$

where

$$R_n^* = (1 + R_{n-1}^*) \frac{g(X_n)}{f(X_n)},$$

and $R_0^* \in [0, \infty)$ has distribution φ chosen by the statistician.

For the right distribution φ_0, the asymptotic optimality of $\tau(A, \varphi_0)$ was proved in Pollak [24]. Later Yakir [43] claimed that it is exactly optimal in the sense of minimizing the "average" detection delay

$$D_g(N) = \sup_{1 \leq k < \infty} \mathbf{E}_k(N - k + 1 | N \geq k - 1) \tag{5.4}$$

among all stopping time N satisfying $\mathbf{E}_\infty N \geq \mathbf{E}_\infty \tau(A, \varphi_0)$.

In Section 5.2, we point out that the proof in Yakir [43] is wrong. Simulation results support our conclusions. It is still an open problem whether $\tau(A, \varphi_0)$ is exactly optimal.

It is natural to extend the theory of change-point problems to the case of dependent observations. Lai [11] showed that the analog of Page's CUSUM procedure is still asymptotically optimal under some conditions which are difficult to verify in general. Pollak [25] showed that the analog of the Shiryayev-Roberts procedure is asymptotically optimal in change-point problems for post-change distributions that are a certain type of mixture. Pollak conjectured that the analogs of the Shiryayev-Roberts procedures are asymptotically optimal for dependent observations in a wide context ([24], [44]). In Section 5.3, we construct a simple counterexample to show that the close relationship between open-ended hypothesis tests and change-point procedures may fail and the analogs of Page's CUSUM or the Shiryayev-Roberts procedures are not in general asymptotically optimal for dependent observations.

5.2 On Yakir's Optimality Proof

In this section, we explain what is wrong with Yakir's proof of the exact optimality of the randomized Shiryayev-Roberts procedure.

5.2.1 Theoretical Results

In order to prove optimality properties of $N_A^* = \tau(A, \varphi_0)$ for the right distribution φ_0, Pollak [24] and Yakir [43] considered the following extended Bayes problem $B(G, p, c)$. Let G be a distribution over the interval $[0, 1]$. Suppose $0 < p < 1$. Assume that a random variable π_0 is sampled from the distribution G before taking any observations. Given the observed value of π_0, suppose the prior distribution of the change-point ν is given by $\mathbf{P}(\nu = 1) = \pi_0$ and $\mathbf{P}(\nu = n) = (1 - \pi_0)p(1 - p)^{n-2}$ for $n \geq 2$. Consider the problem of minimizing the risk

$$\mathbf{P}(N < \nu - 1) + c\mathbf{E}(N - \nu + 1)^+,$$

where $c > 0$ can be thought of as the cost per observation of sampling after a change. It is well known (Shiryayev [30]) that the Bayes solution of this extended Bayes problem $B(G, p, c)$ is of the form

$$M^*_{G,p,c} = \inf\{n \geq 0 : \ R^*_{q,n} \geq A\},$$

where $q = 1 - p$, and

$$R^*_{q,0} = \frac{\pi_0 q}{p(1-\pi_0)} - 1, \quad R^*_{q,n} = (R^*_{q,n-1} + 1)\frac{g(X_n)}{f(X_n)}\frac{1}{q} \quad \text{for} \ \ n \geq 1,$$

where π_0 has a distribution G. Yakir [43] showed that for some sequence of $p \to 0$, there exists a sequence of $G = G_p$ and $c = c_p$ such that $c \to c^*$ and $\pi_0/p \to R^*_0 + 1$ in distribution, and so N^*_A is a limit of Bayes solutions $M^*_{G,p,c}$. Yakir [43] claimed that the Bayes solution $M^*_{G,p,c}$ satisfies

$$\begin{aligned} & \lim_{p \to 0} \frac{1 - \{\text{Expected loss using } M^*_{G,p,c} \text{ for Problem } B(G,p,c)\}}{p} \\ = \ & (1 - c^* \mathbf{E}_1 N^*_A)(\mathbf{E} R^*_0 + 1 + \mathbf{E}_\infty N^*_A). \end{aligned} \tag{5.5}$$

The proof of the exact optimality of the randomized Shiryayev-Roberts procedure in Yakir [43] is based on this equation. However, the next theorem shows that equation (5.5) does not hold in general.

Theorem 5.1.

$$\begin{aligned} & \lim_{p \to 0} \frac{1 - \{\textit{Expected loss using } M^*_{G,p,c} \textit{ for Problem } B(G,p,c)\}}{p} \\ = \ & \mathbf{E}_1\Big((1 - c^* N^*_A)(R^*_0 + 1)\Big) + (1 - c^* \mathbf{E}_1 N^*_A)\mathbf{E}_\infty N^*_A. \end{aligned} \tag{5.6}$$

Proof. For the extended Bayes problem $B(G, p, c)$, any stopping rule N satisfies

$$\frac{1 - \{\text{Expected loss using } N \text{ for Problem } B(G,p,c)\}}{p}$$

$$= \quad \frac{\mathbf{P}(N \geq \nu - 1)}{p}\Big[1 - c\mathbf{E}(N - \nu + 1 | N \geq \nu - 1)\Big]. \tag{5.7}$$

Note that

$$\begin{aligned} \frac{\mathbf{P}(N \geq \nu - 1 | \pi_0)}{p} &= \frac{\pi_0}{p} + \sum_{k=2}^{\infty} (1 - \pi_0)(1 - p)^{k-2} \mathbf{P}(N \geq k - 1 | v = k) \\ &= \frac{\pi_0}{p} + \sum_{k=2}^{\infty} (1 - \pi_0)(1 - p)^{k-2} \mathbf{P}_\infty(N \geq k - 1). \end{aligned}$$

Since $\pi_0/p \to R_0^* + 1$ in distribution, and $M_{G,p,c}^* \to N_A^*$ as $p \to 0$, we have

$$\lim_{p \to 0} \frac{\mathbf{P}(M_{G,p,c}^* \geq \nu - 1)}{p} = \mathbf{E}R_0^* + 1 + \mathbf{E}_\infty N_A^*. \tag{5.8}$$

Arguing as in lemma 13 of Pollak [24], we have

$$\lim_{p \to 0} \mathbf{E}(M_{G,p,c}^* - \nu + 1 | M_{G,p,c}^* \geq \nu - 1) =$$

$$\mathbf{E}_1 N_A^* \frac{\mathbf{E}_\infty N_A^*}{\mathbf{E}R_0^* + 1 + \mathbf{E}_\infty N_A^*} + \frac{\mathbf{E}R_0^* + 1}{\mathbf{E}R_0^* + 1 + \mathbf{E}_\infty N_A^*} \lim_{p \to 0} \mathbf{E}(N_A^* | \nu = 1), \tag{5.9}$$

and the limiting distribution of R_0^* conditional on $\{\nu = 1\}$ has the density $d\varphi_1(x) = (x+1)d\varphi_0(x)/\int(x+1)d\varphi_0(x)$. Since $R_0^* \geq 0$, by definition (5.3),

$$\mathbf{E}(N_A^* | R_0^*, \nu = 1) \leq \mathbf{E}(N_A^* | R_0^* = 0, \nu = 1) = \mathbf{E}_1 N_A,$$

where N_A is defined in (5.1). It is well known (Pollak [24]) that $\mathbf{E}_1 N_A < \infty$. Thus, by bounded converge theorem,

$$\lim_{p \to 0} \mathbf{E}(N_A^* | \nu = 1) = \lim_{p \to 0} \mathbf{E}\Big(\mathbf{E}(N_A^* | R_0^*, \nu = 1)\Big| \nu = 1\Big) = \mathbf{E}_1 \tau(A, \varphi_1), \tag{5.10}$$

since the limiting distribution of R_0^*, given the event $\{\nu = 1\}$, is φ_1. Using the relation between φ_0 and φ_1, it is easy to see that

$$\mathbf{E}_1\tau(A,\varphi_1) = \frac{\mathbf{E}_1\big(N_A^*(R_0^*+1)\big)}{\mathbf{E}R_0^*+1}. \tag{5.11}$$

The theorem follows at once from Equations (5.7) – (5.11). □

Remark: It is interesting to note that N_A^* is a so-called *equalizer rule*, i.e., for all $k \geq 1$,

$$\mathbf{E}_k(N_A^* - k + 1 | N_A^* \geq k-1) = \mathbf{E}_1 N_A^*.$$

Yakir [43] claimed that N_A^* is also an equalizer rule in the context of the extended Bayes problem $B(G,p,c)$, i.e., for all $k \geq 1$,

$$\lim_{p\to 0} \mathbf{E}(N_A^* - \nu + 1 | N_A^* \geq \nu - 1, \nu = k) = \mathbf{E}_1 N_A^*.$$

However, although this is true for all $k \geq 2$, it is false for $k = 1$ in general. Note that (5.10) and (5.11) tell us that for $k = 1$,

$$\lim_{p\to 0} \mathbf{E}(N_A^* - \nu + 1 | N_A^* \geq \nu - 1, \nu = 1) = \mathbf{E}_1\tau(A,\varphi_1) \neq \mathbf{E}_1 N_A^* \quad \text{in general.}$$

Thus, Yakir's claim is wrong.

It is natural to do simulations to confirm that (5.5) fails while (5.6) is correct. However, it is difficult to simulate the value of the left-hand side of these two equations. Now based on (5.5), Yakir [43] also showed that

$$\mathbf{E}_1 N_A^* = \frac{(\mu_0+1)(1-p_0)}{p_0(\mu_0+1)+1}, \tag{5.12}$$

where

$$p_0 = \mathbf{P}(R_0^* \geq A) \qquad \text{and} \qquad \mu_0 = E(R_0^* | R_0^* < A).$$

Yakir is *correct* in deriving (5.12) as a consequence of (5.5). Our result (5.6) and the arguments in Yakir [43] lead to

$$\mathbf{E}_1 N_A^* = (\mu_0 + 1)(1 - p_0) - p_0 \mathbf{E}_1\Big(R_0^* N_A^*\Big). \tag{5.13}$$

Thus, in order to confirm that Yakir's proof is wrong, it suffices to show that (5.12) fails while (5.13) is correct.

5.2.2 Numerical Examples

To illustrate that (5.13) is correct and (5.12) is not, we have performed simulations for the following example, which is considered by Pollak [24] and Yakir [43].

Define $f_0(x) = \exp\{-x\}1(x > 0)$ and $f_1(x) = \theta \exp\{-\theta x\}1(x > 0)$, where $\theta > 1$, and pick an A such that $0 < A < \theta$. Yakir [43] chose the randomized R_0^* as follows. Let the distribution of R^* be φ_0, where

$$\varphi_0(x) = \begin{cases} 0 & \text{if } x \le 0; \\ (x/A)^{1/(\theta-1)} & \text{if } 0 < x \le A; \\ 1 & \text{if } x > A. \end{cases}$$

Let the distribution of Z be the $\mathbf{P}_\infty$ distribution of $f_1(X)/f_0(X)$. Both R^* and Z are independent of the sequence of observations $X_1, X_2, \ldots$. Then $R_0^* = (R^* + 1)Z$.

As shown in Yakir [43], in this example,

$$\mu_0 = \int_0^A x \, d\phi_0(x) = \frac{A}{\theta}, \quad \text{and} \quad p_0 = 1 - \frac{1}{\theta^{1/(\theta-1)}(\theta - 1)} \int_0^A \frac{x^{(2-\theta)/(\theta-1)}}{(x+1)^{1/(\theta-1)}} \, dx.$$

If $\theta = 3$, then $p_0 = 1 - (\log(2A + 1 + 2\sqrt{A^2 + A}))/(2\sqrt{3})$, and if $\theta = 2$, then $p_0 = 1 - (\log(A + 1))/2$.

Table 5.1 compares the theoretical values of $\mathbf{E}_1 N_A^*$ given by (5.12) and (5.13) to Monte Carlo estimates. Our theoretical result (5.13) was based on Monte Carlo estimates of $\mathbf{E}_1(R_0^* N_A^*)$, while Yakir's result (5.12) was calculated exactly. The number of repetitions was $100,000$ in the Monte Carlo experiment.

The results in Table 5.1 suggest that (5.13) gives correct values for $\mathbf{E}_1 N_A^*$ and

Table 5.1: Approximations for $\mathbf{E}_1 N_A^*$

θ	A	Monte Carlo	Our result (5.13)	Yakir's result (5.12)
3	2.9	1.2357 ± 0.0032	1.2382 ± 0.0010	0.9938
3	2.7	1.1652 ± 0.0031	1.1632 ± 0.0010	0.9268
3	2.5	1.0877 ± 0.0029	1.0927 ± 0.0009	0.8617
2	1.98	0.7779 ± 0.0027	0.7715 ± 0.0013	0.5708
2	1.8	0.7005 ± 0.0026	0.6993 ± 0.0012	0.5090
2	1.5	0.5798 ± 0.0023	0.5813 ± 0.0010	0.4115

(5.12) does not. These results support the claim that Yakir's proof of exact optimality of the randomized Shiryayev-Roberts procedures is flawed.

5.3 Pollak's Conjecture on Problems with Dependent Observations

In this section we construct an example to show that the close relationship between open-ended hypothesis tests and change-point procedures no longer holds for dependent observations. It also disproves the conjectures of Pollak [24] and Yakir, Krieger and Pollak [44], which state that Page's CUSUM and the Shiryayev-Roberts procedures are asymptotically optimal for dependent observations. Our example illustrates that in open-ended hypothesis testing problems in which the null hypothesis specifies a mixture of distributions, the SPRT is asymptotically optimal, but Page's CUSUM and the Shiryayev-Roberts procedures are not asymptotically optimal in the corresponding change-point problems.

Consider three given probability densities f_1, f_2 and g such that

$$\mathbf{E}_g\Big(\log\frac{g(X)}{f_j(X)}\Big)^2 < \infty \quad \text{for } j = 1, 2, \quad \text{and} \quad \mathbf{E}_{f_1}\Big(\log\frac{f_1(X)}{f_2(X)}\Big)^2 < \infty, \tag{5.14}$$

and $I_1 > I_2$, where $I_j = I(g, f_j) = \mathbf{E}_g \log(g(X)/f_j(X))$ $(j = 1, 2)$ are the Kullback-Leibler information numbers. Denote by $\mathbf{P}_{f_1}, \mathbf{P}_{f_2}$ and $\mathbf{P}_g$ the probability measures when $X_1, X_2, \cdots$ are independent and identically distributed with densities f_1, f_2 and

g, respectively. Choose a constant $\pi_0 \in (0,1)$, say $\pi_0 = 1/2$. Define

$$\mathbf{P}_f = \pi_0 \mathbf{P}_{f_1} + (1-\pi_0)\mathbf{P}_{f_2}.$$

Under $\mathbf{P}_f$, $X_1, X_2, \cdots, X_n$ have a "mixture" joint density

$$f(x_1, \cdots, x_n) = \pi_0 \prod_{i=1}^n f_1(x_i) + (1-\pi_0)\prod_{i=1}^n f_2(x_i). \tag{5.15}$$

Suppose that $X_1, X_2, \ldots$ are sampled from a true distribution $\mathbf{P}$, and we are interested in testing the null hypothesis $H_0 : \mathbf{P} = \mathbf{P}_f$ against the alternative hypothesis $H_1 : \mathbf{P} = \mathbf{P}_g$. The one-sided SPRT is defined by

$$\begin{aligned} \tau_A &= \inf\left\{n \geq 1 : \prod_{i=1}^n \frac{g(X_i|X_1, \cdots, X_{i-1})}{f(X_i|X_1, \cdots, X_{i-1})} \geq A\right\} \\ &= \inf\left\{n \geq 1 : \frac{\prod_{i=1}^n g(X_i)}{\pi_0 \prod_{i=1}^n f_1(X_i) + (1-\pi_0)\prod_{i=1}^n f_2(X_i)} \geq A\right\} \end{aligned}$$

In change-point problems, we are interested in detecting a change in distribution from $\mathbf{P}_f$ to $\mathbf{P}_g$. Page's CUSUM procedure has stopping time

$$\begin{aligned} T_A &= \inf\left\{n \geq 1 : \max_{1\leq k\leq n} \prod_{i=k}^n \frac{g(X_i|X_1, \cdots, X_{i-1})}{f(X_i|X_1, \cdots, X_{i-1})} \geq A\right\} \\ &= \inf\left\{n \geq 1 : \max_{1\leq k\leq n} \frac{\prod_{i=k}^n g(X_i)}{\pi_{k-1} \prod_{i=k}^n f_1(X_i) + (1-\pi_{k-1})\prod_{i=k}^n f_2(X_i)} \geq A\right\}, \end{aligned}$$

where

$$\pi_k = \frac{\pi_0 \prod_{i=1}^n f_1(X_i)}{\pi_0 \prod_{i=1}^n f_1(X_i) + (1-\pi_0)\prod_{i=1}^n f_2(X_i)}.$$

Similarly, the Shiryayev-Roberts procedure has stopping time

$$\begin{aligned} N_A &= \inf\left\{n \geq 1 : \sum_{k=1}^n \prod_{i=k}^n \frac{g(X_i|X_1, \cdots, X_{i-1})}{f(X_i|X_1, \cdots, X_{i-1})} \geq A\right\} \\ &= \inf\left\{n \geq 1 : \sum_{k=1}^n \frac{\prod_{i=k}^n g(X_i)}{\pi_{k-1} \prod_{i=k}^n f_1(X_i) + (1-\pi_{k-1})\prod_{i=k}^n f_2(X_i)} \geq A\right\}. \end{aligned}$$

The purpose of this section is to establish the asymptotic optimality of τ_A, the SPRT, and to show that T_A and N_A, Page's CUSUM and the Shiryayev-Roberts procedures, are not first-order asymptotically optimal. As a consequence, the conjecture in Pollak [24] and Yakir, Krieger and Pollak [44] is false for dependent observations.

5.3.1 Asymptotic Optimality of the SPRT

The next theorem shows that τ_A, the one-sided SPRT, is asymptotically optimal.

Theorem 5.2. *For any $A > 0$,*

$$\mathbf{P}_f(\tau_A < \infty) \le \frac{1}{A}, \tag{5.16}$$

and as $A \to \infty$

$$\mathbf{E}_g \tau_A = \frac{\log A}{I_2} + O(1). \tag{5.17}$$

Moreover, if $\{N(A)\}$ is a family of stopping times such that (5.16) holds, then

$$\mathbf{E}_g N(A) \ge \frac{\log A}{I_2} + O(1), \quad \text{as } A \to \infty. \tag{5.18}$$

Proof. Relation (5.16) follows at once from Wald's likelihood ratio identity. Note that τ_A can be written as

$$\tau_A = \inf\left\{n \ge 1 : \sum_{k=1}^{n} \log \frac{g(X_k)}{f_2(X_k)} + \log \frac{1}{\pi_0 \gamma_n + (1-\pi_0)} \ge \log A\right\},$$

where $\gamma_n = \prod_{i=1}^{n}(f_1(X_i)/f_2(X_i))$. Since $I_1 > I_2$, we have $\mathbf{E}_g \log \gamma_1 = I_2 - I_1 < 0$. Thus $\gamma_n \to 0$ with probability 1 under $\mathbf{P}_g$. The proof of (5.17) is therefore a direct application of the nonlinear renewal theorem (Theorem 9.28 of Siegmund [33]).

To prove (5.18), note that (5.16) is equivalent to

$$\pi_0 \mathbf{P}_{f_1}(N(A) < \infty) + (1-\pi_0)\mathbf{P}_{f_2}(N(A) < \infty) \le \frac{1}{A},$$

so that we have

$$\mathbf{P}_{f_2}(N(A) < \infty) \le \frac{1}{1-\pi_0} \cdot \frac{1}{A}.$$

Relation (5.18) now follows at once from the well known fact that

$$\mathbf{E}_g N \ge \frac{|\log \mathbf{P}_{f_2}(N < \infty)|}{I_2}.$$

□

5.3.2 Suboptimal Properties in Change-Point Problems

For $1 \le \nu < \infty$, let $\mathbf{P}_\nu$ denote probability when the change in distribution from $\mathbf{P}_f$ to $\mathbf{P}_g$ occurs at the νth observation, so that $X_1, \ldots, X_{\nu-1}$ have joint density f and $X_\nu, X_{\nu+1}, \ldots$ are independent and identically distributed with density g. Let $\mathbf{P}_f$ denote the probability measure when there is no change, i.e. $\nu = \infty$, in which case $X_1, X_2, \ldots$ are distributed with density f. As in the change-point problems for independent observations, we seek a stopping time N which minimizes

$$\overline{\mathbf{E}}_g N$$

subject to the constraint $\mathbf{E}_f N \ge \gamma$.

We first consider the asymptotic behavior of T_A and N_A, Page's CUSUM and the Shiryayev-Robert procedure, and then propose a better procedure, which has much smaller detection delay and roughly the same mean time between false alarms.

Lemma 5.3. *As* $A \to \infty$,

$$\log \mathbf{E}_f N_A \le \log \mathbf{E}_f T_A \le (1 + o(1)) \log A, \tag{5.19}$$

Proof. The first inequality holds by definition, and so it suffices to prove the second inequality. Define a new stopping time

$$t_1(A) = \inf\Big\{n \geq 1 : \prod_{i=1}^{n} \frac{f_2(X_i)}{f_1(X_i)} \leq \log A, \text{ and } \max_{1\leq k\leq n} \prod_{i=k}^{n} \frac{g(X_i)}{f_1(X_i)} \geq K_A\Big\}, \tag{5.20}$$

where $K_A = A(1 + \frac{1-\pi_0}{\pi_0} \log A)$. Note that

$$\begin{aligned}\prod_{i=k}^{n} \frac{g(X_i|X_1, \cdots, X_{i-1})}{f(X_i|X_1, \cdots, X_{i-1})} &= \Big(\prod_{i=k}^{n} \frac{g(X_i)}{f_1(X_i)}\Big)\Big(\frac{\pi_0 + (1-\pi_0)\prod_{i=1}^{k-1}(f_2(X_i)/f_1(X_i))}{\pi_0 + (1-\pi_0)\prod_{i=1}^{n}(f_2(X_i)/f_1(X_i))}\Big) \\ &\geq \Big(\prod_{i=k}^{n} \frac{g(X_i)}{f_1(X_i)}\Big)\Big(1 + \frac{1-\pi_0}{\pi_0}\prod_{i=1}^{n} \frac{f_2(X_i)}{f_1(X_i)}\Big)^{-1},\end{aligned}$$

and hence $T_A \leq t_1(A)$ by definition. By the following lemma, $\mathbf{E}_{f_1} t_1(A) = O(A \log A)$, and so $\mathbf{E}_{f_1} T_A \leq O(A \log A)$. Similarly, $\mathbf{E}_{f_2} T_A \leq O(A \log A)$. The lemma follows, since $\mathbf{E}_f N = \pi_0 \mathbf{E}_{f_1} N + (1-\pi_0)\mathbf{E}_{f_2} N$. □

Lemma 5.4. *As $A \to \infty$,*

$$\mathbf{E}_{f_1} t_1(A) = O(A \log A), \tag{5.21}$$

where $t_1(A)$ is defined in (5.20).

Proof. Let $S_n = \sum_{i=1}^{n} \log(f_1(X_i)/f_2(X_i))$ and $V_n = \sum_{i=1}^{n} \log(g(X_i)/f_1(X_i))$ for $n = 1, 2, \ldots,$ and $V_0 = 0$. Denote $W_n = \max_{0\leq k\leq n-1}(V_n - V_k)$, then $t_1(A)$ can be written as

$$t_1(A) = \inf\{n \geq 1 : S_n \geq -\log\log A \quad \text{and} \quad W_n \geq K_A\}.$$

Using an idea of Kiefer and Sacks [9], let v_1 be the first n such that $W_n \geq K_A$, v_2 the second n such that $W_n \geq K_A$, etc. Let ϕ_t be the indicator function of the set where $S_{v_t} < -\log\log A, t = 1, 2, \ldots$. Then as shown on page 719 of Kiefer and Sacks [9],

$$t_1(A) = v_1 + \sum_{j=1}^{\infty}(v_{j+1} - v_j)\prod_{t=1}^{j} \phi_t.$$

Let $v_{j+1}^* - v_j$ be the first m such that $\max_{1\leq k\leq m}(V_{m+v_j} - V_{k+v_j}) \geq K_A$. Evidently,

$v_{j+1}^* - v_j \geq v_{j+1} - v_j$. Since $v_{j+1}^* - v_j$ depends on X's whose indices are greater than v_j, it follows that $v_{j+1}^* - v_j$ is independent of $\phi_1, \ldots, \phi_j$. Moreover, $v_{j+1}^* - v_j$ has the same distribution as v_1. Consequently

$$\mathbf{E}_{f_1} t_1(A) \leq (\mathbf{E}_{f_1} v_1)\big(1 + \sum_{j=1}^{\infty} \mathbf{E}_{f_1} \prod_{t=1}^{j} \phi_t\big).$$

Let σ be the *last* time $S_n < -\log\log A$. Since $v_j \geq j$, we have

$$\sum_{j=1}^{\infty} \mathbf{E}_{f_1} \prod_{t=1}^{j} \phi_t \leq \sum_{j=1}^{\infty} \mathbf{P}_{f_1}(\sigma \geq v_j) \leq \sum_{j=1}^{\infty} \mathbf{P}_{f_1}(\sigma \geq j) = \mathbf{E}_{f_1}\sigma.$$

Since the summands in S_n have mean $I(f_1, f_2) > 0$ and $\mathrm{Var}(S_1) < \infty$ by assumption (5.14), it is well known that $\mathbf{E}_{f_1}\sigma < \infty$, (see, for example, Theorem D in Kiefer and Sacks [9]). By a property of Page's CUSUM procedure, we know $\mathbf{E}_{f_1} v_1 = O(K_A)$. Thus, $\mathbf{E}_{f_1} t_1(A) \leq O(K_A)O(1) = O(K_A) = O(A \log A)$. □

Lemma 5.5. *As* $A \to \infty$,

$$\overline{\mathbf{E}}_g T_A \geq \overline{\mathbf{E}}_g N_A \geq \mathbf{E}_g N_A \geq \frac{1 + o(1)}{I_2} \log A, \tag{5.22}$$

Proof. The first inequality holds by the definitions of T_A and N_A, and the second inequality holds by the definition of $\overline{\mathbf{E}}_g N$. So it suffices to prove the last inequality. Rewrite N_A as

$$N_A = \inf\Big\{n \geq 1 : \sum_{i=1}^{n} \log \frac{g(X_i)}{f_2(X_i)} + \log \frac{\pi_0 W_n^{(1)} + (1 - \pi_0) W_n^{(2)}}{\pi_0 \gamma_n + (1 - \pi_0)} \geq \log A\Big\},$$

where

$$\gamma_n = \frac{f_{1n}}{f_{2n}}, \qquad \text{and} \quad W_n^{(j)} = 1 + \sum_{i=1}^{n-1} \frac{f_j(X_1) \cdots f_j(X_i)}{g(X_1) \cdots g(X_i)}, \quad j = 1, 2.$$

Since $I_1 > I_2$, $\mathbf{E}_g \log \gamma_1 = I_2 - I_1 < 0$. Thus, under $\mathbf{P}_g$, $\gamma_n \to 0$ with probability 1, and $W_n^{(j)} \to W^{(j)}$, a finite random variable (see Pollak [25]). By the nonlinear renewal theorem, we have that

$$\mathbf{E}_g N_A = \frac{1+o(1)}{I_2} \log A$$

as $A \to \infty$. $\square$

Theorem 5.6. *If $N(\gamma)$ are stopping times such that $\mathbf{E}_f N(\gamma) \geq \gamma$, then*

$$\overline{\mathbf{E}}_g N(\gamma) \geq (1+o(1)) \frac{\log \gamma}{I_1}$$

as $\gamma \to \infty$, and there exist stopping times for which equality holds. Thus T_A and N_A, Page's CUSUM and the Shiryayev-Roberts procedures, are asymptotically suboptimal.

Proof. For any stopping time $N = N(\gamma)$ such that $\mathbf{E}_f N \geq \gamma$, we have

$$\pi_0 \mathbf{E}_{f_1} N + (1-\pi_0) \mathbf{E}_{f_2} N \geq \gamma,$$

thus $\mathbf{E}_{f_1} N \geq \gamma$ or $\mathbf{E}_{f_2} N \geq \gamma$. So by the optimality property of Page's CUSUM

$$\overline{\mathbf{E}}_g N \geq (1+o(1)) \min(\frac{\log \gamma}{I_1}, \frac{\log \gamma}{I_2}) = (1+o(1)) \frac{\log \gamma}{I_1}, \quad \text{as} \ \ \gamma \to \infty$$

since $I_1 > I_2$. Moreover, this bound is achieved by the CUSUM procedure $T_1(\gamma)$, defined by

$$T_1(\gamma) = \inf \left\{ n \geq 1 : \max_{1 \leq k \leq n} \prod_{i=k}^{n} \frac{g(X_i)}{f_1(X_i)} \geq \frac{\gamma}{\pi_0} \right\}. \tag{5.23}$$

This is because $\mathbf{E}_{f_1} T(\gamma) \geq \gamma/\pi_0$ and so $\mathbf{E}_f T(\gamma) = \pi_0 \mathbf{E}_{f_1} T(\gamma) + (1-\pi_0) \mathbf{E}_{f_2} T(\gamma) \geq \gamma$, while $\overline{\mathbf{E}}_g T(\gamma) = (1+o(1))(\log \gamma)/I_1$.

Now it is straightforward to show that T_A and N_A are asymptotically suboptimal. Assume $\mathbf{E}_f T_A = \gamma$. Then by Lemmas 5.3 and 5.5,

$$\overline{\mathbf{E}}_g T_A \geq (1+o(1)) \frac{\log \gamma}{I_2} > (1+o(1)) \frac{\log \gamma}{I_1},$$

since $I_1 > I_2$. Thus T_A is asymptotically suboptimal. The proof for N_A is the same. □

Remarks:

1. The lower bound in the previous theorem can also be achieved by our procedures defined in Chapter 3:

$$M^*(a) = \inf\Big\{n \geq 1 : \min_{1\leq k\leq n} \min_{j=1,2} \sum_{i=k}^{n} (\log \frac{g(X_i)}{f_j(X_i)} - I_j \cdot a) > 0\Big\},$$

or

$$M_1^*(a) = \inf\Big\{n \geq 1 : \min_{j=1,2} \min_{1\leq k\leq n} \sum_{i=k}^{n} (\log \frac{g(X_i)}{f_j(X_i)} - I_j \cdot a) > 0\Big\}. \tag{5.24}$$

2. Page's CUSUM and the Shiryayev-Roberts procedures can effectively detect a change from f_2 to g, but they perform poorly when detecting a change from f_1 to g. However, from the asymptotic viewpoint, the standard formulation is equivalent to the problem of detecting a change from f_1 to g. At a small additional cost of detection delay, our procedures $M^*(a)$ and $M_1^*(a)$ are able to detect both changes very well.

3. This counterexample can be extended to the case where

$$f(x_1, \cdots, x_n) = \int_{\underline{\xi}}^{\theta_1} f_\xi(x_1) \cdots f_\xi(x_n) \pi(\xi) d\xi,$$

and $g = f_\lambda$, where $\{f_\xi\}_{\xi\in\Omega}$ are the densities of a one-parameter exponential family, as defined in (3.4), with natural parameter space $\Omega = (\underline{\xi}, \bar{\xi})$ with respect to a sigma-finite measure F, and $\theta_1 < \lambda$.

5.3.3 A Numerical Example

The purpose of this subsection is to give some indication of Theorem 5.6 and to show that Page's CUSUM procedure T_A and the Shiryayev-Roberts procedure N_A are suboptimal.

Table 5.2 compares the results of a 2500-repetition Monte Carlo experiment in

Table 5.2: Comparisons of four stopping times

	$\mathbf{E}_{f_1}\tau$	$\mathbf{E}_{f_2}\tau$	$\mathbf{E}_f\tau$	$\overline{\mathbf{E}}_g\tau$
Page's CUSUM, T_A	6402 ± 129	14214 ± 289	11610	$\geq 51.9 \pm 0.5$
Shiryayev-Roberts, N_A	1766 ± 35	1336 ± 27	1480	$\geq 34.9 \pm 0.3$
$T_1(\gamma)$ ($\gamma = 2000$)	37815 ± 776	9.4 ± 0.1	12612	17.3 ± 0.1
$M_1^*(a)$ ($a = 17.3$)	36102 ± 744	101 ± 2	12101	23.7 ± 0.2

MATLAB. In Table 5.2 we consider the change-point problem in this section with $f_1 = N(1,1), f_2 = N(-0.5,1), g = N(0,1)$ and $\pi_0 = 1/3$. Note that the expected values of sample means are 0 under both the pre-change distribution f and the post-change distribution g.

Four different procedures are considered in Table 5.2. The first is Page's CUSUM procedure T_A. The second is the Shiryayev-Roberts procedure N_A. For $A = 1000$, we simulated $\mathbf{E}_{f_1}\tau, \mathbf{E}_{f_2}\tau$ and $\mathbf{E}_g\tau$ for these two procedures, giving the mean time between false alarms $\mathbf{E}_f\tau = \pi_0\mathbf{E}_{f_1}\tau + (1-\pi_0)\mathbf{E}_{f_2}\tau$, and a lower bound of the detection delay $\overline{\mathbf{E}}_g\tau$ in $\mathbf{E}_g\tau$.

The third procedure is $T_1(\gamma)$, defined by (5.23), and the fourth is $M_1^*(a)$, defined by (5.24). We simulated $\mathbf{E}_{f_1}\tau, \mathbf{E}_{f_2}\tau$ and $\mathbf{E}_g\tau$ for $\tau = T_1(\gamma)$ with $\gamma = 2000$ and $\tau = M_1^*(a)$ with $a = 17.3$. It is easy to see that $\mathbf{E}_f\tau = \pi_0\mathbf{E}_{f_1}\tau + (1-\pi_0)\mathbf{E}_{f_2}\tau$, and the detection delay $\overline{\mathbf{E}}_g\tau$ equals to $\mathbf{E}_g\tau$ for $T_1(\gamma)$ and $M_1^*(a)$. The thresholds γ and a are determined from the criterion $\mathbf{E}_f\tau \geq \mathbf{E}_f T_A \approx 11610$.

Table 5.2 indicates that $T_1(\gamma)$ performs better than both Page's CUSUM procedure T_A and the Shiryayev-Roberts procedure N_A in the sense that $T_1(\gamma)$ has a larger mean time between false alarms and much smaller detection delay, reflecting that Page's CUSUM and the Shiryayev-Roberts procedures are asymptotically suboptimal. The conclusion still holds if $T_1(\gamma)$ is replaced by $M_1^*(a)$.

Bibliography

[1] Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Applications.* Englewood Cliffs: Prentice-Hall.

[2] Blum, R. S., Kassam, S. A. and Poor H. V. (1997). Distributed Detection wth Multiple Sensors: Part II - Advanceed Topics, *Proceedings of the IEEE* **85**, 64-79.

[3] Beibel, M. and Lerche, H. R. (1997). A new look at optimal stopping problems related to mathematical finance. *Statistica Sinica* **7**, 93-108.

[4] Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Statist.* **30**, 755-770.

[5] Chow, Y. S., Hsiung, C. A. and Lai, T. L. (1979). Extended Renewal Theory and Moment Convergence in Anscombe's Theorem. *Ann. Probability* **7**, 304-318.

[6] Crow, R. W. and Schwartz, S. C. (1996). Quickest detection for sequential decentralized decision systems. *IEEE Trans. Aerosp. Electron. Syst.* **32**, 267-283.

[7] Feller, W. (1971). *An introduction to probability theory and its applications*, vol. II, second edition. New York: John Wiley & Sons.

[8] Govindarajulu, Z. (1975). *Sequential statistical procedures.* Academic Press.

[9] Kiefer, J. and Sacks, J. (1963). Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34**, 705-750.

[10] Lai, T. L. (1995). Sequential change-point detection in quality control and dynamical systems. *J. R. Statist. Soc. B* **57**, 613-658.

[11] Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Information Theory* **44**, 2917-2929.

[12] Lai, T. L. (2000). Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE Trans. Information Theory* **46**, 595-608.

[13] Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica* **11**, 303-408.

[14] Lehmann, E. L. (1959). *Testing Statistical Hypothesis.* New York: Wiley.

[15] Lorden, G. (1970). On excess over the boundary. *Ann. Math. Statist.* **41**, 520-527.

[16] Lorden, G. (1971). Procedure for reacting to a change in distribution. *Ann. Math. Statist.* **42**, 1897-1908.

[17] Lorden, G. (1973). Open-ended tests for Koopman-Darmois families. *Ann. Statist.* **1**, 633-643.

[18] Lorden, G. (1976). 2-SPRT'S and the modified Kiefer-Weiss problem of minimizing an expected sample size. *Ann. Statist.* **4**, 281-291.

[19] Lorden, G. (1980). Structure of sequential tests minimizing an expected sample size. *Z. Wahrscheinlichkeistheorie verw. Gebiete* **51**, 291-302.

[20] Mei, Y. (2002). Asymptotically optimal procedures in sequential decentralized decision systems. Submitted to *IEEE Trans. Information Theory.*

[21] Moustakides, G. (1986). Optimal procedures for detecting changes in distributions. *Ann. Statist.* **14**, 1379-1387.

[22] Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100-115.

[23] Pollak, M. (1978). Optimality and almost optimality of mixture stopping rules. *Ann. Statist.* **6**, 910-916.

[24] Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13**, 206-227.

[25] Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.* **15**, 749-779.

[26] Pollak, M. and Siegmund D. (1985). A diffusion process and its application to detecting a change in the drift of Brownian motion. *Biometrika* **72**, 267-280.

[27] Ritov, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Ann. Statist.* **18**, 1464-1469.

[28] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics* **8**, 411-430.

[29] Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist.* **33**, 224-236.

[30] Shiryayev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Prob. and Applic.* **8**, 22-46.

[31] Shiryayev, A. N. (1978). *Optimal Stopping Rules.* New York: Springer-Verlag.

[32] Siegmund, D. (1969). The variance of one-sided stopping rules. *Ann. Math. Statist.* **40**, 1074-1077.

[33] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* New York: Springer-Verlag.

[34] Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistics for sequential detection of a change-point. *Ann. Statist.* **23**, 255-271.

[35] 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. *J. Amer. Statist. Assoc.* **95**, 992-998.

[36] Tartakovsky, A. G. and Veeravalli, V. V. (2002). An efficient sequential procedure for detecting changes in multichannel and distributed systems. *Proceedings of the 5th International Conference on Information Fusion*, 1-8.

[37] Tsitsiklis, J. N. (1993). Extremal properties of likelihood ratio quantizers. *IEEE Trans. Commun.* **41**, 550-558.

[38] Veeravalli, V. V. (1999). Sequential decision fusion: theory and applications. *J. Franklin Inst.* **336**, 301-322.

[39] Veeravalli, V. V. (2001). Decentralized quickest change detection. *IEEE Trans. Information Theory* **47**, 1657-1665.

[40] Veeravalli, V. V., Basar, T. and Poor, H. V. (1993). Decentralized sequential detection with a fusion center performing the sequential test. *IEEE Trans. Information Theory* **39**, 433-442.

[41] Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19**, 326-339.

[42] Williams, D. (1991). *Probabilty with Martingles.* Cambridge Univeristy Press.

[43] Yakir, B. (1997). A note on optimal detection of a change in distribution. *Ann. Statist.* **25**, 2117-2126.

[44] Yakir, B., Krieger, A. M. and Pollak, M. (1999). Detecting a change in regression: first order optimality. *Ann. Statist.* **27**, 1896-1913.