State-of-the-Art in Sequential Change-Point Detection

Aleksey S. Polunchenko · **Alexander G. Tartakovsky**

Abstract We provide an overview of the state-of-the-art in the area of sequential change-point detection assuming discrete time and known pre- and post-change distributions. The overview spans over all major formulations of the underlying optimization problem, namely, Bayesian, generalized Bayesian, and minimax. We pay particular attention to the latest advances in each. Also, we link together the generalized Bayesian problem with multi-cyclic disorder detection in a stationary regime when the change occurs at a distant time horizon. We conclude with two case studies to illustrate the cutting edge of the field at work.

A.S. Polunchenko
Department of Mathematics, University of Southern California,
3620 S. Vermont Ave., KAP 108
Los Angeles, CA 90089-2532, USA
Tel.: +1-213-821-1892
E-mail: polunche@usc.edu

A.G. Tartakovsky
Department of Mathematics, University of Southern California,
3620 S. Vermont Ave., KAP 108
Los Angeles, CA 90089-2532, USA
Tel.: +1-213-740-2450, Fax: +1-213-740-2424
E-mail: tartakov@usc.edu

1 Introduction

Sequential change-point detection (or quickest change detection, or quickest "disorder" detection) is concerned with the design and analysis of techniques for *quickest* (on-line) detection of a change in the state of a phenomenon, subject to a tolerable limit on the risk of a false detection. Specifically, the substrate of the phenomenon is a time process that may unexpectedly undergo an abrupt change-of-state from "normal" to "abnormal", each defined as deemed appropriate given the physical context at hand. Inference about the current state of the process is drawn by virtue of (quantitative) observations (e.g., measurements). The *sequential setting* assumes the observations are made successively, and, so long as the behavior thereof suggests the process is in the normal state, it is let to continue. However, if the state is believed to have altered, one's aim is to detect the change "as soon as possible", so that an appropriate response can be provided in a timely manner. Thus, with the arrival of every new observation one is faced with the question of whether to let the process continue, or to stop it and raise an alarm (and, e.g., investigate). The decision has to be made in real time based on the available data. The time instance at which the process' state changes is referred to as the *change-point*, and the challenge is that it is not known in advance.

Historically, the subject of change-point detection first began to emerge in the 1920–1930's motivated by considerations of quality control. Shewhart's charts were popular in the past (see Shewhart 1931). Efficient (optimal and quasi-optimal) sequential detection procedures were developed much later in the 1950-1960's, after the emergence of Sequential Analysis, a branch of statistics ushered by Wald (1947). The ideas set in motion by Shewhart and Wald have formed a platform for a vast literature on both theory and practice of sequential change-point detection. See, e.g., Girschick and Rubin (1952), Page (1954), Shiryaev (1961, 1963, 1978), Roberts (1966), Siegmund (1985), Tartakovsky (1991), Brodsky and Darkhovsky (1993), Basseville and Nikiforov (1993), Poor and Hadjiliadis (2008).

The desire to detect the change quickly causes one to be trigger-happy, which, on one hand, will lead to an unacceptably high level of the risk of sounding a *false alarm* – terminating the process prematurely as a result of an erroneous decision that the change did occur, while, in fact, it never did. On the other hand, attempting to avoid false alarms too strenuously will cause a long delay between the actual time of occurrence of the change (i.e., the true change-point) and the time it is detected. Hence, the essence of the problem is to attain a tradeoff between two contradicting performance measures – the loss associated with the delay in 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).

Putting this idea on a rigorous mathematical basis requires formal definition of 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 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. Alone, each has its own history and area(s) of application. This notwithstanding the four approaches are connected and fit together into one big picture shown in Figure 1.

The aim of this paper is to give a brief *exposé* of the above four approaches to quickest change detection. Specifically, the plan is to assess the progress made to date within each with the emphasis on the novel exact and asymptotic optimality results.
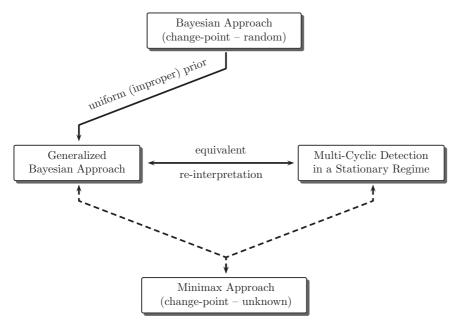
Fig. 1 Four approaches to sequential quickest change-point detection.

2 Change-point models

To formally state the general quickest change-point detection problem, we first have to introduce a change-point model as well as a model for the observations. To this end, a myriad of scenarios is possible; see, e.g., Fuh (2003, 2004), Lai (1995, 1998), Shiryaev (1961, 1963, 1978, 2009, 2010), Tartakovsky (1991, 2009a), Tartakovsky and Moustakides (2010), Tartakovsky and Veeravalli (2005). This section is intended to review the major ones.

A change-point model is characterized by the probabilistic structure of the monitored process (independent, identically or non-identically distributed, correlated, etc.) as well as by that of the change-point (unknown deterministic, random completely or partially dependent on the observed data, random fully independent from the observations).

Consider 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 permitting singular measures to be present, see Shiryaev (2009). For $d = \{0, \infty\}$, write $\mathbb{P}_d^{(n)} = \mathbb{P}_d|_{\mathcal{F}_n}$ for the restriction of measure $\mathbb{P}_d$ to the sigma-algebra $\mathcal{F}_n$, and let $p_d^{(n)}(\cdot)$ be the density of $\mathbb{P}_d^{(n)}$ (with respect to a dominating sigma-finite measure).

Let $\{X_n\}_{n \geqslant 1}$ denote the series of (random) observations; the series is defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and distribution-wise is such that for some time index ν, the observations $X_1, X_2, \ldots, X_\nu$ 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. Note that 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) . Another practice popular in the literature is to define ν as the serial number of the first post-change observation. Although the two definitions map into one another, throughout the remainder of the paper we will follow the former convention.

For every fixed $\nu \geqslant 0$, the change-of-regime in the series $\{X_n\}_{n\geqslant 1}$ gives rise to a new probability measure $\mathbb{P}_\nu$. We will now demonstrate how to 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 will write $p_\nu(\boldsymbol{X}_1^n)$ in the following.

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; recall Figure 1. To get density $p_\nu(\boldsymbol{X}_1^n)$, observe that by the Bayes rule

$$p_\infty(\boldsymbol{X}_1^n) = p_\infty(\boldsymbol{X}_1^\nu) \times p_\infty(\boldsymbol{X}_{\nu+1}^n | \boldsymbol{X}_1^\nu) \ \text{ and } \ p_0(\boldsymbol{X}_1^n) = p_0(\boldsymbol{X}_1^\nu) \times p_0(\boldsymbol{X}_{\nu+1}^n | \boldsymbol{X}_1^\nu),$$

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 assume either independence or homogeneity of observations — the observations may be arbitrary dependent and nonidentically distributed. Furthermore, in certain state-space models and hidden Markov models due to the propagation of the change-point the post-change conditional densities $p_0^{(j,\nu)}(X_j | \boldsymbol{X}_1^{j-1})$, $\nu + 1 \leqslant j \leqslant n$ depend on the change-point ν; see, e.g., Tartakovsky (2009a). Model (1) includes practically all possible scenarios. If, for example, there is a switch of one non-iid model to another non-iid model, which are mutually independent, then the two segments, pre- and post-change, of the observed process are independent, and in (1) the post-change conditional densities $p_0^{(j)}(X_j | \boldsymbol{X}_1^{j-1})$, $j \geqslant \nu + 1$ are replaced with $p_0^{(j)}(X_j | \boldsymbol{X}_j^\nu)$.

Suppose now that the observations $\{X_n\}_{n\geqslant 1}$ 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. For convenience, from now on it will be referred to as the *iid case*, or the *iid model*. It can be seen that 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 (see Figure 1), then the model has to be supplied with the change-point's *prior distribution*. There may be several change-point mechanisms and, as a result, a random variable ν may be dependent on the observations or independent from the observations. To account for these possibilities at once, 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.
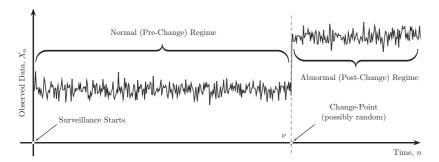
3 Overview of optimality criteria

Contemporary theory of change-point detection is an ensemble of the Bayesian approach, the generalized Bayesian approach, the minimax approach, and the approach related to multi-cyclic detection of a disorder in a stationary regime; see Figure 1. The object of this section is to briefly discuss each problem setting.
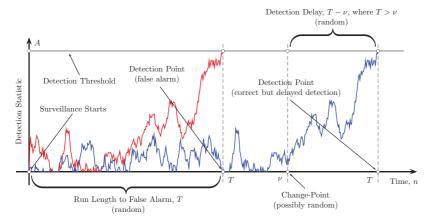
A *sequential detection procedure* is a stopping time T adapted to the filtration $\{\mathcal{F}_n\}_{n\geqslant 0}$ induced by the observations $\{X_n\}_{n\geqslant 1}$, i.e., the event $\{T \leqslant n\} \in \mathcal{F}_n$ for every $n \geqslant 0$. Therefore, after observing $X_1, \ldots, X_T$ it is declared that the change is in effect. That may or may not be the case. If it is not, then $T \leqslant \nu$, and it is said that a *false alarm* has been sounded. Also, note that since $\mathcal{F}_0$ is the trivial sigma-algebra, any $\{\mathcal{F}_n\}$-adapted stopping time T is either strictly positive with probability (w.p.) 1, or $T = 0$ w.p. 1. The latter case is clearly degenerate, and to preclude it, from now on we shall assume $T > 0$ w.p. 1.

Common to the Bayesian, generalized Bayesian, and minimax approaches is that the detection procedure is applied *only once*; the result is either a false alarm, or a correct (may be delayed) detection. Irrespectively, what takes place beyond the stopping point T is of no concern. We will refer to this as the single-run paradigm, which is shown in Figure 2. Figure 2(a) shows an example of the behavior of a certain process of interest as exhibited through the sequence of observations $\{X_n\}_{n\geqslant 1}$. It can be seen that the process undergoes a shift in the mean at some time instant ν, the change-point. Figure 2(b) (red trajectory) gives an example of the corresponding detection statistic trajectory that exceeds the detection threshold prematurely, i.e., before the change occurs. This is a false alarm situation, and T can be regarded as the (random) run length to the false alarm. Another possibility is shown in Figure 2(b) (blue trajectory). This is an example where the detection statistic exceeds the detection threshold past the change-point. Note that the detection delay, captured by the difference $T - \nu$, is random.

In a variety of surveillance applications the detection procedure should be applied *repeatedly*. This requires specification of a renewal mechanism after each alarm (false or true). The simplest renewal strategy is to restart from scratch, in which case the procedure becomes multi-cyclic with similar cycles (in a statistical sense) if the process is homogeneous. In the following sections, we will consider such an approach related to detection of a distant

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

(b) Two possible scenarios of the corresponding detection process: false alarm (red trajectory) and correct detection (blue trajectory).

Fig. 2 Single-run sequential change-point detection.

change in a stationary regime, assuming that the detection procedure is applied repeatedly starting anew after each time the detection statistic exceeds the threshold.

3.1 Bayesian formulation

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

Let $\{\pi_k\}_{k \geqslant 0}$ be the 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

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

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

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

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

We are now in a position to formally introduce 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

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

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

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, to avoid trivialities we have to assume that $\alpha < 1 - \pi$. 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 as well. 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

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

and if $A = A_\alpha$ can be selected in such a way that the PFA is exactly equal to α, i.e., $\mathrm{PFA}^\pi(T_{\mathrm{S}}(A_\alpha)) = \alpha$, then it is strictly optimal in the class $\Delta(\alpha)$, that is,

$$\inf_{T \in \Delta(\alpha)} \mathrm{ADD}^\pi(T) = \mathrm{ADD}^\pi(T_{\mathrm{S}}(A_\alpha)) \;\text{ for any }\; 0 < \alpha < 1 - \pi.$$

Note that Shiryaev's statistic $R_{n,p}$ can be rewritten in the recursive form

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

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 that covers practically arbitrary non-iid models and prior distributions, see Tartakovsky and Veeravalli (2005). Specifically, this work addresses the Bayesian approach assuming only that the prior distribution is independent of the observations. The overall conclusion made by the authors is two-fold: *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 $\mathrm{PFA}^\pi(T_{\mathrm{S}}(A_\alpha)) = \alpha$. 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

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

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)

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

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

$$\Delta(\gamma) = \big\{T \colon \mathbb{E}_\infty[T] \geqslant \gamma\big\}, \qquad (13)$$

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

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

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 $\{R_n^r\}_{n \geqslant 0}$, where $R_n^r = (1 + R_{n-1}^r)\Lambda_n$, $n \geqslant 1$ with $R_0^r = r \geqslant 0$. 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 address 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 $\mathbb{E}_\infty[T]$ (recall the false alarm scenario

from Figure 2(b)). As far as the risk associated with detection delay is concerned, Lorden suggested to use the "worst-worst-case" ADD defined as

$$\mathcal{J}_{\mathrm{L}}(T) = \sup_{0 \leqslant \nu < \infty} \left\{ \operatorname{ess\,sup} \mathbb{E}_{\nu}[(T - \nu)^+ | \mathcal{F}_{\nu}] \right\}. \tag{15}$$

Lorden's minimax optimization problem seeks to

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathcal{J}_{\mathrm{L}}(T_{\mathrm{opt}}) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{L}}(T) \text{ for every } \gamma > 1, \tag{16}$$

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 $\mathcal{J}_{\mathrm{L}}(T)$-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 $\mathcal{J}_{\mathrm{L}}(T)$ with

$$\mathcal{J}_{\mathrm{P}}(T) = \sup_{0 \leqslant \nu < \infty} \mathbb{E}_{\nu}[T - \nu | T > \nu], \tag{17}$$

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

$$\text{find } T_{\mathrm{opt}} \in \Delta(\gamma) \text{ such that } \mathcal{J}_{\mathrm{P}}(T_{\mathrm{opt}}) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) \text{ for every } \gamma > 1. \tag{18}$$

It is our opinion that $\mathcal{J}_{\mathrm{P}}(T)$ 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 $\mathcal{J}_{\mathrm{P}}(T) \leqslant \mathcal{J}_{\mathrm{L}}(T)$. Second, it is directly connected to the conventional decision theoretic approach — the optimization problem (18) 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 $\sup_{\pi} \mathrm{ADD}^{\pi}(T) = \mathcal{J}_{\mathrm{P}}(T)$, where $\mathrm{ADD}^{\pi}(T)$ is defined in (4). In addition, unlike Lorden's minimax problem (16), Pollak's minimax problem (18) is still not solved. For these reasons, from now on, when considering the minimax approach, we focus on Pollak's supremum ADD measure $\mathcal{J}_{\mathrm{P}}(T)$. 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 Sections 7 and 8.

We conclude this section with presenting yet another way to gauge the risk of raising a false alarm, namely, by means of 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 probability is a better option than the ARL to false alarm, which is more global. Specifically, let

$$\Delta_{\alpha}^{m} = \left\{ T \colon \sup_{k \geqslant 0} \mathbb{P}_{\infty}(k < T \leqslant k + m | T > k) \leqslant \alpha \right\}, \tag{19}$$

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 when $\alpha \to 0$.

Let T be the stopping time associated with a generic detection procedure. The appropriateness of the ARL to false alarm $\mathbb{E}_\infty[T]$ as an exhaustive measure of the risk of raising a false alarm is questionable, 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 $\mathbb{E}_\infty[T]$. 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).

For the iid model (2), 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 $p \geqslant 1$. See also Tartakovsky et al (2008). 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, which suggests that alternative measures of the false alarm rate are in order.

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). In Section 8 we take the work of Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010) one step further and present a procedure that solves the optimization problem (18) in the class (19) in a specific example.

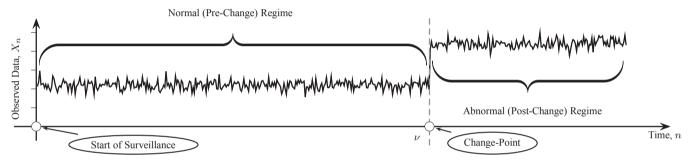3.4 Multi-cyclic detection of a disorder in a stationary regime

Common to all of the above approaches is that the detection procedure is applied only once. This is the single-run paradigm. The result is either a correct (though usually delayed) detection or a false alert; recall Figure 2. Yet another formulation may be derived by abandoning the single-run paradigm for the *multi-run* or the *multi-cyclic* one.

Specifically, 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 schematically shown in Figure 3. As one can see, the ARL to false alarm in this case is the mean time between (consecutive) false alarms, and therefore may be thought of the false alarm rate (or frequency).
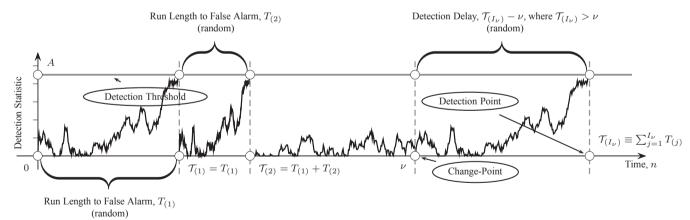
As argued by Pollak and Tartakovsky (2009b), the multi-cyclic approach is instrumental in many surveillance applications, in particular in the areas concerned with intrusion/anomaly detection, e.g., cybersecurity and particularly detection of attacks in computer networks.

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

$$\mathcal{J}_{\mathrm{ST}}(T) = \lim_{\nu \to \infty} \mathbb{E}_\nu[\mathcal{T}_{(I_\nu)} - \nu] \qquad (20)$$

(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.

Fig. 3 Multi-cyclic change-point detection in a stationary regime.

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

We are now in a position to formalize the notion of optimality in the multi-cyclic setup:

$$\text{find } T_{\text{opt}} \in \Delta(\gamma) \text{ such that } \mathcal{J}_{\text{ST}}(T_{\text{opt}}) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\text{ST}}(T) \text{ for every } \gamma > 1 \qquad (21)$$

(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 $\mathcal{J}_{\text{ST}}(T)$ is the same as $\text{RIADD}(T)$ 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., we 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 $\leqslant k < \infty$ 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

$$p(\boldsymbol{X}_1^n | \mathcal{H}_\infty) = \prod_{j=1}^{n} f(X_j),$$

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

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 then "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 stated in (14) and in the multi-cyclic disorder detection in a stationary regime stated in (21). We have already remarked that for the iid model in question the SR procedure solves both these problems. We preface the presentation of the exact results with the introduction of the SR procedure.

4.1 The Shiryaev–Roberts procedure

The SR procedure is due to the independent work of Shiryaev (1961, 1963) and 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 given by Pollak (1985), and it has become the convention.

Formally, the SR procedure is defined as the stopping time

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

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

$$R_n = (1 + R_{n-1})\,\Lambda_n, \quad n \geqslant 1 \ \text{ with } \ R_0 = 0 \tag{23}$$

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

4.2 Optimality properties

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 facts allow us to conjecture that the SR procedure is optimal in the generalized Bayesian sense. In addition, as we stated in Subsection 3.4, the RIADD is equal to the STADD of the multi-cyclic procedure, so that we expect that the repeated SR procedure is optimal for detecting distant changes. The exact result is due to Pollak and Tartakovsky (2009b) and is given next.

Theorem 1 (Pollak and Tartakovsky 2009b) *Let $\mathcal{S}_A$ be the SR procedure defined by (22) and (23). 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 $\mathbb{E}_\infty[T] \geqslant \gamma$, that is,*

$$\mathrm{RIADD}(\mathcal{S}_{A_\gamma}) = \inf_{T \in \Delta(\gamma)} \mathrm{RIADD}(T) \ \text{ for every } \ \gamma > 1.$$

(ii) *For any stopping time T, $\mathrm{RIADD}(T) = \mathcal{J}_{\mathrm{ST}}(T)$. Therefore, 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.,*

$$\mathcal{J}_{\mathrm{ST}}(\mathcal{S}_{A_\gamma}) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{ST}}(T) \ \text{ for every } \ \gamma > 1.$$

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 (32) 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 (18), 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 $\mathcal{J}_{\mathrm{P}}(T^*)$ and the (unknown) optimum $\inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T)$ 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

$$\frac{\mathcal{J}_{\mathrm{P}}(T^*)}{\inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T)} = 1 + o(1), \quad \text{as} \ \ \gamma \to \infty,$$

where $o(1) \to 0$ 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

$$\mathcal{J}_{\mathrm{P}}(T^*) - \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) = O(1), \quad \text{as} \ \ \gamma \to \infty,$$

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

$$\mathcal{J}_{\mathrm{P}}(T^*) - \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) = o(1), \quad \text{as} \ \ \gamma \to \infty.$$

5.1 The Shiryaev–Roberts–Pollak procedure

The question of what procedure minimizes Pollak's measure of detection delay $\mathcal{J}_{\mathrm{P}}(T)$ is an open issue. As an attempt to resolve the issue, Pollak (1985) proposed to "tweak" the SR procedure (22). 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 (18) 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; see Figure 4 below. 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 (23), 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

$$Q_A(x) = \lim_{n \to \infty} \mathbb{P}_\infty(R_n \leqslant x | \mathcal{S}_A > n). \tag{24}$$

Therefore, the SRP procedure is defined as the stopping time

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

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

$$R_n^Q = (1 + R_{n-1}^Q)\,\Lambda_n, \quad n \geqslant 1, \quad R_0^Q \sim Q_A(x) \tag{26}$$

is the detection statistic.

We reiterate that, by design, the SRP procedure (25) and (26) is an equalizer: it delivers the same conditional average detection delay for any change-point ν, 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 $\mathcal{J}_\mathrm{P}(T)$. More specifically, the following is true.

Theorem 2 (Pollak 1985) *Let $\mathbb{E}_0[(\log \Lambda_1)^+] < \infty$. Suppose that in the SRP procedure $\mathcal{S}_A^Q$ the detection threshold $A = A_\gamma$ is selected in such a way that $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^Q] = \gamma$. Then*

$$\mathcal{J}_\mathrm{P}(\mathcal{S}_{A_\gamma}^Q) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_\mathrm{P}(T) + o(1), \ \ as \ \ \gamma \to \infty.$$

Recently, Tartakovsky et al (2011) obtained the following asymptotic approximation for $\mathcal{J}_\mathrm{P}(\mathcal{S}_A^Q)$ under the second moment condition $\mathbb{E}_0[\log \Lambda_1]^2 < \infty$:

$$\mathbb{E}_0[\mathcal{S}_A^Q] = \frac{1}{I}(\log A + \varkappa - C_\infty) + o(1), \ \ as \ \ A \to \infty,$$

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 (2011).

Note that for sufficiently large γ,

$$\mathbb{E}_\infty[\mathcal{S}_A^Q] \approx (A/\zeta) - \mu_Q, \ \ where \ \ \mu_Q = \int_0^A y\, dQ_A(y), \tag{27}$$

i.e., μ_Q is the mean of the quasi-stationary distribution, and ζ is a constant defined in (32) 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 2011).

5.2 The Shiryaev–Roberts–r procedure

The third-order asymptotic optimality of the SRP procedure makes the latter practically appealing. On the flip size, the SRP rule requires the knowledge of the quasi-stationary distribution (24). It is rare that this distribution can be expressed in a closed form; for examples where this is possible, see 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 implementable, 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

$$\mathcal{S}_A^r = \inf\{n \geqslant 1 \colon R_n^r \geqslant A\}, \tag{28}$$

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

$$R_n^r = (1 + R_{n-1}^r)\,\Lambda_n, \quad n \geqslant 1, \quad \text{with} \quad R_0^r = r \geqslant 0 \tag{29}$$

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 $T_{A_1}^r$).

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 (28) 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. To understand why, let us look at Figure 4 which shows the average delay to detection $\mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ vs. ν for several initialization values $R_0^r = r$. This figure was obtained using the integral equations and the numerical technique of Moustakides et al (2011). For $r = 0$, this is the classical SR procedure (with $R_0 = 0$) whose average delay to detection is monotonically decreasing to its minimum (steady state value) that is attained at infinity. Note that in fact this steady state is attained for essentially finite values of the change-point ν. It is seen that there exists a value $r = r_A$ that depends on the threshold A for which the worst point ν is at infinity, i.e., $\mathcal{J}_\mathrm{P}(\mathcal{S}_A^{r_A}) = \mathrm{ADD}_\infty(\mathcal{S}_A^{r_A})$. The value of r_A is the minimal value for which this happens and it is also the value that delivers the minimum to the difference between $\mathcal{J}_\mathrm{P}(\mathcal{S}_A^r)$ and the lower bound for $\inf_{T \in \Delta(\gamma)} \mathcal{J}_\mathrm{P}(T)$ derived by Moustakides et al (2011) and by Polunchenko and Tartakovsky (2010). This is a very important observation, since it allows us to build a proof of asymptotic optimality based on an estimate of $\mathrm{ADD}_\infty(\mathcal{S}_A^r)$. We also note that for the SR–r procedure with initialization $r = r_A$ (pink line) the average detection delay at the beginning and at infinity are approximately equal, $\mathbb{E}_0[\mathcal{S}_A^{r_A}] \approx \mathrm{ADD}_\infty(\mathcal{S}_A^{r_A})$. This allows us to conjecture that an "optimal" SR–r is an equalizer at the beginning ($\nu = 0$) and at sufficiently large values of ν, so that initialization r_A should be selected to achieve this property. The following theorem, whose proof can be found in Polunchenko and Tartakovsky (2010), shows that the lower bound for the "minimax risk" can be expressed via the integral average detection delay of the SR–r procedure, which partially explains the issue.
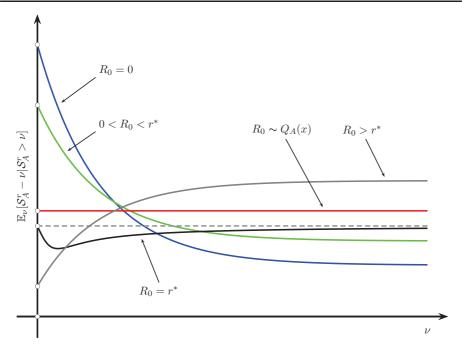
Fig. 4 Typical behavior of the conditional expected detection delay $\mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ of the SR–r procedure as a function of the change-point ν for various initialization strategies.

Theorem 3 *Let $\mathcal{S}_A^r$ be defined as in* (28) *and* (29), *and let $A = A_\gamma$ be selected so that $\mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^r] = \gamma$. Then, for* every *$r \geqslant 0$,*

$$\inf_{T \in \Delta(\gamma)} \mathcal{J}_P(T) \geqslant \frac{r\,\mathbb{E}_0[\mathcal{S}_{A_\gamma}^r] + \sum_{\nu=0}^\infty \mathbb{E}_\nu[(\mathcal{S}_{A_\gamma}^r - \nu)^+]}{r + \mathbb{E}_\infty[\mathcal{S}_{A_\gamma}^r]} = \mathcal{J}_B(\mathcal{S}_{A_\gamma}^r). \qquad (30)$$

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 (30) is equal to $\mathbb{E}_0[\mathcal{S}_A^r]$, which, in turn, is equal to $\sup_\nu \mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu] = \mathcal{J}_P(\mathcal{S}_A^r)$. Therefore, we have the following corollary that will be used in Section 7 for proving that the SR–r procedure with a specially designed $r = r_A$ is strictly optimal for two specific models.

Corollary 1 *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.,*

$$\inf_{T \in \Delta(\gamma)} \mathcal{J}_P(T) = \mathcal{J}_P(\mathcal{S}_{A_\gamma}^{r(\gamma)}). \qquad (31)$$

While the SR–r is not strictly minimax in general, it is almost obvious that this procedure is almost minimax. In fact, Moustakides et al (2011) conjecture that the SR–r procedure is third-order asymptotically minimax and Tartakovsky et al (2011) 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

$$\varkappa = \lim_{a \to \infty} \mathbb{E}_0[\kappa_a], \quad \zeta = \lim_{a \to \infty} \mathbb{E}_0\left[e^{-\kappa_a}\right]. \tag{32}$$

The constants $\varkappa > 0$ and $0 < \zeta < 1$ depend on the model and can be computed numerically. Next, 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

$$C_\infty = \mathbb{E}[\log(1 + R_\infty + \tilde{V}_\infty)] = \int_0^\infty \int_0^\infty \log(1 + x + y)\, dQ_{\mathrm{ST}}(x)\, d\tilde{Q}(y), \tag{33}$$

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

Theorem 4 ([Tartakovsky et al 2011]) *Let $\mathbb{E}_0[\log \Lambda_1]^2 < \infty$ and let $\log \Lambda_1$ be non-arithmetic. Then the following assertions hold.*

(i) *As $\gamma \to \infty$,*

$$\inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) \geqslant \frac{1}{I}[\log(\gamma\zeta) + \varkappa - C_\infty] + o(1). \tag{34}$$

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

$$\mathrm{ADD}_\infty(\mathcal{S}_A^r) = \mathbb{E}_0[\mathcal{S}_A^Q] = \frac{1}{I}(\log A + \varkappa - C_\infty) + o(1), \quad as \ \ A \to \infty. \tag{35}$$

(iii) *Furthermore, if in the SR–r procedure $A = A_\gamma = \gamma\zeta$ and the initialization point $r = o(\gamma)$ is selected so that $\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^r) = \mathrm{ADD}_\infty(\mathcal{S}_A^r)$, then $\mathbb{E}_\infty[\mathcal{S}_A^r] = \gamma(1 + o(1))$ and*

$$\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^r) = \frac{1}{I}[\log(\gamma\zeta) + \varkappa - C_\infty] + o(1) \ \ as \ \ \gamma \to \infty. \tag{36}$$

Therefore, the SR–r procedure is third-order asymptotically optimal:

$$\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^r) - \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) = o(1), \quad as \ \ \gamma \to \infty.$$

Also,

$$\mathrm{ADD}_0(\mathcal{S}_A^r) = \frac{1}{I}[\log A + \varkappa - C(r)] + o(1), \quad as \ \ A \to \infty, \tag{37}$$

where

$$C(r) = \mathbb{E}[\log(1 + r + \tilde{V}_\infty)] = \int_0^\infty \log(1 + r + y)\, d\tilde{Q}(y). \tag{38}$$

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 (35) and (37) 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 for the conventional SR procedure that starts from zero

$$\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A) = \mathrm{ADD}_0(\mathcal{S}_A) = \frac{1}{I}[\log A + \varkappa - C(0)] + o(1), \text{ as } A \to \infty.$$

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 (27), for sufficiently large γ,

$$\mathbb{E}_\infty[\mathcal{S}_A^r] \approx (A/\zeta) - r. \tag{39}$$

Polunchenko and Tartakovsky (2010) and Tartakovsky and Polunchenko (2010) offer two scenarios where the SR–r procedure is *strictly* minimax. Both are discussed (and extended) in Section 7. In addition, in Section 8 we present 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.

6 Numerical performance evaluation

Recall that each of the four approaches adduced above is characterized by its own optimality criterion. Together they bring about a variety of performance measures. Hence, to judge about the efficiency of a detection procedure (with respect to one performance measure or another), one has to be able to compute the procedure's corresponding operating characteristics (OC). In this section, we present integral equations for a multitude of OC-s that are of interest in various problem settings (Bayesian, minimax, etc.) and practical applications. Usually these equations cannot be solved analytically and numerical techniques are needed; cf. Moustakides et al (2011); Tartakovsky et al (2009).

Consider a generic detection procedure described by the stopping time

$$\mathcal{T}_A^s = \inf\{n \geqslant 1 : V_n^s \geqslant A\}, \tag{40}$$

where $A > 0$ is the detection threshold and $\{V_n^s\}_{n \geqslant 0}$ is a Markov detection statistic that satisfies the recursive relation

$$V_n^s = \xi(V_{n-1}^s)\,\Lambda_n, \quad n \geqslant 1 \text{ with } V_0^s = s \geqslant 0, \tag{41}$$

where $\xi(x)$ is a positive-valued function and s is a fixed parameter referred to as the starting point or the "head start". Observe first that this detection scheme constitutes a rather broad class of detection schemes that includes, e.g., CUSUM, Shiryaev's procedure, SR–r, and EWMA (exponentially weighted moving average). Indeed, for the Shiryaev procedure $\xi(V) = (1 + V)/(1 - p)$ and for the SR–r procedure $\xi(V) = 1 + V$. (We will not consider other procedures such as CUSUM and EWMA in this paper).

Let $P_d^\Lambda(t) = \mathbb{P}_d(\Lambda_1 \leqslant t)$ denote the cdf of the LR Λ_1 under the measure $\mathbb{P}_d$, $d = \{0, \infty\}$. Define

$$\mathcal{K}_d(x, y) = \frac{\partial}{\partial y}\mathbb{P}_d(V_{n+1}^s \leqslant y | V_n^s = x) = \frac{\partial}{\partial y}P_d^\Lambda\left(\frac{y}{\xi(x)}\right), \quad d = \{0, \infty\},$$

where $\{V_n^s\}_{n\geqslant 0}$ is as in (41). Here and in the following we assume that Λ_1 is continuous under both hypotheses.

Let $\ell(x) = \mathbb{E}_\infty[\mathcal{T}_A^x]$ and $\delta_0(x) = \mathbb{E}_0[\mathcal{T}_A^x]$, where $\mathcal{T}_A^x$ is as in (40). Observe that $\ell(x)$ and $\delta_0(x)$ are conditional expectations of the form $\mathbb{E}_d[\,\cdot\,|V_0^x = x]$ (for $d = \infty$ and $d = 0$ respectively), where $V_0^x = x$ is the starting point of the generic detection statistic (41). It is not difficult to see that $\ell(x)$ and $\delta_0(x)$ are governed by the equations

$$\ell(x) = 1 + \int_0^A \mathcal{K}_\infty(x,y)\,\ell(y)\,dy, \tag{42}$$

and

$$\delta_0(x) = 1 + \int_0^A \mathcal{K}_0(x,y)\,\delta_0(y)\,dy, \tag{43}$$

respectively; cf. Moustakides et al (2011).

Next, for any $\nu \geqslant 0$, let $\delta_\nu(x) = \mathbb{E}_\nu[(\mathcal{T}_A^x - \nu)^+]$ and $\rho_\nu(x) = \mathbb{P}_\infty(\mathcal{T}_A^x > \nu)$. By Moustakides et al (2011),

$$\delta_{\nu+1}(x) = \int_0^A \mathcal{K}_\infty(x,y)\,\delta_\nu(y)\,dy, \quad \rho_{\nu+1}(x) = \int_0^A \mathcal{K}_\infty(x,y)\,\rho_\nu(y)\,dy, \tag{44}$$

where $\delta_0(x)$ is governed by (43) and $\rho_0(x) = 1$ for all x, since $\mathbb{P}_\infty(\mathcal{T}_A^x > 0) = 1$. Consider now the conditional average delays to detection $\mathbb{E}_\nu[\mathcal{T}_A^x - \nu|\mathcal{T}_A^x > \nu] = \mathbb{E}_\nu[(\mathcal{T}_A^x - \nu)^+]/\mathbb{P}_\nu(\mathcal{T}_A^x > \nu), \nu \geqslant 0$. Since $\mathbb{P}_\nu(\mathcal{T}_A^x > \nu) = \mathbb{P}_\infty(\mathcal{T}_A^x > \nu)$, we obtain

$$\mathbb{E}_\nu[\mathcal{T}_A^x - \nu|\mathcal{T}_A^x > \nu] = \frac{\mathbb{E}_\nu[(\mathcal{T}_A^x - \nu)^+]}{\mathbb{P}_\infty(\mathcal{T}_A^x > \nu)} = \frac{\delta_\nu(x)}{\rho_\nu(x)}, \quad \nu \geqslant 0,$$

where $\delta_\nu(x)$ and $\rho_\nu(x)$ are given by (44). Thus, the conditional average detection delays can be computed for any $\nu \geqslant 0$, which allows one to evaluate $\sup_{\nu\geqslant 0}\mathbb{E}_\nu[\mathcal{T}_A^x - \nu|\mathcal{T}_A^x > \nu]$.

Now, let $\psi(x) = \sum_{\nu=0}^\infty \mathbb{E}_\nu[(\mathcal{T}_A^x - \nu)^+] = \sum_{\nu=0}^\infty \delta_\nu(x)$. By Theorem 1,

$$\mathcal{J}_{\text{ST}}(\mathcal{T}_A^x) = \text{RIADD}(\mathcal{T}_A^x) = \frac{\sum_{\nu=0}^\infty \mathbb{E}_\nu[(\mathcal{T}_A^x - \nu)^+]}{\mathbb{E}_\infty[\mathcal{T}_A^x]} = \frac{\psi(x)}{\ell(x)},$$

so that in order to compute the STADD $\mathcal{J}_{\text{ST}}(\mathcal{T}_A^x)$ we have to be able to compute $\psi(x)$. As shown by Moustakides et al (2011), $\psi(x)$ is determined by the equation

$$\psi(x) = \delta_0(x) + \int_0^A \mathcal{K}_\infty(x,y)\,\psi(y)\,dy, \tag{45}$$

where $\delta_0(x)$ is governed by equation (43).

Note that the lower bound (30) for the minimax risk given in Theorem 3 can be computed as

$$\mathcal{J}_{\text{B}}(\mathcal{T}_A^x) = \frac{x\delta_0(x) + \psi(x)}{x + \ell(x)},$$

where $\ell(x)$, $\delta_0(x)$, and $\psi(x)$ are governed by equations (42), (43), and (45).

The local conditional probabilities of false alarm $\mathbb{P}_\infty(\mathcal{T}_A^x \leqslant k + m|\mathcal{T}_A^x > k)$, $k \geqslant 0$ inside a fixed "window" of size $m = 1, 2, \ldots$ can also be evaluated noting that $\mathbb{P}_\infty(\mathcal{T}_A^x \leqslant k + m|\mathcal{T}_A^x > k) = 1 - \rho_{k+m}(x)/\rho_k(x)$, where $\rho_k(x)$ are as in (44). Having $\mathbb{P}_\infty(\mathcal{T}_A^x \leqslant$

$k + m | \mathcal{T}_A^x > k)$ evaluated for sufficiently many k's, one can easily find $\sup_k \mathbb{P}_\infty \mathcal{T}_A^x \leqslant k + m | \mathcal{T}_A^x > k)$ for any fixed m.

The next step is to extend the obtained equations to the case when $\mathcal{T}_A^x$ is randomized similarly to the SRP procedure (25) and (26). To this end, let $Q_A(y) = \lim_{n \to \infty} \mathbb{P}_\infty(V_n^s \leqslant y | \mathcal{T}_A^s > n)$ be the quasi-stationary distribution. Note that this distribution does not depend on the starting point $V_0^s = s$ and exists whenever the LR is continuous; cf. (Harris 1963, Theorem III.10.1).

It can be shown that the quasi-stationary pdf $q_A(x) = dQ_A(x)/dx$ satisfies the equation

$$\lambda_A \, q_A(y) = \int_0^A q_A(x) \, \mathcal{K}_\infty(x, y) \, dx, \quad \text{subject to} \quad \int_0^A q_A(x) \, dx = 1, \qquad (46)$$

whence one can conclude that $q_A(x)$ is the *left dominant eigenvector* of the linear integral operator induced by the kernel $\mathcal{K}_\infty(x, y)$, and $\lambda_A \in (0, 1)$ is the corresponding eigenvalue; cf. Moustakides et al (2011) and Pollak (1985). We also note that both $q_A(x)$ and λ_A are *unique*.

Consider now $\mathcal{T}_A^Q$ defined as the above generic procedure $\mathcal{T}_A^x$ with the starting point being random and sampled from the quasi-stationary distribution. Specifically,

$$\mathcal{T}_A^Q = \inf\{n \geqslant 1 \colon V_n^Q \geqslant A\}, \qquad (47)$$

where $A > 0$ is the detection threshold, and $\{V_n^Q\}_{n \geqslant 0}$ is a generic detection statistic computed recursively

$$V_n^Q = \xi(V_{n-1}^Q) \, \Lambda_n, \quad n \geqslant 1 \quad \text{with} \quad V_0^Q \sim Q_A. \qquad (48)$$

We note that the SRP procedure is the special case of $\mathcal{T}_A^Q$ with $\xi(x) = 1 + x$.

Once $q_A(x)$ and λ_A are available, one can compute the ARL to false alarm and the detection delay (which is independent from the change-point) for this randomized variant $\mathcal{T}_A^Q$ of the generic procedure $\mathcal{T}_A^x$. Indeed,

$$\mathbb{E}_\infty[\mathcal{T}_A^Q] = \int_0^A \ell(x) \, q_A(x) \, dx = \frac{1}{1 - \lambda_A} \quad \text{and} \quad \mathbb{E}_0[\mathcal{T}_A^Q] = \int_0^A \delta_0(x) \, q_A(x) \, dx.$$

To understand the second equality in the formula for $\mathbb{E}_\infty[\mathcal{T}_A^Q]$, note that $\mathcal{T}_A^Q$ is $\mathbb{P}_\infty$-geometrically distributed with the "probability of success" $1 - \lambda_A$. We also remark that, by design, the randomized variant $\mathcal{T}_A^Q$ of the generic procedure $\mathcal{T}_A^x$ is an equalizer, i.e., $\mathbb{E}_0[\mathcal{T}_A^Q] = \mathbb{E}_\nu[\mathcal{T}_A^Q - \nu | \mathcal{T}_A^Q > \nu]$ for all $\nu \geqslant 1$.

Finally, we present Bayesian operating characteristics — the average detection delay

$$\mathrm{ADD}^\pi(\mathcal{T}_A^x) = \frac{\sum_{k=0}^\infty \pi_k \, \mathbb{E}_k(\mathcal{T}_A^x - k)^+}{1 - \mathrm{PFA}^\pi(\mathcal{T}_A^x)}$$

and the probability of false alarm, $\mathrm{PFA}^\pi(\mathcal{T}_A^x) = \sum_{k=1}^\infty \pi_k \mathbb{P}_\infty(\mathcal{T}_A^x \leqslant k)$. Assuming the geometric prior distribution (6), we obtain

$$\mathrm{PFA}^\pi(\mathcal{T}_A^x) = (1 - \pi) \left\{ 1 - p \sum_{k=0}^\infty (1 - p)^k \rho_k(x) \right\},$$

$$\sum_{k=0}^\infty \pi_k \, \mathbb{E}_k[(\mathcal{T}_A^x - k)^+] = \pi \delta_0(r) + (1 - \pi) p \sum_{k=0}^\infty (1 - p)^k \delta_k(r)$$

(cf. Tartakovsky and Moustakides 2010). Let $\psi_p(x) = \sum_{k=0}^{\infty}(1-p)^k\delta_k(x)$ and $\chi_p(x) = \sum_{k=0}^{\infty}(1-p)^k\rho_k(x)$. Using the Markov property of the statistic V_n^x, it is readily seen that $\psi_p(x)$ and $\chi_p(x)$ satisfy the following integral equations

$$\psi_p(x) = \delta_0(x) + (1-p)\int_0^A \mathcal{K}_\infty(x,y)\,\psi_p(y)\,dy,$$

$$\chi_p(x) = 1 + (1-p)\int_0^A \mathcal{K}_\infty(x,y)\,\chi_p(y)\,dy.$$

The PFA and ADD are then computed, respectively, as

$$\mathrm{PFA}^\pi(\mathcal{T}_A^x) = (1-\pi)\left\{1 - p\chi_p(x)\right\} \quad \text{and} \quad \mathrm{ADD}^\pi(\mathcal{T}_A^x) = \frac{\pi\delta_0(x) + (1-\pi)p\psi_p(x)}{\pi + (1-\pi)p\chi_p(x)}.$$

The above equations are Fredholm integral equations of the second kind. As a rule, such equations do not allow for an (exact) analytical solution. For a few exceptions from the rule see Pollak (1985), Mevorach and Pollak (1991), Polunchenko and Tartakovsky (2010), and Tartakovsky and Polunchenko (2010). The results of the last two papers are summarized and extended in Section 7. Hence, a numerical technique may be in order. A simple numerical interpolation-projection type scheme has been suggested by Moustakides et al (2011). The scheme is effectively a piecewise collocation method with interpolating polynomials being of degree zero (constants). Using, e.g., (Atkinson and Han 2009, Theorem 12.1.2) we can conclude that the corresponding rate of convergence is at worst linear.

The above performance evaluation methodology can now be applied to any particular scenario we may be interested in. A few such scenarios are worked out in Sections 7 and 8.

7 Exact optimality of the Shiryaev–Roberts–r procedure

As we have pointed out earlier, the question of what solves Pollak's version of the minimax optimization problem (18) has been open since its inception in 1985. Because of the third-order asymptotic optimality and the fact that it is an equalizer it was conjectured that the SRP procedure might be the sought optimum. In this subsection, we suggest two counterexamples that disprove this conjecture. These examples show that a) the SRP procedure is not optimal, and b) that the SR–r procedure is optimal. We stress that the SR–r procedure is optimal in these examples, but not in general.

As a starting point, observe that equations (42), (43), (45) and (46) are special cases of the more general equation

$$u(x) = v(x) + \int_0^A \mathcal{K}(x,y)\,u(y)\,dy, \tag{49}$$

where $v(x)$ is a given function, $u(x)$ is the sought (unknown) function, and $\mathcal{K}(x,y)$, which is called the *kernel* of this equation, is of the form

$$\mathcal{K}(x,y) = \frac{\partial}{\partial y}P^\Lambda\left(\frac{y}{1+x}\right),$$

with $P^\Lambda(x)$ being the cdf of the LR $\Lambda_n = g(X_n)/f(X_n)$.

To see that (49) is an "umbrella" equation for all equations we are interested in, note that to obtain equation (42), which determines the ARL to false alarm, it suffices to take $v(x) =$

1 for all x and $\mathcal{K}(x, y) = \mathcal{K}_\infty(x, y)$. Likewise, equation (43), which governs the ADD at $\nu = 0$, can be obtained from (49) by assuming $v(x) = 1$ for all x and $\mathcal{K}(x, y) = \mathcal{K}_0(x, y)$. By a similar argument, one can also verify that equations (45) and (46) are instances of (49). Thus, if one is able to solve equation (49), one is also able to solve any of the equations of interest.

Suppose now that we have a change-point scenario for which the cdf $P^\Lambda(t)$ is such that

$$P^\Lambda \left(\frac{y}{1+x} \right) = \mathcal{X}(x) \, \mathcal{Y}(y)$$

for some sufficiently smooth functions $\mathcal{X}(x)$ and $\mathcal{Y}(y)$. In this case, the kernel $\mathcal{K}(x, y)$ is separable, i.e.,

$$\mathcal{K}(x, y) = \frac{\partial}{\partial y} P^\Lambda \left(\frac{y}{1+x} \right) = \mathcal{X}(x) \frac{d}{dy} \mathcal{Y}(y) = \mathcal{X}(x) \, \mathcal{Y}'(y),$$

so that the variables x and y are separated.

If the kernel is separable and the interval of integration has constant limits, the above equation can be solved analytically, and the solution is $u(x) = v(x) + M \mathcal{X}(x)$, where

$$M = \left(\int_0^A v(t) \, \mathcal{Y}'(t) \, dt \right) \bigg/ \left(1 - \int_0^A \mathcal{X}(t) \, \mathcal{Y}'(t) \, dt \right),$$

which is a function of A only.

More important is the fact that in this case

$$\mathbb{P}(R_1^r \leqslant y | \mathcal{S}_A^r > 1) = \mathbb{P}(R_1^r \leqslant y | R_1^r < A, R_0^r = r) = \mathcal{Y}(y) / \mathcal{Y}(A),$$

i.e., $\mathbb{P}(R_1^r \leqslant y | \mathcal{S}_A^r > 1)$ does not depend on the starting point $R_0^r = r$. This means that the quasi-stationary distribution "kicks in" as early as the first observation becomes available. As a result, the SR–r procedure is an equalizer for $\nu \geqslant 1$, and the only "degree of freedom" is $\nu = 0$. If one now designs the starting point R_0^r so as to equate the performance of the SR–r procedure at $\nu = 0$ to that at $\nu \geqslant 1$, then the SR–r procedure will be an equalizer for all $\nu \geqslant 0$. Therefore, by Corollary 1, in this case it is minimax. Note that this equalizer is different from the SRP rule, which is also an equalizer. The SR–r is an equalizer and minimax not in general but only in this particular case, i.e., in the case when the kernel is separable. Thus, we now have to find examples where this is true. This will prove that the SRP procedure is not strictly minimax in general.

Suppose now that the observations' distribution is $\mathsf{uniform}(0, 1)$ pre-change and $\mathsf{beta}(2, 1)$ post-change, that is, $f(x) = \mathbb{1}_{\{0 < x < 1\}}$ and $g(x) = 2x \, \mathbb{1}_{\{0 < x < 1\}}$. The LR is $\Lambda_n = 2X_n \, \mathbb{1}_{\{0 < X_n < 1\}}$; observe that $\Lambda_n \in (0, 2)$, since $X_n \in (0, 1)$. Hence,

$$P_\infty^\Lambda(t) = \begin{cases} 1, & \text{if } t \geqslant 2; \\ t/2, & \text{if } 0 \leqslant t < 2; \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad P_0^\Lambda(t) = \begin{cases} 1, & \text{if } t \geqslant 2; \\ (t/2)^2, & \text{if } 0 \leqslant t < 2; \\ 0, & \text{otherwise}. \end{cases}$$

It is apparent that both these distributions are monomial and therefore separable. As a result, one can compute the required operating characteristics of any SR-type procedure analytically. This was done by Tartakovsky and Polunchenko (2010). Another example, where $f(x) = e^{-x} \mathbb{1}_{\{x \geqslant 0\}}$ and $g(x) = 2e^{-2x} \mathbb{1}_{\{x \geqslant 0\}}$, was considered by Polunchenko and Tartakovsky (2010). Although this model may seem very different from the $\mathsf{uniform}(0, 1)$-to-$\mathsf{beta}(2, 1)$ model, it has exactly the same distributions $P_d^\Lambda(t)$, $d = \{0, \infty\}$. Both papers

established the following theorem the proof of which can be found in Polunchenko and Tartakovsky (2010).

Theorem 5 (Polunchenko and Tartakovsky 2010) *Let* $\bar{\gamma} = 1/(1 - 0.5 \log 3) \approx 2.2$.

(i) *If the starting point* r *in the SR–r procedure is chosen as* $r_A = \sqrt{1 + A} - 1$ *and the detection threshold* $A = A_\gamma$ *is set to the solution of the transcendental equation*

$$A + (\gamma - 1)\sqrt{1 + A} \log(1 + A) - 2(\gamma - 1)\sqrt{1 + A} = 0,$$

then, for every $\gamma \in (1, \bar{\gamma})$, *the ARL to false alarm* $\mathbb{E}_\infty[\mathcal{S}_A^r]$ *is exactly* γ *and the SR–r procedure is strictly minimax. That is,*

$$\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^r) = \inf_{T \in \Delta(\gamma)} \mathcal{J}_{\mathrm{P}}(T) \ \text{for every} \ \gamma \in (1, \bar{\gamma}).$$

(ii) *If the detection threshold in the SRP procedure* $\mathcal{S}_B^Q$ *is set to* $B = B_\gamma = \exp\{2(1 - 1/\gamma)\} - 1$, *then the ARL to false alarm* $\mathbb{E}_\infty[\mathcal{S}_B^Q]$ *is exactly* γ *and* $\mathcal{J}_{\mathrm{P}}(\mathcal{S}_B^Q)$ *is strictly greater than* $\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^r)$ *for every* $\gamma \in (1, \bar{\gamma})$. *Hence, the SRP procedure is suboptimal.*
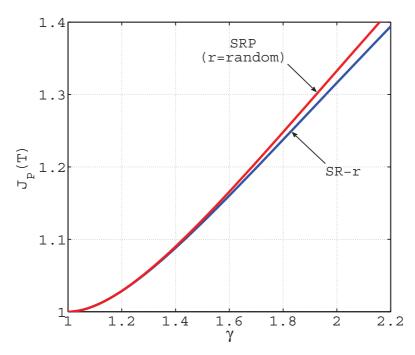

Fig. 5 Performance of the SRP procedure vs. that of the SR–r rule for the uniform$(0, 1)$-to-beta$(2, 1)$ model. The detection threshold for either procedure is between 0 and 2.

This theorem is illustrated in Figure 5. Note that the curves in the picture are *exact*. We stress again that while the SR–r procedure is exactly minimax in this example, it is still an open question what minimizes Pollak's $\mathcal{J}_{\mathrm{P}}(T)$ in general. We conjecture that, in the general case, the optimal procedure is based on the deterministic initialization $\{r_n\}_{n \geqslant 1}$ that

depends on time, in which case a threshold $A = A_n$ may also be a function of time. We also note that even if this fact is proved rigorously, finding the sequences $\{r_n(\gamma)\}$ and $\{A_n(\gamma)\}$ (in every particular case) is an extremely difficult problem. Solving this problem may not be worth trying since the difference between the lower bound (30) and the supremum ADD is usually small, at least for a moderate (and of course low) false alarm rate. See Moustakides et al (2011) and Figure 9 below.

We conclude this subsection with a remark concerning exact optimality of the SR–r procedure in the class $\Delta_\alpha^m = \{T\colon \sup_k \mathbb{P}_\infty(k < T \leqslant k + m | T > k) \leqslant \alpha\}$, where $\alpha \in (0,1)$ and $m \geqslant 1$. We first discussed this class in Subsection 3.3, where we mentioned that it is "stronger" than the class $\Delta(\gamma)$, i.e., in general $\Delta(\gamma)$ contains Δ_α^m. It can be easily verified that the $\mathbb{P}_\infty$-distribution of the SR–r stopping time $\mathcal{S}_A^r$ is zero-modified geometric:

$$\mathbb{P}_\infty(k < \mathcal{S}_A^r \leqslant k + m | \mathcal{S}_A^r > k) = 1 - \begin{cases} \left[\dfrac{1}{2}\log(1+A)\right]^m & \text{for } k \geqslant 1; \\[3ex] \dfrac{A}{2(1+r)}\left[\dfrac{1}{2}\log(1+A)\right]^{m-1} & \text{for } k = 0, \end{cases}$$

where $m \geqslant 1$. Thus, there is a one-to-one correspondence between the classes Δ_α^m and $\Delta(\gamma)$. As a result, the SR–r procedure is minimax in the class Δ_α^m as well. The same is true for the exponential model considered by Polunchenko and Tartakovsky (2010). We believe that this is the first exact optimality result in the class Δ_α^m.

8 Case studies

This section dissects two specific cases of the iid model (2) to illustrate the performance margin between the SR$-r$ and SRP procedures $\mathcal{S}_A^r$ and $\mathcal{S}_A^Q$, defined in (28), (29) and (25), (26), respectively.

8.1 Example 1: A beta-to-beta model

Suppose the pre- and post-change densities are, respectively,

$$f(x) = \frac{x^{\delta-1}(1-x)^\delta}{\mathtt{B}(\delta, \delta+1)}\,\mathbb{1}_{\{0<x<1\}} \quad\text{and}\quad g(x) = \frac{x^\delta(1-x)^{\delta-1}}{\mathtt{B}(\delta+1, \delta)}\,\mathbb{1}_{\{0<x<1\}},$$

where $\delta > 0$ is a given constant and $\mathtt{B}(\cdot, \cdot)$ is the Beta function. That is, the observations X_n, $n \geqslant 1$ are iid $\mathtt{beta}(\delta, \delta+1)$-distributed pre-change and iid $\mathtt{beta}(\delta+1, \delta)$-distributed post-change. This $\mathtt{beta}(\delta, \delta+1)$-to-$\mathtt{beta}(\delta+1, \delta)$ model is of interest in the context of studying the accuracy of the asymptotic expansions for the performance of the two competing SR-type procedures. Specifically, recall that for sufficiently large detection thresholds,

$$\mathbb{E}_\infty[\mathcal{S}_A^Q] \approx A/\zeta - \mu_Q \quad\text{and}\quad \mathrm{ADD}_\nu(\mathcal{S}_A^Q) \approx \frac{1}{I}(\log A + \varkappa - C_\infty) \quad\text{for all } \nu \geqslant 0,$$

$$\mathbb{E}_\infty[\mathcal{S}_A^r] \approx A/\zeta - r \quad\text{and}\quad \mathrm{ADD}_\infty(\mathcal{S}_A^r) \approx \frac{1}{I}(\log A + \varkappa - C_\infty),$$

where $I = \mathbb{E}_0[\log \Lambda_1]$ is the Kullback–Leibler information number, ζ and $\varkappa$ are defined in (32), μ_Q is the mean of the quasi-stationary distribution, and the constant C_∞ is defined in (33).

For the $\mathtt{beta}(\delta, \delta + 1)$-to-$\mathtt{beta}(\delta + 1, \delta)$ model, C_∞, $\varkappa$, ζ and I are all computable *analytically* for any $\delta > 0$. This is of much aid in the context of testing the accuracy of the asymptotic approximations. Specifically, we first present the exact, explicit formulas for each of the needed quantities, assuming arbitrary $\delta > 0$. We then evaluate the performance of the procedures of interest using the methodology of Section 6 and compare the obtained performance against that predicted by the asymptotic approximations.

Observe that $\Lambda_n = X_n/(1 - X_n)$ for any $\delta > 0$, whence one can readily deduce $P_d^\Lambda(t) = \mathbb{P}_d(\Lambda_1 \leqslant t)$, $d = \{0, \infty\}$. Specifically, the densities $p_d^\Lambda(t) = dP_d^\Lambda(t)/dt$, $d = \{0, \infty\}$, can be seen to be

$$p_\infty^\Lambda(t) = \frac{t^{\delta-1}(1 + t)^{-2\delta-1}}{\mathtt{B}(\delta, \delta + 1)} \, \mathbb{1}_{\{t > 0\}} \quad \text{and} \quad p_0^\Lambda(t) = \frac{t^\delta (1 + t)^{-2\delta-1}}{\mathtt{B}(\delta + 1, \delta)} \, \mathbb{1}_{\{t > 0\}}, \quad (50)$$

i.e., under either measure $\mathbb{P}_d$, $d = \{0, \infty\}$, the LR's distribution is Beta of type II (also known as the Beta prime distribution); the parameters are δ and $\delta + 1$ under measure $\mathbb{P}_\infty$, and $\delta + 1$ and δ under measure $\mathbb{P}_0$. The fact that $p_\infty^\Lambda(t)$ and $p_0^\Lambda(t)$ are both Beta prime with "mirrored" parameters suggests a certain symmetry embedded in the $\mathtt{beta}(\delta, \delta + 1)$-to-$\mathtt{beta}(\delta + 1, \delta)$ model. Specifically, consider the "dual" $\mathtt{beta}(\delta + 1, \delta)$-to-$\mathtt{beta}(\delta, \delta + 1)$ model. That is, suppose the pre- and post-change distributions – $f(x)$ and $g(x)$ – are swapped so that the former is not $\mathtt{beta}(\delta, \delta + 1)$, but $\mathtt{beta}(\delta + 1, \delta)$, and the latter is not $\mathtt{beta}(\delta + 1, \delta)$, but $\mathtt{beta}(\delta, \delta + 1)$. A special case of this swapped model (with $\delta = 1$) was considered by Tartakovsky et al (2011). It can be shown, exploiting properties of the Beta and Beta prime distributions, that for the swapped $\mathtt{beta}(\delta + 1, \delta)$-to-$\mathtt{beta}(\delta, \delta + 1)$ model, the densities $p_d^\Lambda(t)$, $d = \{0, \infty\}$, are exactly the same as those we just derived for the original $\mathtt{beta}(\delta, \delta+1)$-to-$\mathtt{beta}(\delta+1, \delta)$ model; see (50). Put otherwise, the $\mathtt{beta}(\delta, \delta+1)$-to-$\mathtt{beta}(\delta + 1, \delta)$ model and the $\mathtt{beta}(\delta + 1, \delta)$-to-$\mathtt{beta}(\delta, \delta + 1)$ model are statistically indistinguishable, for any $\delta > 0$. This symmetry entails a few consequences to be demonstrated next.

Consider the stationary distribution $Q_{\mathrm{ST}}(x) = \lim_{n \to \infty} \mathbb{P}_\infty(R_n \leqslant x)$ of the SR statistic $\{R_n\}_{n \geqslant 0}$. The quasi-stationary pdf $q_{\mathrm{ST}}(x) = dQ_{\mathrm{ST}}(x)/dx$ is governed by the equation

$$q_{\mathrm{ST}}(x) = \int_0^\infty \frac{\partial}{\partial x} P_\infty^\Lambda \left(\frac{x}{1 + y} \right) q_{\mathrm{ST}}(y) \, dy,$$

which can be derived from equation (46) for the quasi-stationary pdf, $q_A(x)$, by letting $A \to \infty$ and noticing that $\lim_{A \to \infty} \lambda_A = 1$ and $\lim_{A \to \infty} q_A(x) = q_{\mathrm{ST}}(x)$ (cf. Pollak and Siegmund 1986). Using (50), we obtain

$$q_{\mathrm{ST}}(x) = \frac{x^{\delta-1}}{\mathtt{B}(\delta + 1, \delta)} \int_0^\infty \frac{(1 + y)^{\delta+1}}{(1 + x + y)^{1+2\delta}} q_{\mathrm{ST}}(y) \, dy,$$

and the (exact) solution is

$$q_{\mathrm{ST}}(x) = \frac{x^{\delta-1}(1 + x)^{-1-\delta}}{\mathtt{B}(\delta, 1)} \, \mathbb{1}_{\{x > 0\}} = \delta x^{\delta-1}(1 + x)^{-1-\delta} \, \mathbb{1}_{\{x > 0\}},$$

which is the pdf of a Beta prime distribution with parameters δ and 1. Note that $q_{\mathrm{ST}}(x) \sim x^{-2}$ as $x \to \infty$, which agrees with Kesten (1973).

Next, it can be shown that the pdf $\tilde{q}(x) = d\tilde{Q}(x)/dx$ of distribution $\tilde{Q}(x) = \mathbb{P}_0(\tilde{V}_\infty \leqslant x)$ is governed by the equation

$$\tilde{q}(x) = -\int_0^\infty \frac{\partial}{\partial x} P_0^\Lambda \left(\frac{1 + y}{x} \right) \tilde{q}(y) \, dy,$$

which can be established in a manner similar to that used to derive the above equation for $q_{\mathrm{ST}}(x)$. However, due to the symmetry of the model one can immediately conclude that $\tilde{q}(x) \equiv q_{\mathrm{ST}}(x)$, so that

$$\tilde{q}(x) = q_{\mathrm{ST}}(x) = \delta x^{\delta-1}(1+x)^{-1-\delta} \, \mathbb{1}_{\{x>0\}} \, . \tag{51}$$

We now can find

$$C_\infty = \delta \Psi_1(\delta) + \Psi_0(\delta) - \Psi_0(1),$$

where $\Psi_n(x) = d^{n+1} \log \Gamma(x)/dx^{n+1}$ ($n \geqslant 0$) is the polygamma function and $\Gamma(x)$ is the Gamma function; also note that $\Psi_0(1) = -0.577\ldots$ is the negative Euler's constant.

To find ζ and $\varkappa$, we use the formulas

$$\zeta = \frac{1}{I} \exp\left\{ -\sum_{k=1}^{\infty} \frac{1}{k} \big[\mathbb{P}_\infty(S_k > 0) + \mathbb{P}_0(S_k \leqslant 0) \big] \right\},$$

$$\varkappa = \frac{\mathbb{E}_0[Z_1^2]}{2I} - \sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}_0[S_k^-],$$

where $x^- = -\min(0,x)$; cf., e.g., (Woodroofe 1982, Chapters 2 & 3) and (Siegmund 1985, Chapter VIII). Using the work of Springer and Thompson (1970), after certain manipulations we obtain that

$$\mathbb{P}_0(S_k \leqslant 0) = \frac{1}{\Gamma^k(\delta)\Gamma^k(\delta+1)} G_{k+1,k+1}^{k+1,k} \left(1 \left| \begin{array}{c} \overbrace{-\delta,\ldots,-\delta}^{k \text{ times}},1 \\ 0, \underbrace{\delta,\ldots,\delta}_{k \text{ times}} \end{array} \right. \right), \quad k \geqslant 1,$$

where $G_{\cdot,\cdot}^{\cdot,\cdot}(\cdot|\cdot)$ is the Meijer G-function. Note that due to the symmetry of the $\mathtt{beta}(\delta,\delta+1)$-to-$\mathtt{beta}(\delta+1,\delta)$ model, $\mathbb{P}_\infty(S_k > 0) = \mathbb{P}_0(S_k \leqslant 0)$ for all $k \geqslant 1$. Hence,

$$\zeta = \delta \exp\left\{ -2 \sum_{k=1}^{\infty} \frac{1}{k\Gamma^k(\delta)\Gamma^k(\delta+1)} G_{k+1,k+1}^{k+1,k} \left(1 \left| \begin{array}{c} -\delta,\ldots,-\delta,1 \\ 0,\delta,\ldots,\delta \end{array} \right. \right) \right\},$$

which can be evaluated numerically for any $\delta > 0$ and with any desired accuracy.

Next, it can be shown that $\mathbb{E}_0[Z_1^2] = 2\Psi_1(\delta)$ and

$$\mathbb{E}_0[S_k^-] = \frac{1}{\Gamma^k(\delta)\Gamma^k(\delta+1)} G_{k+2,k+2}^{k+2,k} \left(1 \left| \begin{array}{c} -\delta,\ldots,-\delta,1,1 \\ 0,0,\delta,\ldots,\delta \end{array} \right. \right), \quad k \geqslant 1,$$

whence

$$\varkappa = \delta \Psi_1(\delta) - \sum_{k=0}^{\infty} \frac{1}{k\Gamma^k(\delta)\Gamma^k(\delta+1)} G_{k+2,k+2}^{k+2,k} \left(1 \left| \begin{array}{c} -\delta,\ldots,-\delta,1,1 \\ 0,0,\delta,\ldots,\delta \end{array} \right. \right).$$

Consider now starting the SR–r procedure off the point $R_0^r = r^*$ for which $\mathrm{ADD}_0(\mathcal{S}_A^{r^*})$ and $\mathrm{ADD}_\infty(\mathcal{S}_A^{r^*})$ are the same (at least approximately). This idea was first brought up in Subsection 5.2; recall Figure 4. By (35) and (37), when the ARL to false alarm is sufficiently large,

$$\mathrm{ADD}_\infty(\mathcal{S}_A^r) \approx \frac{1}{I}(\log A + \varkappa - C_\infty) \quad \text{and} \quad \mathrm{ADD}_0(\mathcal{S}_A^r) \approx \frac{1}{I}[\log A + \varkappa - C(r)], \tag{52}$$

where $C(r) = \mathbb{E}[\log(1+r+\tilde{V}_\infty)]$ (see (38)). Hence, equating $\mathrm{ADD}_0(\mathcal{S}_A^r)$ and $\mathrm{ADD}_\infty(\mathcal{S}_A^r)$ is equivalent to requiring $C_\infty = C(r)$, and setting R_0^r to r that solves the equation $C_\infty = C(r)$ results in the desired effect of $\mathrm{ADD}_0(\mathcal{S}_A^r) \approx \mathrm{ADD}_\infty(\mathcal{S}_A^r)$ (asymptotically). Since C_∞ is already computed, it is left to find $C(r)$. To this end, using (51), we obtain

$$C(r) = \Phi\left(\frac{r}{1+r}, 1, \delta\right) + \Psi_0(\delta) - \Psi_0(1),$$

where $\Phi(\cdot, \cdot, \cdot)$ is the Lerch transcendent. Hence, the equation $C_\infty = C(r)$, where r is the unknown, reduces to

$$\Phi\left(\frac{r}{1+r}, 1, \delta\right) = \delta\Psi_1(\delta), \quad r \geqslant 0,$$

which can be solved numerically for any desired $\delta > 0$ and with any pleased precision.

We are now in a position to perform particular computations. To remind, we would like to test the accuracy of the asymptotic approximations (35) and (36). Clearly, the accuracy is the better, the higher the desired level of the ARL to false alarm $\mathbb{E}_\infty[T] = \gamma$. First, we intend to try a relatively small value of $\gamma = 10^2$, which corresponds to practically high chances of sounding a false alarm. We do not expect the asymptotics to kick in for γ lower than a few hundreds. Suppose that $\delta = 1$. For this choice of δ we have: $C_\infty = \pi^2/6 \approx 1.64$, $I = 1$, $\zeta \approx 0.425$, $\varkappa \approx 1.25$, and $r^* \approx 2$ (so that $C(r^*) = C_\infty \approx 1.64$).

The first step is to set thresholds to guarantee the given ARL to false alarm γ. For the SR–r procedure, the detection threshold, A, should be set to the solution of the equation $\gamma = A/\zeta - r$, which follows from the corresponding asymptotics for $\mathbb{E}_\infty[\mathcal{S}_A^r]$. Since in our case $r = r^* \approx 2$ and $\zeta \approx 0.425$, we find that A must be set to about 43. The actual (evaluated numerically with very high accuracy) ARL to false alarm with this A is 100.1. Hence, the approximation $\mathbb{E}_\infty[\mathcal{S}_A^r] \approx A/\zeta - r$ is very accurate, even when $\gamma = 10^2$, which is equivalent to a relatively high risk of raising a false alarm. For the SRP procedure to have $\mathbb{E}_\infty[\mathcal{S}_A^Q] = 10^2$ the detection threshold, A, should be set to 43 as well; the actual ARL to false alarm for this choice of A is 99.6, and the mean, μ_Q, of the quasi-stationary distribution is around 2.6. Hence, the approximation $\mathbb{E}_\infty[\mathcal{S}_A^Q] \approx A/\zeta - \mu_Q$ is also very accurate.

We now proceed to examining $\mathrm{ADD}_\nu(T) = \mathbb{E}_\nu[T - \nu|T > \nu]$ as a function of $\nu \geqslant 0$ for the two procedures in consideration. Figure 6 depicts how the sequence $\mathrm{ADD}_\nu(T)$, indexed by ν, evolves as ν runs from 0 to 20 for the SR–r procedure (with $R_0^r = r^* \approx 2$) and for the SRP procedure. It can be seen that $\mathrm{ADD}_0(\mathcal{S}_A^r) \approx \mathrm{ADD}_\infty(\mathcal{S}_A^r)$, as we planned. More importantly, note that the SR–r procedure is uniformly (i.e., for all $\nu \geqslant 0$) better than the SRP rule, while the difference is small. Starting the SR–r procedure from the point that equates the average detection delays at zero and at infinity is practically more convenient, as it does not require one to know the lower bound (not to mention the quasi-stationary distribution). As this example illustrates, it may also be sufficient to outperform the SRP procedure (though for this example the gain is practically negligible).

We now turn to the accuracy of the asymptotic approximations for the average detection delays (52). According to these approximations for both procedures the worst ADD is about 2.9 (note that both procedures have the same threshold). However, the actual ADD-s are 3.54 for the SRP procedure and 3.52 for the SR–r procedure. Hence, the approximations are not too accurate, which is because the ARL to false alarm is only 100.

Consider now setting δ to 5. Since $I = 1/\delta$ this is a less contrast change than $\delta = 1$. Consequently, the ADD-s should be higher, which can be used to better illustrate the
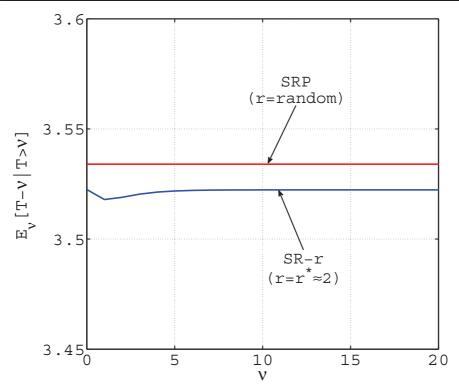
Fig. 6 Conditional average detection delay $\mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ vs. change-point ν for the SRP procedure and for the SR–r procedure with $r = r^* \approx 2$ for the beta$(\delta, \delta + 1)$-to-beta$(\delta + 1, \delta)$ model with $\delta = 1$. The ARL to false alarm $\mathbb{E}_\infty[T] = \gamma$ is approximately 100 for each procedure.

accuracy of their respective approximations. For $\delta = 5$, we have $I = 0.2$, $C_\infty \approx 3.19$, $\zeta \approx 0.685$, $\varkappa \approx 0.435$, and $r^* \approx 11$. Let $\gamma = 5 \times 10^3$. To have this level of the ARL to false alarm, the threshold for the SR–r procedure should be set to 3452 (the actual ARL to false alarm for this threshold is 4999.3), and for the SRP procedure – to 3462 (the actual ARL to false alarm for this threshold is 5000.1, and $\mu_Q \approx 26.1$). Again, both approximations $\mathbb{E}_\infty[\mathcal{S}_A^r] \approx A/\zeta - r$ and $\mathbb{E}_\infty[\mathcal{S}_A^Q] \approx A/\zeta - \mu_Q$ are highly accurate. We now look at the delays. Figure 7 shows the average delay to detection $\mathrm{ADD}_\nu(T)$ versus the changepoint ν for the SR–r procedure with $R_0^r = r^* \approx 11$ and for the SRP procedure. It can be seen that again $\mathrm{ADD}_0(\mathcal{S}_A^r) \approx \mathrm{ADD}_\infty(\mathcal{S}_A^r)$. Furthermore, the SR–$r$ procedure is almost an equalizer: there is a tiny mound raising above the SRP's flat line, though the mound is comparable in magnitude to the numerical error, and therefore, can be disregarded from a practical point of view. Both procedures are equally efficient, but since the SR–r procedure is easier to initialize it is preferable for practical purposes.

In terms of the accuracy the actual $\mathrm{ADD}_0(\mathcal{S}_A^r)$ is 27, while that predicted by the approximation is 27. The actual value of $\mathrm{ADD}_\infty(\mathcal{S}_A^r)$ is 27.1 versus the approximated value 27 (which is the same as the value predicted for $\mathrm{ADD}_0(\mathcal{S}_A^r)$, because $C(r^*) = C_\infty$). Lastly, for the SRP procedure the actual average delay is 27.1, while the predicted using the asymptotic approximation value is 27. As we can see, the approximations for the ADD-s are now accurate. The reason is that the ARL to false alarm is relatively high.
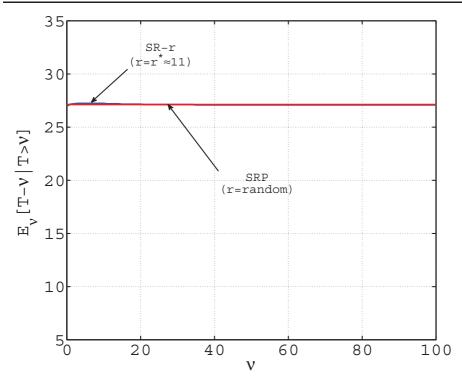
Fig. 7 Conditional average detection delay $\mathbb{E}_\nu[\mathcal{S}_A^r - \nu | \mathcal{S}_A^r > \nu]$ vs. change-point ν for the SRP procedure and for the SR–r procedure with $r = r^* \approx 11$ for the $\mathtt{beta}(\delta, \delta+1)$-to-$\mathtt{beta}(\delta+1, \delta)$ model with $\delta = 5$. The ARL to false alarm $\mathbb{E}_\infty[T] = \gamma$ is approximately 5×10^3 for each procedure.

To draw a line under this example, the main conclusion is that the SR–r procedure is almost equalizer, and its performance is almost indistinguishable from that of the SRP procedure. However, it is easier to implement in practice, which is contrary to the SRP procedure. Hence, we recommend the SR–r procedure for practical purposes.

8.2 Example 2: An exponential scenario

Suppose the sequence $\{X_n\}_{n \geqslant 1}$ is comprised by the exponentially distributed random variables that undergo a shift in the mean from 1 to $1 + \theta$, where $\theta > 0$. Formally, the pre- and post-change densities in this case are

$$f(x) = \exp\{-x\}\, \mathbb{1}_{\{x \geqslant 0\}} \quad \text{and} \quad g(x) = \frac{1}{1+\theta} \exp\left\{-\frac{x}{1+\theta}\right\} \mathbb{1}_{\{x \geqslant 0\}},$$

respectively. We refer to this model as the $\mathcal{E}(1)$-to-$\mathcal{E}(1+\theta)$ model.

This model was considered by Tartakovsky et al (2009) for $\theta = 0.1$, which corresponds to a small, not easily detectable change. Using the numerical framework of Moustakides et al (2011), also presented in Section 6, they carried out a performance analysis of CUSUM, the SRP procedure and the SR–r procedure comparing each against the other. They also computed the lower bound. We present an excerpt of results for the SRP and SR–r procedures along with the lower bound. The accuracy is within 0.5%.

Figure 8 shows operating characteristics in terms of Pollak's supremum conditional average detection delay $\mathcal{J}_{\mathrm{P}}(T) = \sup_\nu \mathbb{E}_\nu[T - \nu | T > \nu]$ as a function of the ARL to false alarm $\mathbb{E}_\infty[T] = \gamma$, plus the lower bound $\mathcal{J}_{\mathrm{B}}(T)$. It can be seen that the best performance is delivered by the SR–r procedure. This is expected since by design the SR–r rule is the closest to the lower bound $\mathcal{J}_{\mathrm{B}}(T)$. This suggests that the (unknown) optimal procedure can offer only a practically insignificant improvement over the SR–r procedure.

Next, Figure 9 shows the behavior of the stationary average detection delay $\mathcal{J}_{\mathrm{ST}}(T)$ against the ARL to false alarm. Since the SR procedure is exactly optimal with respect to $\mathcal{J}_{\mathrm{ST}}(T)$ its performance is the best among the three procedures, but the difference is relatively small. Note also that for the SRP procedure $\mathcal{J}_{\mathrm{P}}(\mathcal{S}_A^Q)$ is the same as $\mathcal{J}_{\mathrm{ST}}(\mathcal{S}_A^Q)$, since the SRP procedure is an equalizer.

References

Atkinson K, Han W (2009) Theoretical Numerical Analysis: A Functional Analysis Framework, Texts in Applied Mathematics, vol 39, 3rd edn. Springer, DOI 10.1007/978-1-4419-0458-4

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

Brodsky BE, Darkhovsky BS (1993) Nonparameteric methods in change point problems, Mathematics and Its Applications, vol 243. Kluwer Academic Publishers

Feinberg EA, Shiryaev AN (2006) Quickest detection of drift change for Brownian motion in generalized Bayesian and minimax settings. Statistics & Decisions 24(4):445–470, DOI 10.1524/stnd.2006.24.4.445

Ferguson TS (1967) Mathematical Statistics – A Decision Theoretic Approach. Academic Press, New York

Fuh CD (2003) SPRT and CUSUM in hidden Markov models. The Annals of Statistics 31(3):942–977, DOI 10.1214/aos/1056562468

Fuh CD (2004) Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. The Annals of Statistics 32(5):2305–2339, DOI 10.1214/009053604000000580

Girschick MA, Rubin H (1952) A Bayes approach to a quality control model. The Annals of Mathematical Statistics 23(1):114–125, DOI 10.1214/aoms/1177729489

Harris TE (1963) The Theory of Branching Processes. Springer-Verlag, Berlin

Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Mathematica 131(1):207–248, DOI 10.1007/BF02392040

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

Lai TL (1998) Information bounds and quick detection of parameter changes in stochastic systems. IEEE Transactions on Information Theory 44:2917–2929, DOI 10.1109/18.737522

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

Mevorach Y, 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

Moustakides GV (1986) Optimal stopping times for detecting changes in distributions. The Annals of Statistics 14(4):1379–1387

Moustakides GV (2008) Sequential change detection revisited. The Annals of Statistics 36(2):787–807, DOI 10.1214/009053607000000938

Moustakides GV, Polunchenko AS, Tartakovsky AG (2011) A numerical approach to performance analysis of quickest change-point detection procedures. Statistica Sinica 21(2):571–596

Page ES (1954) Continuous inspection schemes. Biometrika 41(1):100–115

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

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

Pollak M, Siegmund D (1986) Convergence of quasi-stationary to stationary distributions for stochastically monotone Markov processes. Journal of Applied Probability 23(1):215–220

Pollak M, Tartakovsky AG (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
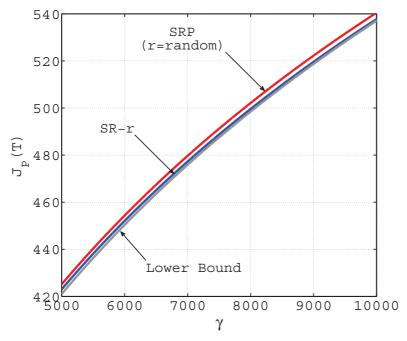
Fig. 8 The lower bound $\mathcal{J}_\mathrm{B}(T)$ and Pollak's $\mathcal{J}_\mathrm{P}(T)$ for the SRP and SR–r procedures for the $\mathcal{E}(1)$-to-$\mathcal{E}(1+\theta)$ model with $\theta = 0.1$. The ARL to false alarm is between 5×10^3 and 10^4.

Fig. 9 The stationary average detection delay $\mathcal{J}_\mathrm{ST}(T)$ for the SRP and SR–r procedures for the $\mathcal{E}(1)$-to-$\mathcal{E}(1+\theta)$ model with $\theta = 0.1$. The ARL to false alarm is between 5×10^3 and 10^4.
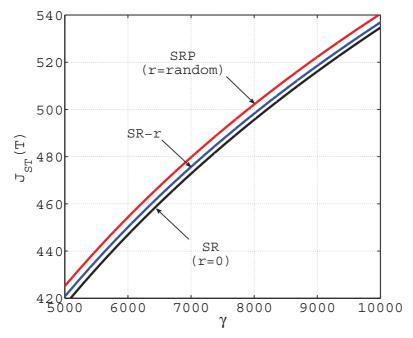
Pollak M, Tartakovsky AG (2009b) Optimality properties of the Shiryaev-Roberts procedure. Statistica Sinica 19:1729–1739

Polunchenko AS, Tartakovsky AG (2010) On optimality of the Shiryaev-Roberts procedure for detecting a change in distribution. The Annals of Statistics 36(6):3445–3457, DOI 10.1214/09-AOS775

Poor HV, Hadjiliadis O (2008) Quickest Detection. Cambridge University Press

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

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

Shewhart WA (1931) Economic control of quality of manufactured product. D. Van Nostrand Company, Inc., New York

Shiryaev AN (1961) The problem of the most rapid detection of a disturbance in a stationary process. Soviet Math Dokl 2:795–799

Shiryaev AN (1963) On optimum methods in quickest detection problems. Theory of Probability and Its Applications 8(1):22–46, DOI 10.1137/1108002

Shiryaev AN (1978) Optimal Stopping Rules. Springer-Verlag, New York

Shiryaev AN (2006) From "disorder" to nonlinear filtering and martingale theory. In: Bolibruch A, Osipov Y, Sinai Y (eds) Mathematical Events of the Twentieth Century, Springer Berlin Heidelberg, pp 371–397, DOI 10.1007/3-540-29462-7_18

Shiryaev AN (2009) On the stochastic models and optimal methods in the quickest detection problems. Theory of Probability and Its Applications 53(3):385–401, DOI 10.1137/S0040585X97983717

Shiryaev AN (2010) Quickest detection problems: Fifty years later. Sequential Analysis 29:345–385, DOI 10.1080/07474946.2010520580

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

Springer MD, Thompson WE (1970) The distribution of products of Beta, Gamma and Gaussian random variables. SIAM Journal on Applied Mathematics 18(4):721–737, DOI 10.1137/0118065

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

Tartakovsky AG (2005) Asymptotic performance of a multichart CUSUM test under false alarm probability constraint. In: Proceedings of the 2005 IEEE Conference on Decision and Control, vol 44, pp 320–325

Tartakovsky AG (2008) Discussion on "Is average run length to false alarm always an informative criterion?" by Yajun Mei. Sequential Analysis 27(4):396–405, DOI 10.1080/07474940802446046

Tartakovsky AG (2009a) Asymptotic optimality in Bayesian changepoint detection problems under global false alarm probability constraint. Theory of Probability and Its Applications 53:443–466, DOI 10.1137/S0040585X97983754

Tartakovsky AG (2009b) Discussion on "Optimal sequential surveillance for finance, public health, and other areas" by Marianne Frisén. Sequential Analysis 28(3):365–371, DOI 10.1080/07474940903041704

Tartakovsky AG, Moustakides GV (2010) State-of-the-art in Bayesian changepoint detection. Sequential Analysis 29(2):125–145, DOI 10.1080/07474941003740997

Tartakovsky AG, Polunchenko AS (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 AG, Veeravalli VV (2005) General asymptotic Bayesian theory of quickest change detection. Theory of Probability and Its Applications 49(3):458–497, DOI 10.1137/S0040585X97981202

Tartakovsky AG, Pollak M, Polunchenko AS (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 AG, Polunchenko AS, Moustakides GV (2009) Design and comparison of Shiryaev–Roberts- and CUSUM-type change-point detection procedures. In: Proceedings of the 2nd International Workshop in Sequential Methodologies, University of Technology of Troyes, Troyes, France

Tartakovsky AG, Pollak M, Polunchenko AS (2011) Third-order asymptotic optimality of the generalized Shiryaev-Roberts changepoint detection procedures. Theory of Probability and Its Applications

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

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