Quickest Change-Point Detection: A Bird's Eye View

Aleksey S. Polunchenko* Grigory Sokolov† Wenyu Du*

Abstract

We provide a bird's eye view onto the area of sequential change-point detection. We focus on the discrete-time case with known pre- and post-change data distributions and offer a summary of the forefront asymptotic results established in each of the four major formulations of the underlying optimization problem: Bayesian, generalized Bayesian, minimax, and multi-cyclic.

Key Words: CUSUM chart, Quickest change detection, Sequential analysis, Sequential change-point detection, Shiryaev's procedure, Shiryaev–Roberts procedure, Shiryaev–Roberts–Pollak procedure, Shiryaev–Roberts–*r* procedure

1. Introduction

Quickest change-point detection is concerned with the design and analysis of procedures for "on-the-go" detection of possible changes in the characteristics of a running (random) process. Specifically, the process is assumed to be monitored continuously through sequentially made observations (e.g., measurements), and should their behavior suggest the process may have statistically changed, the aim is to conclude so within the fewest observations possible, subject to a tolerable level of the risk of false detection. See, e.g., Wald (1947); Shiryaev (1978); Siegmund (1985); Poor and Hadjiliadis (2008). For nonparametric change-point detection theory see, e.g., Brodsky and Darkhovsky (1993). The area finds applications across many branches of science and engineering: industrial quality and process control (see, e.g., Ryan, 2011; Montgomery, 2012; Wetherill and Brown, 1991; Kenett and Zacks, 1998; Shewhart, 1931), biostatistics (see, e.g., Cohen, 1987), clinical trials (see, e.g., Siegmund, 1985), econometrics (see, e.g., Broemeling and Tsurumi, 1987), seismology (see, e.g., Basseville and Nikiforov, 1993), forensics, navigation, cybersecurity (see, e.g., Tartakovsky et al., 2006 and Polunchenko et al., 2012; Tartakovsky et al., 2013), and communication systems (see, e.g., Basseville and Nikiforov, 1993; Tartakovsky, 1991) – to name a few. See also, e.g., Chernoff (1972). A sequential change-point detection procedure, a rule whereby one stops and declares that (apparently) a change is in effect, is defined as a stopping time, T, adapted to the observed data, $\{X_n\}_{n\geqslant 1}$.

The desire to detect the change quickly causes one to be trigger-happy. That is, if one is too hasty, i.e., too quick to stop, the risk of a false detection is high. On the other hand, however, if one is too wary, i.e., too slow to stop, the delay to (correct) detection is substantial. Hence, there is a loss in either case and the essence of the problem is to attain a tradeoff between two contradicting performance measures – the loss associated with the delay to detection of a true change and that associated with raising a false alarm. A good sequential detection policy is expected to minimize the average loss related to the detection delay, subject to a constraint on the loss associated with false alarms (or vice versa).

To put this idea on rigorous mathematical grounds one is to first formally define both the "detection delay" and the "risk of raising a false alarm". To this end, contemporary theory of sequential change-point detection distinguishes four different approaches: the

*Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, NY 13902–6000

†Department of Mathematics, University of Southern California, Los Angeles, CA 90089–2532

minimax approach, the Bayesian approach, the generalized Bayesian approach, and the approach related to multi-cyclic detection of a distant change in a stationary regime. The aim of this paper is to give a brief overview of all four. For a more detailed overview see, e.g., Polunchenko and Tartakovsky (2012), Tartakovsky and Moustakides (2010), and Tartakovsky and Veeravalli (2005).

2. Change-Point Models

To formally state the general quickest change-point detection problem, one is to first introduce a change-point model, i.e., describe a probabilistic structure of the observations (independent, identically or non-identically distributed, correlated, etc.) as well as that of the change-point (unknown deterministic, random completely or partially dependent on the observed data, random fully independent from the observations). To this end, a myriad of scenarios is possible; see, e.g., Fuh (2003, 2004), Tartakovsky (1991, 2009a), Tartakovsky and Moustakides (2010), Lai (1995, 1998), Shiryaev (1961, 1963, 1978, 2009, 2010), Tartakovsky and Veeravalli (2005), and Polunchenko and Tartakovsky (2012). This section is intended to review the major ones.

Fix a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F} = \vee_{n\geqslant 0}\mathcal{F}_n$, $\mathcal{F}_n$ is the sigma-algebra generated by the first $n \geqslant 1$ observations ($\mathcal{F}_0 = \{\varnothing, \Omega\}$ is the trivial sigma-algebra), and $\mathbb{P}\colon \mathcal{F} \mapsto [0,1]$ is a probability measure. Let $\mathbb{P}_\infty$ and $\mathbb{P}_0$ be two mutually locally absolutely continuous (i.e., equivalent) probability measures; for a general case with singular measures present see Shiryaev (2009). For $d = \{0, \infty\}$, write $\mathbb{P}_d^{(n)} = \mathbb{P}_d|_{\mathcal{F}_n}$ for the restriction of $\mathbb{P}_d$ to $\mathcal{F}_n$, and let $p_d^{(n)}(\cdot)$ be the density of $\mathbb{P}_d^{(n)}$ (with respect to a dominating sigma-finite measure).

Let $\{X_n\}_{n\geqslant 1}$ be the series of observations such that $X_1, X_2, \ldots, X_\nu$, for some ν, adhere to measure $\mathbb{P}_\infty$ ("normal" regime), but $X_{\nu+1}, X_{\nu+2}, \ldots$ follow measure $\mathbb{P}_0$ ("abnormal" regime). That is, at an unknown time instant ν (change-point), the observations undergo a change-of-regime from "normal" to "abnormal". Hence, ν is the serial number of the last normal observation, so that if $\nu = 0$, then the entire series $\{X_n\}_{n\geqslant 1}$ is in the abnormal regime admitting measure $\mathbb{P}_0$, while if $\nu = \infty$, then $\{X_n\}_{n\geqslant 1}$ is in the normal regime admitting measure $\mathbb{P}_\infty$ (i.e., there is no change).

For every fixed $\nu \geqslant 0$, the change-of-regime in the series $\{X_n\}_{n\geqslant 1}$ generates a new probability measure $\mathbb{P}_\nu$. We will now construct the pdf $p_\nu^{(n)}(\boldsymbol{X}_1^n)$ of $\mathbb{P}_\nu^{(n)}$ for $n \geqslant 1$ and $\nu \geqslant 0$ in the most general case. For the sake of brevity, we will omit the superscript and write $p_\nu(\boldsymbol{X}_1^n)$.

For $1 \leqslant i \leqslant j$, let $\boldsymbol{X}_i^j = (X_i, X_{i+1}, \ldots, X_j)$, that is, $\boldsymbol{X}_i^j$ is a sample of $j - i + 1$ successive observations indexed from i through j. Hence, if the sample $\boldsymbol{X}_1^n = (X_1, X_2, \ldots, X_n)$ is observed, then $\boldsymbol{X}_1^k = (X_1, \ldots, X_k)$ is the vector of the first k observations in this sample and $\boldsymbol{X}_{k+1}^n = (X_{k+1}, \ldots, X_n)$ is the vector of the rest of the observations in the sample, from $k+1$ to n.

First, suppose ν is deterministic unknown. This is the main assumption of the minimax approach. To get density $p_\nu(\boldsymbol{X}_1^n)$, observe that by the Bayes rule $p_\infty(\boldsymbol{X}_1^n) = p_\infty(\boldsymbol{X}_1^\nu) \times p_\infty(\boldsymbol{X}_{\nu+1}^n|\boldsymbol{X}_1^\nu)$ and $p_0(\boldsymbol{X}_1^n) = p_0(\boldsymbol{X}_1^\nu) \times p_0(\boldsymbol{X}_{\nu+1}^n|\boldsymbol{X}_1^\nu)$, whence by combining the first factor of the pre-change density, $p_\infty(\boldsymbol{X}_1^n)$, with the second one of the post-change density, $p_0(\boldsymbol{X}_1^n)$, we obtain $p_\nu(\boldsymbol{X}_1^n) = p_\infty(\boldsymbol{X}_1^\nu) \times p_0(\boldsymbol{X}_{\nu+1}^n|\boldsymbol{X}_1^\nu)$, or, after some more algebra using the Bayes rule,

$$p_\nu(\boldsymbol{X}_1^n) = \left(\prod_{j=1}^{\nu} p_\infty^{(j)}(X_j|\boldsymbol{X}_1^{j-1})\right) \times \left(\prod_{j=\nu+1}^{n} p_0^{(j)}(X_j|\boldsymbol{X}_1^{j-1})\right), \tag{1}$$

where $p_\infty^{(j)}(X_j|\boldsymbol{X}_1^{j-1})$ and $p_0^{(j)}(X_j|\boldsymbol{X}_1^{j-1})$ are the conditional densities of the j-th observation,

X_j, given the past information $\boldsymbol{X}_1^{j-1}$, $j \geqslant 1$. Note that in general these densities depend on j. Hereafter it is understood that $\prod_{j=k+1}^{n} p_d^{(j)}(X_j|\boldsymbol{X}_1^{j-1}) = 1$ for $k \geqslant n$.

Model (1) is very general: it does not require the observations to be independent or homogeneous. Suppose now that $\{X_n\}_{n\geqslant 1}$ are *independent* and such that $X_1, \ldots, X_\nu$ are each distributed according to a common density $f(x)$, while $X_{\nu+1}, X_{\nu+2}, \ldots$ each follow a common density $g(x) \not\equiv f(x)$. This is the simplest and most prevalent case. From now on it will be referred to as the *iid case*, or the *iid model*. In this case, model (1) reduces to

$$p_\nu(\boldsymbol{X}_1^n) = \left(\prod_{j=1}^{\nu} f(X_j) \right) \times \left(\prod_{j=\nu+1}^{n} g(X_j) \right), \tag{2}$$

and it will be referenced repeatedly throughout the paper.

If the change-point, ν, is random, which is the ground assumption of the Bayesian approach, then any change-point model has to be supplied with a change-point's *prior distribution*. To this end, let $\pi_0 = \mathbb{P}(\nu \leqslant 0)$ and $\pi_n = \mathbb{P}(\nu = n|\boldsymbol{X}_1^n)$, $n \geqslant 1$, and observe that the series $\{\pi_n\}_{n\geqslant 0}$ is $\{\mathcal{F}_n\}$-adapted. That is, the probability of the change occurring at time instance $\nu = k$ depends on $\boldsymbol{X}_1^k$, the observations' history accumulated up to (and including) time moment $k \geqslant 1$. With the so defined prior distribution one can describe very general change-point models, including those that assume ν is a $\{\mathcal{F}_n\}$-adapted stopping time; see Moustakides (2008).

To conclude this section, we note that when the probability series $\{\pi_n\}_{n\geqslant 0}$ depends on the observed data $\{X_n\}_{n\geqslant 1}$, it is argumentative whether $\{\pi_n\}_{n\geqslant 0}$ can be referred to as the change-point's *prior* distribution: it can just as well be viewed as the change-point's *a posteriori* distribution. However, a deeper discussion of this subject is out of scope to this paper, and from now on, we will assume that $\{\pi_n\}_{n\geqslant 0}$ do not depend on $\{X_n\}_{n\geqslant 1}$, in which case it represents the "true" prior distribution.

3. Overview of Optimality Criteria

3.1 Bayesian Formulation

The signature assumption of the Bayesian formulation is that the change-point is a random variable with a prior distribution. This is instrumental in certain applications (see, e.g., Shiryaev, 2006, 2010 and Tartakovsky and Veeravalli, 2005), but mostly of interest since the limiting versions of Bayesian solutions lead to optimal or asymptotically optimal procedures in more practical minimax problems.

Let $\{\pi_k\}_{k\geqslant 0}$ be a prior distribution of the change-point, ν, where $\pi_0 = \mathbb{P}(\nu \leqslant 0)$ and $\pi_k = \mathbb{P}(\nu = k)$ for $k \geqslant 1$. From the Bayesian point of view, the risk of sounding a false alarm is reasonable to measure by the Probability of False Alarm (PFA), which is defined as

$$\mathrm{PFA}^\pi(T) = \mathbb{P}^\pi(T \leqslant \nu) = \sum_{k=1}^{\infty} \pi_k \mathbb{P}_k(T \leqslant k), \tag{3}$$

where $\mathbb{P}^\pi(\mathcal{A}) = \sum_{k=0}^{\infty} \pi_k \mathbb{P}_k(\mathcal{A})$ and the π in the superscript emphasizes the dependence on the prior distribution. Note that summation in (3) is over $k \geqslant 1$ since by convention $\mathbb{P}_k(T \geqslant 1) = 1$, so that $\mathbb{P}_k(T \leqslant 0) = 0$. The most popular and practically reasonable way to benchmark the detection delay is through the Average Detection Delay (ADD), which is defined as

$$\mathrm{ADD}^\pi(T) = \mathbb{E}^\pi[T - \nu | T > \nu] = \mathbb{E}^\pi[(T-\nu)^+]/\mathbb{P}^\pi(T > \nu), \tag{4}$$

where hereafter $x^+ = \max\{0, x\}$ and $\mathbb{E}^\pi$ denotes expectation with respect to $\mathbb{P}^\pi$.

We now formally define the notion of Bayesian optimality. Let $\Delta_\alpha = \{T\colon \text{PFA}^\pi(T) \leqslant \alpha\}$ be the class of detection procedures (stopping times) for which the PFA does not exceed a preset (desired) level $\alpha \in (0,1)$. Then under the Bayesian approach one's aim is to

$$\text{find } T_{\text{opt}} \in \Delta_\alpha \text{ such that } \text{ADD}^\pi(T_{\text{opt}}) = \inf_{T\in\Delta_\alpha} \text{ADD}^\pi(T) \text{ for every } \alpha \in (0,1). \tag{5}$$

For the iid model (2) and under the assumption that the change-point ν has a *geometric* prior distribution this problem was solved by Shiryaev (1961, 1963, 1978). Specifically, Shiryaev assumed that ν is distributed according to the zero-modified geometric distribution

$$\mathbb{P}(\nu < 0) = \pi \;\text{ and }\; \mathbb{P}(\nu = n) = (1-\pi)p(1-p)^n, \;\; n \geqslant 0, \tag{6}$$

where $\pi \in [0,1)$ and $p \in (0,1)$. This is equivalent to choosing the series $\{\pi_n\}_{n\geqslant 0}$ as $\pi_0 = \mathbb{P}(\nu \leqslant 0) = \pi + (1-\pi)p$ and $\pi_n = \mathbb{P}(\nu = n) = (1-\pi)p(1-p)^n$, $n \geqslant 1$.

Observe now that if $\alpha \geqslant 1-\pi$, then problem (5) can be solved by simply stopping right away. This clearly is a trivial solution, since for this strategy the ADD is exactly zero, and $\text{PFA}^\pi(T) = \mathbb{P}(\nu > 0) = 1-\pi$, so that the constraint $\text{PFA}^\pi(T) \leqslant \alpha$ is satisfied. Therefore, assume that $\alpha < 1-\pi$ and in this case, Shiryaev (1961, 1963, 1978) proved that the optimal detection procedure is based on testing the posterior probability of the change currently being in effect, $\mathbb{P}(\nu < n|\mathcal{F}_n)$, against a certain detection threshold. The procedure stops as soon as $\mathbb{P}(\nu < n|\mathcal{F}_n)$ exceed the threshold. This strategy is known as the Shiryaev procedure. To guarantee its strict optimality the detection threshold should be set so as to guarantee that the PFA is exactly equal to the selected level α, which is rarely possible.

The Shiryaev procedure will play an important role in the sequel when considering non-Bayes criteria. It is more convenient to express Shiryaev's procedure through the average likelihood ratio (LR) statistic

$$R_{n,p} = \frac{\pi}{(1-\pi)p}\prod_{j=1}^{n}\left(\frac{\Lambda_j}{1-p}\right) + \sum_{k=1}^{n}\prod_{j=k}^{n}\left(\frac{\Lambda_j}{1-p}\right), \tag{7}$$

where $\Lambda_n = g(X_n)/f(X_n)$ is the "instantaneous" LR for the n-th data point, X_n. Indeed, by using the Bayes rule, one can show that

$$\mathbb{P}(\nu < n|\mathcal{F}_n) = \frac{R_{n,p}}{R_{n,p} + 1/p}, \tag{8}$$

whence it is readily seen that "thresholding" the posterior probability $\mathbb{P}(\nu < n|\mathcal{F}_n)$ is the same as "thresholding" the process $\{R_{n,p}\}_{n\geqslant 1}$. Therefore, the Shiryaev detection procedure has the form

$$T_{\text{S}}(A) = \inf\{n \geqslant 1\colon R_{n,p} \geqslant A\}, \tag{9}$$

and if $A = A_\alpha$ can be selected in such a way that the PFA is exactly equal to α, i.e., $\text{PFA}^\pi(T_{\text{S}}(A_\alpha)) = \alpha$, then it is strictly optimal in the class $\Delta(\alpha)$, that is, $\inf_{T\in\Delta(\alpha)} \text{ADD}^\pi(T) = \text{ADD}^\pi(T_{\text{S}}(A_\alpha))$ for any $0 < \alpha < 1-\pi$. Note that Shiryaev's statistic $R_{n,p}$ can be rewritten in the recursive form

$$R_{n,p} = (1 + R_{n-1,p})\frac{\Lambda_n}{1-p}, \;\; n \geqslant 1, \;\; \text{with } R_{0,p} = \frac{\pi}{(1-\pi)p}. \tag{10}$$

We also note that (7) and (8) remain true under the geometric prior distribution (6) even in the general non-iid case (1), with $\Lambda_n = g(X_n|\boldsymbol{X}_1^{n-1})/f(X_n|\boldsymbol{X}_1^{n-1})$. However, in order for the recursion (10) to hold in this case, $\{\Lambda_n\}_{n\geqslant 1}$ should be independent of the change-point.

As $p \to 0$, where p is the parameter of the geometric prior (6), the Shiryaev detection statistic (10) converges to what is known as the *Shiryaev–Roberts (SR) detection statistic*. The latter is the basis for the so-called *SR procedure*. As we will see, the SR procedure is a "bridge" between all four different approaches to change-point detection mentioned above.

For a general asymptotic Bayesian change-point detection theory in discrete time see Tartakovsky and Veeravalli (2005). Specifically, this work addresses the Bayesian approach assuming merely that the prior distribution is independent of the observations, and the overall conclusion is twofold: *a*) the Shiryaev procedure is asymptotically (as $\alpha \to 0$) optimal in a very broad class of change-point models and prior distributions, and *b*) depending on the behavior of the prior distribution at the right tail, the SR procedure may or may not be asymptotically optimal. Specifically, if the tail is exponential, the SR procedure is not asymptotically optimal, though it is asymptotically optimal if the tail is heavy. When the prior distribution is arbitrary and depends on the observations, we are not aware of any strict or asymptotic optimality results.

3.2 Generalized Bayesian Formulation

The generalized Bayesian approach is the limiting case of the Bayesian formulation, presented in the preceding section. Specifically, in the generalized Bayesian approach the change-point ν is assumed to be a "generalized" random variable with a uniform (improper) prior distribution.

First, return to the Bayesian constrained minimization problem (5). Specifically, consider the iid model (2) and assume that the change-point ν is distributed according to zero-modified geometric distribution (6). Then the Shiryaev procedure defined in (10) and (9) is optimal if the threshold $A = A_\alpha$ is chosen so that $\mathrm{PFA}^\pi(T_S(A_\alpha)) = \alpha$. Suppose now that $\pi = 0$ and $p \to 0$; this is turning the geometric prior (6) to an improper uniform distribution. It can be seen that in this case $\{R_{n,p}\}_{n\geqslant 0}$ becomes $\{R_{n,0}\}_{n\geqslant 0}$, where $R_{0,0} = 0$ and $R_{n,0} = (1+R_{n-1,0})\,\Lambda_n$, $n \geqslant 1$ with $\Lambda_n = g(X_n)/f(X_n)$. The limit $\{R_{n,0}\}_{n\geqslant 0}$ is known as the SR statistic, and is customarily denoted as $\{R_n\}_{n\geqslant 0}$, i.e., $R_n = R_{n,0}$ for all $n \geqslant 0$; in particular, note that $R_0 = 0$.

Next, when $\pi = 0$ and $p \to 0$ it can also be shown that

$$\frac{\mathbb{P}(T > \nu)}{p} \to \mathbb{E}_\infty[T] \quad \text{and} \quad \frac{\mathbb{E}[(T-\nu)^+]}{p} \to \sum_{k=0}^{\infty} \mathbb{E}_k[(T-k)^+], \tag{11}$$

where T is an arbitrary stopping time. As a result, one may conjecture that the SR procedure minimizes the *Relative Integral Average Detection Delay* (RIADD)

$$\mathrm{RIADD}(T) = \frac{\sum_{k=0}^{\infty} \mathbb{E}_k[(T-k)^+]}{\mathbb{E}_\infty[T]} \tag{12}$$

over all detection procedures for which the *Average Run Length (ARL) to false alarm*, $\mathbb{E}_\infty[T]$, is no less than $\gamma > 1$, an *a priori* set level.

Let

$$\Delta(\gamma) = \{T:\ \mathbb{E}_\infty[T] \geqslant \gamma\}, \tag{13}$$

be the class of detection procedures (stopping times) for which the ARL to false alarm $\mathbb{E}_\infty[T]$ is "no worse" than $\gamma > 1$. Then under the generalized Bayesian formulation one's goal is to

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathrm{RIADD}(T_{\mathrm{opt}}) = \inf_{T\in\Delta(\gamma)} \mathrm{RIADD}(T) \text{ for every } \gamma > 1. \tag{14}$$

We have already hinted that this problem is solved by the SR procedure. This was formally demonstrated by Pollak and Tartakovsky (2009b) in the discrete-time iid case, and by Shiryaev (1963) and Feinberg and Shiryaev (2006) in continuous time for detecting a shift in the mean of a Brownian motion.

We conclude this subsection with two remarks. First, observe that if the assumption $\pi = 0$ is replaced with $\pi = rp$, where $r \geqslant 0$ is a fixed number, then, as $p \to 0$, the Shiryaev statistic $\{R_{n,p}\}_{n\geqslant 0}$ converges to $\{R^r_n\}_{n\geqslant 0}$, where $R^r_n = (1 + R^r_{n-1})\,\Lambda_n$, $n \geqslant 1$ with $R^r_0 = r \geqslant 0$. This is the so-called *Shiryaev–Roberts–r (SR–r) detection statistic*, and it is the basis for the SR–r detection procedure that starts from an arbitrary deterministic point r. This procedure is due to Moustakides et al. (2011). The SR–r procedure possesses certain minimax properties (cf. Polunchenko and Tartakovsky, 2010 and Tartakovsky and Polunchenko, 2010). We will discuss this procedure at greater length later.

Secondly, though the generalized Bayesian formulation is the limiting (as $p \to 0$) case of the Bayesian approach, it may also be equivalently re-interpreted as a completely different approach – *multi-cyclic disorder detection in a stationary regime*. We will consider this approach in Subsection 3.4.

3.3 Minimax formulation

Contrary to the Bayesian formulation the minimax approach posits that the change-point is an unknown not necessarily random number. Even if it is random its distribution is unknown. The minimax approach has multiple optimality criteria.

First minimax theory is due to Lorden (1971) who proposed to measure the risk of raising a false alarm by the ARL to false alarm $\mathbb{E}_\infty[T]$. As far as the risk associated with detection delay is concerned, Lorden suggested to use the "worst-worst-case" ADD defined as

$$\mathrm{ESADD}(T) = \sup_{0\leqslant\nu<\infty}\left\{\operatorname{ess\,sup}\mathbb{E}_\nu[(T-\nu)^+|\mathcal{F}_\nu]\right\}.$$

Lorden's minimax optimization problem seeks to

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathrm{ESADD}(T_{\mathrm{opt}}) = \inf_{T\in\Delta(\gamma)} \mathrm{ESADD}(T) \text{ for every } \gamma > 1, \qquad (15)$$

where $\Delta(\gamma)$ is the class of detection procedures with the lower bound γ on the ARL to false alarm defined in (13).

For the iid scenario (2), Lorden (1971) showed that Page's (1954) Cumulative Sum (CUSUM) procedure is first-order asymptotically minimax as $\gamma \to \infty$. For any $\gamma > 1$, this problem was solved by Moustakides (1986), who showed that CUSUM is exactly optimal (see also Ritov (1990) who reestablished Moustakides' (1986) finding using a different decision-theoretic argument).

Though the strict ESADD(T)-optimality of the CUSUM procedure is a strong result, it is more natural to construct a procedure that minimizes the average (conditional) detection delay, $\mathbb{E}_\nu[T-\nu|T>\nu]$, for all $\nu \geqslant 0$ simultaneously. As no such uniformly optimal procedure is possible, Pollak (1985) suggested to revise Lorden's version of minimax optimality by replacing ESADD(T) with

$$\mathrm{SADD}(T) = \sup_{0\leqslant\nu<\infty} \mathbb{E}_\nu[T-\nu|T>\nu],$$

the worst conditional expected detection delay. Thus, Pollak's version of the minimax optimization problem seeks to

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathrm{SADD}(T_{\mathrm{opt}}) = \inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) \text{ for every } \gamma > 1. \qquad (16)$$

It is our opinion that SADD(T) is better suited for practical purposes for two reasons. First, Lorden's criterion is effectively a double-minimax approach, and therefore, is overly pessimistic in the sense that SADD(T) $\leqslant$ ESADD(T). Second, it is directly connected to the conventional decision theoretic approach — the optimization problem (16) can be solved by finding the least favorable prior distribution. More specifically, since by the general decision theory the minimax solution corresponds to the (generalized) Bayesian solution with the least favorable prior distribution, it can be shown that $\sup_\pi \mathrm{ADD}^\pi(T) = \mathrm{SADD}(T)$, where $\mathrm{ADD}^\pi(T)$ is defined in (4). In addition, unlike Lorden's minimax problem (15), Pollak's minimax problem (16) is still not solved. For these reasons, from now on, when considering the minimax approach, we focus on Pollak's supremum ADD measure SADD(T). Some light as to the possible solution (in the iid case) is shed in the work of Polunchenko and Tartakovsky (2010); Tartakovsky and Polunchenko (2010), and Moustakides et al. (2011). A synopsis of the results is given in the sequel.

Yet another way to gauge the false alarm risk is through the worst local (conditional) probability of sounding a false alarm within a time "window" of a given length. As argued by Tartakovsky (2005, 2008), in many surveillance applications (e.g., target detection) this may be a better option than the ARL to false alarm: the latter is more global. Specifically, the concern is that for a generic detection procedure, T, the ARL to false alarm, $\mathbb{E}_\infty[T]$, is not an exhaustive measure of the false alarm risk, unless the $\mathbb{P}_\infty$-distribution of T is geometric (at least approximately); see Tartakovsky (2005, 2008). The geometric distribution is characterized entirely by a single parameter, which a) uniquely determines $\mathbb{E}_\infty[T]$, and b) is uniquely determined by $\mathbb{E}_\infty[T]$. For the iid model (2), Tartakovsky et al. (2008); Pollak and Tartakovsky (2009a) showed that under mild assumptions the $\mathbb{P}_\infty$-distribution of the stopping times associated with detection schemes from a certain class is asymptotically (as $\gamma \to \infty$) exponential with parameter $1/\mathbb{E}_\infty[T]$; the convergence is in the L^p sense, where $p \geqslant 1$. The class includes all of the most popular procedures. Hence, for the iid model (2), the ARL to false alarm is an acceptable measure of the false alarm rate. However, for a general non-iid model this is not necessarily true. Hence, alternative measures of the false alarm rate are in order. As a result, if T is geometric, one can evaluate $\mathbb{P}_\infty(k < T \leqslant k+m | T > k)$ for any $k \geqslant 0$ (in fact, for all $k \geqslant 0$ at once). Specifically, let

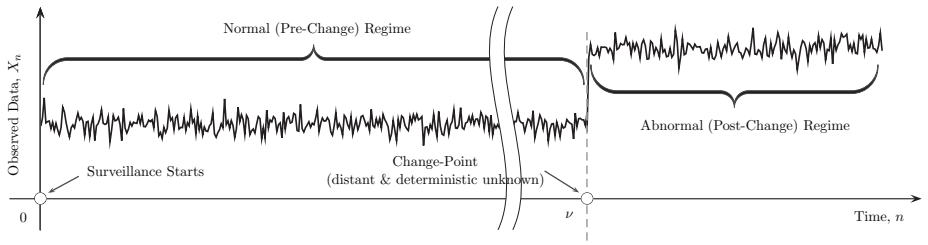
$$\Delta_\alpha^m = \left\{ T : \sup_{k \geqslant 0} \mathbb{P}_\infty(k < T \leqslant k+m | T > k) \leqslant \alpha \right\}, \tag{17}$$

be the class of detection procedures for which $\mathbb{P}_\infty(k < T \leqslant k+m | T > k)$, the conditional probability of raising a false alarm inside a sliding window of $m \geqslant 1$ observations is "no worse" than a certain *a priori* chosen level $\alpha \in (0,1)$. The size of the window m may either be fixed or go to infinity as $\alpha \to 0$.

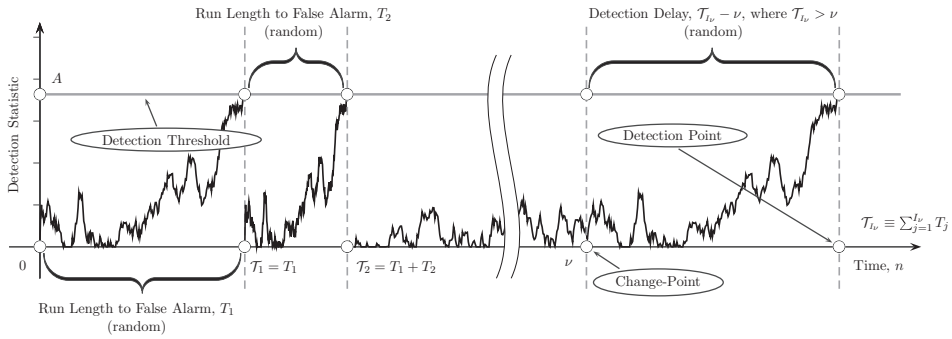
As argued by Tartakovsky (2005), in general, $\sup_k \mathbb{P}_\infty(k < T \leqslant k+m | T > k) \leqslant \alpha$ is a *stronger* condition than $\mathbb{E}_\infty[T] \geqslant \gamma$. Hence, in general, $\Delta_\alpha^m \subset \Delta(\gamma)$. See also Tartakovsky (2009b). For a specific example where the optimization problem (16) is solved in the class (17) see Polunchenko and Tartakovsky (2012).

3.4 Multi-cyclic detection of a disorder in a stationary regime

Consider a context in which it is of utmost importance to detect the change as quickly as possible, even at the expense of raising many false alarms (using a repeated application of the same stopping rule) before the change occurs. This is equivalent to saying that the change-point ν is substantially larger than the tolerable level of false alarms γ. That is, the change "strikes" in a distant future and is preceded by a *stationary flow of false alarms*. This scenario is shown in Figure 1. As one can see, the ARL to false alarm in this case



(a) An example of the behavior of a process of interest as exhibited through the series of observations $\{X_n\}_{n\geqslant 1}$.



(b) An example of the behavior of the detection statistic when the decision to terminate surveillance is made *past* the change-point.

Figure 1: Multi-cyclic change-point detection in a stationary regime.

is the mean time between (consecutive) false alarms, and therefore may be thought of the false alarm rate (or frequency).

As argued by Pollak and Tartakovsky (2009b), the multi-cyclic approach is instrumental in many surveillance applications, in particular in the areas concerned with intrusion/anomaly detection, e.g., cybersecurity and particularly detection of attacks in computer networks.

Formally, let $T_1, T_2, \ldots$ denote sequential independent repetitions of the same stopping time T, and let $\mathcal{T}_{(j)} = T_{(1)} + T_{(2)} + \cdots + T_{(j)}$ be the time of the j-th alarm. Define $I_\nu = \min\{j \geqslant 1 \colon \mathcal{T}_{(j)} > \nu\}$. Put otherwise, $\mathcal{T}_{(I_\nu)}$ is the time of detection of the true change that occurs at the time instant ν after $I_\nu - 1$ false alarms have been raised. Write

$$\mathrm{STADD}(T) = \lim_{\nu\to\infty} \mathbb{E}_\nu[\mathcal{T}_{(I_\nu)} - \nu]$$

for the limiting value of the ADD that we will refer to as the *stationary ADD* (STADD).

We now formally state the multi-cyclic change-point detection problem:

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathrm{STADD}(T_{\mathrm{opt}}) = \inf_{T\in\Delta(\gamma)} \mathrm{STADD}(T) \text{ for every } \gamma > 1 \qquad (18)$$

(among all multi-cyclic procedures).

For the iid model (2), this problem was solved by Pollak and Tartakovsky (2009b), who showed that the solution is the multi-cyclic SR procedure by arguing that $\mathrm{STADD}(T) \equiv \mathrm{RIADD}(T)$ defined in (12). This suggests that the optimal solution of the problem of multi-cyclic change-point detection in a stationary regime is completely equivalent to the solution of the generalized Bayesian problem. The exact result is stated in the next section.

4. Optimality Properties of the Shiryaev–Roberts Detection Procedure

From now on we will confine ourselves to the iid scenario (2), i.e., assume that *a*) the observations $\{X_n\}_{n\geqslant 1}$ are independent throughout their history, and *b*) $X_1, \ldots, X_\nu$ are distributed

according to a common known pdf $f(x)$ and $X_{\nu+1}, X_{\nu+2}, \ldots$ are distributed according to a common pdf $g(x) \not\equiv f(x)$, also known.

Let $\mathcal{H}_k\colon \nu = k$ for $0 \leqslant k < \infty$ and $\mathcal{H}_\infty\colon \nu = \infty$ be, respectively, the hypotheses that the change takes place at the time moment $\nu = k$, $k \geqslant 0$, and that no change ever occurs. The densities of the sample $\boldsymbol{X}_1^n = (X_1, \ldots, X_n)$, $n \geqslant 1$ under these hypotheses are given by

$$p(\boldsymbol{X}_1^n|\mathcal{H}_\infty) = \prod_{j=1}^{n} f(X_j), \quad \text{and} \quad p(\boldsymbol{X}_1^n|\mathcal{H}_k) = \prod_{j=1}^{k} f(X_j) \prod_{j=k+1}^{n} g(X_j) \quad \text{for } k < n,$$

and $p(\boldsymbol{X}_1^n|\mathcal{H}_\infty) = p(\boldsymbol{X}_1^n|\mathcal{H}_k)$ for $k \geqslant n$, so that the corresponding LR is

$$\Lambda_n^k = \frac{p(\boldsymbol{X}_1^n|\mathcal{H}_k)}{p(\boldsymbol{X}_1^n|\mathcal{H}_\infty)} = \prod_{j=k+1}^{n} \Lambda_j \quad \text{for} \quad k < n,$$

where $\Lambda_n = g(X_n)/f(X_n)$ is the "instantaneous" LR for the n-th observation X_n.

To decide in favor of one of the hypotheses $\mathcal{H}_k$ or $\mathcal{H}_\infty$, the likelihood ratios are "fed" to an appropriate sequential detection procedure, which is chosen according to the particular version of the optimization problem. In this section we are interested in the generalized Bayesian problem (14) and in the multi-cyclic disorder detection in a stationary regime (18). We have already remarked that for the iid model the SR procedure solves both of these problems. We preface the presentation of the exact results with the introduction of the SR procedure.

The SR procedure is due to the independent work of Shiryaev (1961, 1963) and that of Roberts (1966). Specifically, Shiryaev considered the problem of detecting a change in the drift of a Brownian motion; Roberts focused on the case of detecting a shift in the mean of an iid Gaussian sequence. The name "Shiryaev–Roberts" was coined by Pollak (1985). See Pollak (2009) for a brief account of the SR procedure's history.

Formally, the SR procedure is defined as the stopping time

$$\mathcal{S}_A = \inf\{n \geqslant 1\colon R_n \geqslant A\}, \tag{19}$$

where $A > 0$ is the detection threshold, and

$$R_n = (1 + R_{n-1})\,\Lambda_n, \quad n \geqslant 1 \quad \text{with} \quad R_0 = 0 \tag{20}$$

is the SR detection statistic. As usual, we set $\inf\{\varnothing\} = \infty$, i.e., $\mathcal{S}_A = \infty$ if R_n never crosses A.

Recall first that $R_n = \lim_{p\to 0} R_{n,p}$, where $R_{n,p}$ is the Shiryaev statistic given by recursion (10). Recall also that the limiting relations (11) hold. These allow us to conjecture that the SR procedure is optimal in the generalized Bayesian sense. In addition, since the RIADD is equal to the STADD of the multi-cyclic procedure, the repeated SR procedure should be optimal for detecting distant changes. The exact result is given next.

Theorem 1 (Pollak and Tartakovsky, 2009b)**.** *Let $\mathcal{S}_A$ be the SR procedure defined by* (19) *and* (20)*. Suppose the detection threshold $A = A_\gamma$ is selected from the equation $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}] = \gamma$, where $\gamma > 1$ is the desired level of the ARL to false alarm.*

(i) *Then the SR procedure $\mathcal{S}_{A_\gamma}$ minimizes* $\mathrm{RIADD}(T) = \sum_{k=0}^{\infty} \mathbb{E}_k[(T-k)^+]/\,\mathbb{E}_\infty[T]$ *over all stopping times T that satisfy $\mathbb{E}_\infty[T] \geqslant \gamma$, i.e.,* $\mathrm{RIADD}(\mathcal{S}_{A_\gamma}) = \inf_{T\in\Delta(\gamma)} \mathrm{RIADD}(T)$ *for every $\gamma > 1$.*

(ii) *Since* $\mathrm{RIADD}(T) \equiv \mathrm{STADD}(T)$ *for any stopping time T, the SR procedure $\mathcal{S}_{A_\gamma}$ minimizes the stationary average detection delay among all multi-cyclic procedures in the class $\Delta(\gamma)$, i.e.,* $\mathrm{STADD}(\mathcal{S}_{A_\gamma}) = \inf_{T\in\Delta(\gamma)} \mathrm{STADD}(T)$ *for every $\gamma > 1$.*

It is worth noting that the ARL to false alarm of the SR procedure satisfies the inequality $\mathbb{E}_\infty[\mathcal{S}_A] \geqslant A$ for all $A > 0$, which can be easily obtained by noticing that $R_n - n$ is a $\mathbb{P}_\infty$-martingale with mean zero. Also, asymptotically (as $A \to \infty$), $\mathbb{E}_\infty[\mathcal{S}_A] \approx A/\zeta$, where the constant $0 < \zeta < 1$ is given by (28) below (see Pollak, 1987). Hence, setting $A_\gamma = \gamma\zeta$ yields $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}] \approx \gamma$, as $\gamma \to \infty$.

5. Optimal and Nearly Optimal Minimax Detection Procedures

In this section, we will be concerned exclusively with the minimax problem in Pollak's setting (16), assuming that the change-point ν is deterministic unknown. As of today, this problem is not solved in general. As has been indicated earlier, the usual way around this is to consider it asymptotically by allowing the ARL to false alarm $\gamma \to \infty$. The hope is to design such procedure $T^* \in \Delta(\gamma)$ that $\mathrm{SADD}(T^*)$ and the (unknown) optimum $\inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T)$ will be in some sense "close" to each other in the limit, as $\gamma \to \infty$. To this end, the following three different types of asymptotic optimality are usually distinguished.

Definition 1 (First-Order Asymptotic Optimality)**.** A procedure $T^* \in \Delta(\gamma)$ is said to be *first-order asymptotically* optimal in the class $\Delta(\gamma)$ if $\mathrm{SADD}(T^*) = \inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T)[1 + o(1)]$, as $\gamma \to \infty$, where from now on $o(1) \to 0$, as $\gamma \to \infty$.

Definition 2 (Second-Order Asymptotic Optimality)**.** A procedure $T^* \in \Delta(\gamma)$ is said to be *second-order asymptotically* optimal in the class $\Delta(\gamma)$ if $\mathrm{SADD}(T^*) - \inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) = O(1)$, as $\gamma \to \infty$, where $O(1)$ stays bounded, as $\gamma \to \infty$.

Definition 3 (Third-Order Asymptotic Optimality)**.** A procedure $T^* \in \Delta(\gamma)$ is said to be *third-order asymptotically* optimal in the class $\Delta(\gamma)$ if $\mathrm{SADD}(T^*) - \inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) = o(1)$, as $\gamma \to \infty$.

5.1 The Shiryaev–Roberts–Pollak procedure

The question of what procedure minimizes Pollak's measure of detection delay $\mathrm{SADD}(T)$ is an open issue. As an attempt to resolve the issue, Pollak (1985) proposed to "tweak" the SR procedure (19). This led to the new procedure that we will refer to as the Shiryaev–Roberts–Pollak (SRP) procedure. To facilitate the presentation of the latter, we first explain the heuristics.

As known from the general decision theory (see, e.g., Ferguson, 1967, Theorem 2.11.3), an $\mathcal{F}_n$-adapted stopping time T solves (16) if *a*) T is an extended Bayes rule, *b*) it is an equalizer, and *c*) it satisfies the false alarm constraint with equality. A procedure is said to be an equalizer if its conditional risk (which we measure through $\mathbb{E}_\nu[T - \nu | T > \nu]$) is constant for all $\nu \geqslant 0$, that is, $\mathbb{E}_0[T] = \mathbb{E}_\nu[T - \nu | T > \nu]$ for all $\nu \geqslant 1$. Of the three conditions the one that requires T to be an equalizer poses the most challenge. Pollak (1985) came up with an elegant solution.

It turns out that the sequence $\mathbb{E}_\nu[\mathcal{S}_A - \nu | \mathcal{S}_A > \nu]$ indexed by ν eventually *stabilizes*, i.e., it remains the same for all sufficiently large ν. This happens because the SR detection statistic enters the quasi-stationary mode, which means that the conditional distribution $\mathbb{P}_\infty(R_n \leqslant x | \mathcal{S}_A > n)$ no longer changes with time. If one could get to the quasi-stationary mode immediately, then the resulting procedure would have the same expected conditional detection delay for all $\nu \geqslant 0$, i.e., it would be the equalizer. Thus, Pollak's (1985) idea was to start the SR detection statistic $\{R_n\}_{n\geqslant 0}$, defined in (20), not from zero ($R_0 = 0$), but from a random point $R_0 = R_0^Q$, where R_0^Q is sampled from the *quasi-stationary distribution* of the

SR statistic under the hypothesis $\mathcal{H}_\infty$ (which is a Markov Harris-recurrent process under $\mathcal{H}_\infty$). Specifically, the quasi-stationary cdf, $Q_A(x)$, is defined as

$$Q_A(x) = \lim_{n\to\infty} \mathbb{P}_\infty(R_n \leqslant x | \mathcal{S}_A > n). \tag{21}$$

Therefore, the SRP procedure is defined as the stopping time

$$\mathcal{S}_A^Q = \inf\{n \geqslant 1 \colon R_n^Q \geqslant A\}, \tag{22}$$

where $A > 0$ is a detection threshold, and

$$R_n^Q = (1 + R_{n-1}^Q)\,\Lambda_n, \quad n \geqslant 1, \quad R_0^Q \sim Q_A(x) \tag{23}$$

is the detection statistic.

We reiterate that, by design, the SRP procedure (22) and (23) is an equalizer: it delivers the same conditional average detection delay for any change-point $\nu \geqslant 0$, that is, $\mathbb{E}_0[\mathcal{S}_A^Q] = \mathbb{E}_\nu[\mathcal{S}_A^Q - \nu | \mathcal{S}_A^Q > \nu]$ for all $\nu \geqslant 1$. Pollak (1985) was able to demonstrate that the SRP procedure is third-order asymptotically optimal with respect to SADD(T). We now state his result.

Theorem 2 (Pollak, 1985)**.** *Let* $\mathbb{E}_0[(\log \Lambda_1)^+] < \infty$. *Suppose the detection threshold, A, of the SRP procedure,* $\mathcal{S}_A^Q$*, is set to the solution,* A_γ*, of the equation* $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^Q] = \gamma$. *Then* $\mathrm{SADD}(\mathcal{S}_{A_\gamma}^Q) = \inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) + o(1)$, *as* $\gamma \to \infty$.

Recently, Tartakovsky et al. (2012) proved that $\mathbb{E}_0[\mathcal{S}_A^Q] = (1/I)[\log A + \varkappa - C_\infty] + o(1)$, as $A \to \infty$, provided $\mathbb{E}_0[\log \Lambda_1]^2 < \infty$, where $\varkappa$ is the limiting average overshoot in the one-sided sequential test, which is a subject of renewal theory (see, e.g., Woodroofe, 1982), and C_∞ is a constant that can be computed numerically (e.g., by Monte Carlo simulations). Both $\varkappa$ and C_∞ are formally defined in the next subsection, where we reiterate the exact result of Tartakovsky et al. (2012).

Note that for sufficiently large γ,

$$\mathbb{E}_\infty[\mathcal{S}_A^Q] \approx (A/\zeta) - \mu_Q, \quad \text{where} \quad \mu_Q = \int_0^A y\, dQ_A(y), \tag{24}$$

i.e., μ_Q is the mean of the quasi-stationary distribution, and ζ is a constant defined in (28) below. This approximation can be obtained by first noticing that for a fixed $R_0^Q = r$ the process $R_n^Q - r - n$ is a zero-mean $\mathbb{P}_\infty$-martingale, and then applying optional sampling theorem to this martingale as well as a renewal theoretic argument (cf. Tartakovsky et al., 2012).

5.2 The Shiryaev–Roberts–r procedure

Though the SRP procedure is practically appealing due to its third-order asymptotic optimality, it requires the knowledge of the quasi-stationary distribution (21) to implement. It is rare that this distribution can be expressed in a closed form; for examples where this is possible, see, e.g., Pollak (1985), Mevorach and Pollak (1991), Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010). As a result, the SRP procedure has not been used in practice.

To make the SRP procedure practical, Moustakides et al. (2011) proposed a numerical framework. More importantly, Moustakides et al. (2011) offered numerical evidence that there exist procedures that are uniformly better than the SRP procedure. Specifically, they

regard starting off the original SR procedure at a fixed (but specially designed) $R_0 = r$, $0 \leqslant r < A$, and defining the stopping time with this new deterministic initialization. Because of the importance of the starting point, they dubbed their procedure the SR–r procedure.

Formally, the SR–r procedure is defined as the stopping time

$$\mathcal{S}_A^r = \inf\{n \geqslant 1 \colon R_n^r \geqslant A\}, \tag{25}$$

where $A > 0$ is the detection threshold, and

$$R_n^r = (1 + R_{n-1}^r)\,\Lambda_n, \quad n \geqslant 1, \quad \text{with} \quad R_0^r = r \geqslant 0 \tag{26}$$

is the SR–r detection statistic.

Moustakides et al. (2011) show numerically that for certain values of the starting point, $R_0^r = r$, apparently, $\mathbb{E}_\nu[\mathcal{S}_{A_1}^r - \nu | \mathcal{S}_{A_1}^r > \nu]$ is strictly less than $\mathbb{E}_\nu[\mathcal{S}_{A_2}^Q - \nu | \mathcal{S}_{A_2}^Q > \nu]$ for all $\nu \geqslant 0$, where A_1 and A_2 are such that $\mathbb{E}_\infty[\mathcal{S}_{A_1}^r] = \mathbb{E}_\infty[\mathcal{S}_{A_2}^Q]$ (although the maximal expected delay is only slightly smaller for $\mathcal{S}_{A_1}^r$).

It turns out that using the ideas of Moustakides et al. (2011) we are able to design the initialization point $r = r(\gamma)$ in the SR–r procedure (25) so that this procedure is also third-order asymptotically optimal. In this respect, the average delay to detection at infinity $\mathrm{ADD}_\infty(\mathcal{S}_A^r) = \lim_{\nu\to\infty} \mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ plays the critical role. The following theorem, whose proof can be found in Polunchenko and Tartakovsky (2010), is important.

Theorem 3. *Let $\mathcal{S}_A^r$ be defined as in* (25) *and* (26)*, and let $A = A_\gamma$ be selected so that $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^r] = \gamma$. Then, for* every *$r \geqslant 0$,*

$$\inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) \geqslant \frac{r\,\mathbb{E}_0[\mathcal{S}_{A_\gamma}^r] + \sum_{\nu=0}^{\infty} \mathbb{E}_\nu[(\mathcal{S}_{A_\gamma}^r - \nu)^+]}{r + \mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^r]} = \mathcal{J}_{\mathrm{B}}(\mathcal{S}_{A_\gamma}^r). \tag{27}$$

Note that Theorem 3 suggests that if r can be chosen so that the SR–r procedure is an equalizer (i.e., $\mathbb{E}_0[\mathcal{S}_A^r] = \mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ for all $\nu \geqslant 0$), then it is *exactly* optimal. This is because the right-hand side in (27) is equal to $\mathbb{E}_0[\mathcal{S}_A^r]$, which, in turn, is equal to $\sup_\nu \mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu] = \mathrm{SADD}(\mathcal{S}_A^r)$. Therefore, we have the following corollary.

Corollary. *Let $A = A_\gamma$ be selected so that $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^r] = \gamma$. Assume that $r = r(\gamma)$ is chosen in such a way that the SR–r procedure $\mathcal{S}_{A_\gamma}^{r(\gamma)}$ is an equalizer. Then it is strictly minimax in the class $\Delta(\gamma)$, i.e.,* $\inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) = \mathrm{SADD}(\mathcal{S}_{A_\gamma}^{r(\gamma)})$.

Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010) used this Corollary to prove that the SR–r procedure with a specially designed $r = r_A$ is strictly optimal for two specific models. In general, Moustakides et al. (2011) conjecture that the SR–r procedure is third-order asymptotically minimax, and Tartakovsky et al. (2012) show that this conjecture is true. We will state the exact result after we introduce some additional notation.

Let $S_n = \log\Lambda_1 + \cdots + \log\Lambda_n$ and, for $a \geqslant 0$, introduce the one-sided stopping time $\tau_a = \inf\{n \geqslant 1 \colon S_n \geqslant a\}$. Let $\kappa_a = S_{\tau_a} - a$ be an overshoot (excess over the level a at stopping), and let

$$\varkappa = \lim_{a\to\infty} \mathbb{E}_0[\kappa_a], \quad \zeta = \lim_{a\to\infty} \mathbb{E}_0\left[e^{-\kappa_a}\right]. \tag{28}$$

The constants $\varkappa > 0$ and $0 < \zeta < 1$ depend on the model and can be computed numerically. Let $I = \mathbb{E}_0[\log\Lambda_1]$ denote the Kullback–Leibler information number, and

let $\tilde{V}_\infty = \sum_{j=1}^{\infty} e^{-S_j}$. Also, let R_∞ be a random variable that has the $\mathbb{P}_\infty$-limiting (stationary) distribution of R_n, as $n \to \infty$, i.e., $Q_{\mathrm{ST}}(x) = \lim_{n\to\infty} \mathbb{P}_\infty(R_n \leqslant x) = \mathbb{P}_\infty(R_\infty \leqslant x)$. Let

$$C_\infty = \mathbb{E}[\log(1 + R_\infty + \tilde{V}_\infty)] = \int_0^\infty \int_0^\infty \log(1 + x + y)\, dQ_{\mathrm{ST}}(x)\, d\tilde{Q}(y),$$

where $\tilde{Q}(y) = \mathbb{P}_0(\tilde{V}_\infty \leqslant y)$.

Theorem 4 (Tartakovsky et al., 2012)**.** *Let* $\mathbb{E}_0[\log \Lambda_1]^2 < \infty$ *and let* $\log \Lambda_1$ *be non-arithmetic. Then the following assertions hold.*

(i) $\inf_{T\in\Delta(\gamma)} \mathrm{SADD}(T) \geqslant (1/I)[\log(\gamma\zeta) + \varkappa - C_\infty] + o(1)$, *as* $\gamma \to \infty$.

(ii) *For any* $r \geqslant 0$,

$$\mathrm{ADD}_\infty(\mathcal{S}_A^r) = \mathbb{E}_0[\mathcal{S}_A^Q] = \frac{1}{I}(\log A + \varkappa - C_\infty) + o(1), \quad as \;\; A \to \infty. \tag{29}$$

(iii) *Furthermore, if in the SR–r procedure* $A = A_\gamma = \gamma\zeta$ *and the initialization point* $r = o(\gamma)$ *is selected so that* $\mathrm{SADD}(\mathcal{S}_A^r) = \mathrm{ADD}_\infty(\mathcal{S}_A^r)$, *then* $\mathbb{E}_\infty[\mathcal{S}_A^r] = \gamma(1 + o(1))$ *and* $\mathrm{SADD}(\mathcal{S}_A^r) = (1/I)[\log(\gamma\zeta) + \varkappa - C_\infty] + o(1)$, *as* $\gamma \to \infty$.

Hence, the SR–r procedure is third-order asymptotically optimal.

Also,

$$\mathrm{ADD}_0(\mathcal{S}_A^r) = \frac{1}{I}[\log A + \varkappa - C(r)] + o(1), \quad \text{as} \;\; A \to \infty, \tag{30}$$

where $C(r) = \mathbb{E}[\log(1 + r + \tilde{V}_\infty)]$. As we mentioned above, it is desirable to make the SR–r procedure to look like equalizer by choosing the head start r, which can be achieved by equalizing ADD_0 and ADD_∞. Comparing (29) and (30) we see that this property approximately holds when r is selected from the equation $C(r^*) = C_\infty$. This shows that asymptotically (as $\gamma \to \infty$) the "optimal" value r^* is a fixed number that does not depend on γ. Clearly, this observation is important since it allows us to design the initialization point effectively and make the resulting procedure approximately optimal.

It is worth mentioning that $\mathrm{SADD}(\mathcal{S}_A) = \mathrm{ADD}_0(\mathcal{S}_A) = (1/I)[\log A + \varkappa - C(0)] + o(1)$, as $A \to \infty$, is true for the conventional SR procedure that starts from zero. Therefore, the SR procedure is only second-order asymptotically optimal. For sufficiently large γ, the difference between the supremum ADD-s of the SR procedure and the optimized SR–r is given by $(C(0) - C_\infty)/I$, which can be quite large if the Kullback–Leibler information number I is small.

Note that similar to (24), for sufficiently large γ, we have $\mathbb{E}_\infty[\mathcal{S}_A^r] \approx (A/\zeta) - r$. For an example where distributions $Q_{\mathrm{ST}}(x)$ and $\tilde{Q}(x)$ and the constants $\varkappa$, ζ, C_∞, and $C(r)$ can be computed analytically see Polunchenko and Tartakovsky (2012).

References

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric Methods in Change-Point Problems*. Kluwer, Dordrecht.

Broemeling, L. D. and Tsurumi, H. (1987). *Econometrics and Structural Change*, volume 74 of *Statistics, textbooks and monographs*. CRC Press.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. Society for Industrial and Applied Mathematics, Philadelphia.

Cohen, A. (1987). *Biomedical Signal Processing*. CRC Press, Boca Raton, FL.

Feinberg, E. A. and Shiryaev, A. N. (2006). Quickest detection of drift change for Brownian motion in generalized Bayesian and minimax settings. *Statistics & Decisions*, 24(4):445–470.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

Fuh, C.-D. (2003). SPRT and CUSUM in hidden Markov models. *Annals of Statistics*, 31(3):942–977.

Fuh, C.-D. (2004). Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. *Annals of Statistics*, 32(5):2305–2339.

Kenett, R. S. and Zacks, S. (1998). *Modern Industrial Statistics: Design and Control of Quality and Reliability*. Duxbury Press, first edition.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B. Methodological*, 57(4):613–658.

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44:2917–2929.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908.

Mevorach, Y. and Pollak, M. (1991). A small sample size comparison of the Cusum and the Shiryayev-Roberts approaches to changepoint detection. *American Journal of Mathematical and Management Sciences*, 11:277–298.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control*. Wiley, seventh edition.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387.

Moustakides, G. V. (2008). Sequential change detection revisited. *Annals of Statistics*, 36(2):787–807.

Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2011). A numerical approach to performance analysis of quickest change-point detection procedures. *Statistica Sinica*, 21(2):571–596.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1):100–115.

Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Statistics*, 13(1):206–227.

Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15(2):749–779.

Pollak, M. (2009). The Shiryaev–Roberts changepoint detection procedure in retrospect - theory and practice. In *Proceedings of the 2nd International Workshop in Sequential Methodologies*, University of Technology of Troyes, Troyes, France.

Pollak, M. and Tartakovsky, A. G. (2009a). Asymptotic exponentiality of the distribution of first exit times for a class of markov processes with applications to quickest change detection. *Theory of Probability and Its Applications*, 53(3):430–442.

Pollak, M. and Tartakovsky, A. G. (2009b). Optimality properties of the Shiryaev–Roberts procedure. *Statistica Sinica*, 19:1729–1739.

Polunchenko, A. S. and Tartakovsky, A. G. (2010). On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *Annals of Statistics*, 38(6):3445–3457.

Polunchenko, A. S. and Tartakovsky, A. G. (2012). State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability*, 14(3):649–684.

Polunchenko, A. S., Tartakovsky, A. G., and Mukhopadhyay, N. (2012). Nearly optimal change-point detection with an application to cybersecurity. *Sequential Analysis*, 31(3):409–435.

Poor, H. V. and Hadjiliadis, O. (2008). *Quickest Detection*. Cambridge University Press.

Ritov, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Annals of Statistics*, 18(3):1464–1469.

Roberts, S. (1966). A comparison of some control chart procedures. *Technometrics*, 8(3):411–430.

Ryan, T. P. (2011). *Statistical Methods for Quality Improvement*. Wiley, third edition.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Bell Telephone Laboratories series. D. Van Nostrand Company, Inc., Princeton, New Jersey.

Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl.*, 2:795–799. Translation from Dokl. Akad. Nauk SSSR 138:1039–1042, 1961.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8(1):22–46.

Shiryaev, A. N. (1978). *Optimal Stopping Rules*. Springer-Verlag, New York.

Shiryaev, A. N. (2006). From "disorder" to nonlinear filtering and martingale theory. In Bolibruch, A., Osipov, Y., and Sinai, Y., editors, *Mathematical Events of the Twentieth Century*, pages 371–397. Springer Berlin Heidelberg.

Shiryaev, A. N. (2009). On the stochastic models and optimal methods in the quickest detection problems. *Theory of Probability and Its Applications*, 53(3):385–401.

Shiryaev, A. N. (2010). Quickest detection problems: Fifty years later. *Sequential Analysis*, 29:345–385.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer-Verlag, New York.

Tartakovsky, A. G. (1991). *Sequential Methods in the Theory of Information Systems*. Radio & Communications, Moscow, Russia.

Tartakovsky, A. G. (2005). Asymptotic performance of a multichart CUSUM test under false alarm probability constraint. In *Proceedings of the 2005 IEEE Conference on Decision and Control*, volume 44, pages 320–325.

Tartakovsky, A. G. (2008). Discussion on "Is average run length to false alarm always an informative criterion?" by Yajun Mei. *Sequential Analysis*, 27(4):396–405.

Tartakovsky, A. G. (2009a). Asymptotic optimality in Bayesian changepoint detection problems under global false alarm probability constraint. *Theory of Probability and Its Applications*, 53:443–466.

Tartakovsky, A. G. (2009b). Discussion on "Optimal sequential surveillance for finance, public health, and other areas" by Marianne Frisén. *Sequential Analysis*, 28(3):365–371.

Tartakovsky, A. G. and Moustakides, G. V. (2010). State-of-the-art in Bayesian changepoint detection. *Sequential Analysis*, 29(2):125–145.

Tartakovsky, A. G., Pollak, M., and Polunchenko, A. S. (2008). Asymptotic exponentiality of first exit times for recurrent Markov processes and applications to changepoint detection. In *Proceedings of the 2008 International Workshop on Applied Probability*, Compiégne, France.

Tartakovsky, A. G., Pollak, M., and Polunchenko, A. S. (2012). Third-order asymptotic optimality of the Generalized Shiryaev–Roberts changepoint detection procedures. *Theory of Probability and Its Applications*, 56(3):457–484.

Tartakovsky, A. G. and Polunchenko, A. S. (2010). Minimax optimality of the Shiryaev–Roberts procedure. In *Proceedings of the 5th International Workshop on Applied Probability*, Universidad Carlos III of Madrid, Spain.

Tartakovsky, A. G., Polunchenko, A. S., and Sokolov, G. (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):4–11.

Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B., and Kim, H. (2006). Detection of intrusions in information systems by sequential changepoint methods (with discussion). *Statistical Methodology*, 3(3):252–340.

Tartakovsky, A. G. and Veeravalli, V. V. (2005). General asymptotic Bayesian theory of quickest change detection. *Theory of Probability and Its Applications*, 49(3):458–497.

Wald, A. (1947). *Sequential Analysis*. J. Wiley & Sons, Inc., New York.

Wetherill, G. B. and Brown, D. W. (1991). *Statistical Process Control: Theory and Practice*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, third edition.

Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA.