Sequential Analysis, 29: 386–393, 2010

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474946.2010.520581

Discussion on "Quickest Detection Problems: Fifty Years Later" by Albert N. Shiryaev

Alexander G. Tartakovsky[1] **and George V. Moustakides**[2]
[1]Department of Mathematics, University of Southern California, Los Angeles, California, USA
[2]Department of Electrical and Computer Engineering, University of Patras, Patras, Greece

Abstract: Quickest detection is a fascinating area of sequential analysis that spans across various branches of science and engineering. It is a pleasure to welcome Professor Albert Shiryaev's article, which provides a comprehensive overview (both scientific and historic) of this area. In this discussion, we expand on some of the issues raised in the article that we believe require further elaboration.

Keywords: Average detection delay; Average run length to false alarm; Change-point detection; Cusum; False alarm rate; Shiryaev–Roberts procedure.

Subject Classifications: 62L15; 60G40; 62F12; 62F15.

1. OPTIMALITY PROPERTIES OF THE SHIRYAEV–ROBERTS PROCEDURE

Shiryayev (1963) was the first to propose the "multi-cycle" problem of detecting a change that takes place in a distant future and is preceded by a stationary flow of false alarms. The analysis of the problem for the Brownian Motion (BM) case, which is presented in Sections 5 and 6 of the review article, led to the natural definition of the Stationary Average Delay to Detection (STADD) as the appropriate performance measure. This criterion was then optimized within the class $\{\tau : \mathsf{E}_\infty \tau \geq T\}$ of procedures for which the average run length (ARL) between

Received May 27, 2010, Revised June 1, 2010, Accepted August 31, 2010
Recommended by Nitis Mukhopadhyay
Address correspondence to Alexander G. Tartakovsky, Department of Mathematics, University of Southern California, KAP-108, Los Angeles, CA 90089-2532, USA; E-mail: tartakov@math.usc.edu or George V. Moustakides, Department of Electrical and Computer Engineering, University of Patras, Patras 26500, Rio, Greece; E-mail: moustaki@upatras.gr

false alarms is forced to exceed some prescribed value $T > 0$. The optimization gave rise to one of the two most popular change detection strategies, known as the Shiryaev–Roberts (SR) procedure (see Roberts, 1966, for the discrete-time case). The SR procedure and related optimality results are presented in a very lucid and detailed way in the review article. In particular, Theorem 5.1 establishes the optimality of the multi-cycle SR procedure in detecting a change in the drift of the Brownian motion that occurs in a distant time-horizon, provided that the procedure starts anew after each false alarm.

Prof. Shiryaev also relates the STADD to the generalized Bayesian problem, namely the minimization of the relative integral average delay $T^{-1}\int_0^\infty \mathsf{E}_\theta(\tau-\theta)^+ d\theta$ over procedures that satisfy $\mathsf{E}_\infty\tau = T$. Observing that the two objectives in the generalized Bayesian problem and in the multi-cycle detection actually coincide allows for the immediate conclusion that the SR procedure is optimal not only for the latter but also for the former problem.

Prof. Shiryaev details the important applied problem of sequentially detecting an abruptly appearing target with the help of a radar system. This problem is clearly one application area where this mathematical formulation is relevant. Many surveillance applications where an "intruder" must be detected as soon as possible exhibit very similar features. Indeed, an intrusion is usually preceded by a long interval of surveillance with multiple false alarms. This scenario of course presumes a *low cost* of false alarms and, therefore, the existence of an additional (higher level) mechanism capable of distinguishing between false alarms and true detections. A modern and challenging application where such ideas can find fertile grounds is the problem of network security—rapid detection of computer attacks based on real-time analysis of network traffic. Change-point detection methods are very efficient in providing small detection delays if detection thresholds can be lowered. However, this will unavoidably lead to an intense flow of false alarms. False detections can then be filtered by an independent algorithm, e.g., by spectral analysis. The implication of Shiryaev's results is that one has reason to prefer the multi-cycle SR procedure over other multi-cycle surveillance schemes, including CUSUM.

Continuous-time/Gaussian analysis presents definite mathematical advantages and, most of the time, offers fascinating results. However, in the majority of applications, observations are obtained discretely with a certain sampling rate; furthermore, the Gaussian assumption does not necessarily apply. It is fortunate that results similar to the ones obtained for the BM, also apply in the general discrete-time model case. This fact has been recently established by Pollak and Tartakovsky (2009). Specifically, the discrete-time version of the SR procedure minimizes the integral average delay to detection $(\mathsf{E}_\infty\tau)^{-1}\sum_{\theta=0}^{\infty}\mathsf{E}_\theta(\tau-\theta)^+$ in the class of procedures with $\mathsf{E}_\infty\tau \geq T$, where $T > 1$. As in continuous-time, the integral average detection delay is equivalent to the STADD (the limit as $\theta \to \infty$ of the average delay to detection of a repeated detection procedure). Consequently, the aforementioned double optimality property of the SR procedure for continuous time is also valid in discrete time.

The multi-cycle formulation of the change detection problem is definitely interesting, mathematically appealing and useful in a multitude of applications. We would like to point out that it relies on a particular assumption that the change will take place at an unknown time in the *distant future*. The resulting optimization criterion and the optimal procedure (i.e., the SR procedure) are clearly tailored to this specific assumption.

A different point of view is proposed in the single-run Bayesian formulation of the problem in Section 6. We note that due to the exponential decay of the prior distribution the change is more likely to occur in the *near* rather than in the far future as long as the parameter of the exponential prior λ is not very small. Consequently, the emphasis may be placed on changes that occur relatively soon after surveillance begins. In fact, by appropriately selecting the exponential decay rate λ of the prior distribution, we can control the expected value of the changepoint. In addition, we note that a multi-cycle setting can also be considered in a Bayesian context, in which case a multi-cycle Bayesian Shiryaev procedure is optimal in the sense of minimizing the average detection delay for a given average number of false alarms. As $\lambda \to 0$, this Bayesian procedure converges to the multi-cycle SR procedure.

Both approaches are based on specific prior assumptions (a change occurring in a distant future and a low cost of false alarms for the multi-cycle setup with no prior distribution or knowledge of the prior distribution in the single-run Bayesian approach). If such knowledge is available, one should certainly use the corresponding optimal procedures. However, when no a priori information regarding the time of change exists or when the change can occur at any time instant with no emphasis on any particular time interval, clearly, both procedures lose their optimality properties. In this case one should resort to detection strategies that can respond efficiently to all possible changes. The minimax formulations (Variants C and D), mentioned in Sections 8 and 9 of the review article, are capable of handling such possibilities. The most popular detection strategy of this type is the CUSUM procedure, which is *exactly* optimal in the sense described by Variant D. Regarding the SR procedure; interestingly, one can find simple modifications that affect only the *initialization* of the SR statistic, which exhibit a very strong *asymptotic* minimax property in the sense of Variant C. In Section 3 we briefly describe two such modifications proposed by Pollak (1985) and Moustakides et al. (2011).

In conclusion, it is evident from the previous discussion that the SR procedure possesses great versatility. Furthermore, changing the initialization of the SR statistic from zero (standard SR) to a specially designed point allows us to vary the desired performance, e.g., to achieve a fast initial response to changes occurring soon after surveillance begins or to get the best possible performance in worst-case scenarios (cf. Moustakides et al., 2011; Pollak, 1985). In other words, depending on the available prior knowledge about the time of change, it is possible to come up with a version of the SR procedure that can respond to the change in a very effective way. Similar flexibility is not encountered in other detection structures.

1.1. Comparison of CUSUM and Shiryaev–Roberts Procedures

When comparing the operating characteristics of the optimal SR procedure with a similar multi-cycle CUSUM, Prof. Shiryaev assumes (for the sake of simplicity) that the "signal-to-noise ratio" $q = r^2/2\sigma^2 = 1$, in which case the difference between the stationary average delays to detection, for the same ARL to false alarm T, is approximately equal to $\mathbf{C} - 0.5 \approx 0.077$ for large T, i.e., it is negligible. If $q \neq 1$, the difference becomes $(\mathbf{C} - 0.5)/q \approx 0.077/q$, which can take non-negligible values for small values of q corresponding to low-contrast changes. Note that the relative difference (asymptotic efficiency) is $0.077/\log T$, i.e., both procedures perform almost identically.

On the other hand, CUSUM is minimax with respect to Lorden's essential supremum criterion and, therefore, it outperforms the SR procedure for $\theta = 0$ (the worst point for both procedures in terms of the conditional average delays $\mathsf{E}_\theta(\tau - \theta \mid \tau \geq \theta)$). In this case it is assumed that there is only a single run. The difference between $\mathsf{E}_0\tau$ for CUSUM and SR for large T is approximately equal to $(1 - \mathbf{C})/q \approx 0.427/q$, and again the relative efficiency is $0.427/\log T$, i.e., negligible.

1.2. Comparison of the Shiryaev–Roberts and Neyman–Pearson Multi-Cycle Procedures

In Section 3 of the review article, Prof. Shiryaev analyzes the multi-cycle, multi-stage Neyman–Pearson (NP) procedure with a fixed batch size m and shows that if $m \sim \log T$, where T is the ARL to false alarm constraint value, then with a certain optimal selection of the threshold the average detection delay of this procedure is asymptotically equal to $\frac{3}{2}\log T$ as $T \to \infty$. This suggests that the average delay is 1.5 times larger than the corresponding stationary average delay to detection of the CUSUM and SR procedures. This result, however, is true *on the average* after imposing a uniform prior distribution on the time of change in the interval $[0, m]$. Specifically, $m^{-1}\int_0^m \mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta)d\theta \sim 1.5q^{-1}\log T$ as $T \to \infty$, where τ_{NP} denotes the stopping time of the NP procedure.

One may be interested in evaluating the conditional average detection delay $\mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta)$ for all $\theta \in [0, m]$ and identifying the least favorable changepoint that maximizes $\mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta)$. Intuitively, if the change occurs close to the beginning or the end of the interval $[(k-1)m, km]$, then the NP procedure will detect it with an average delay of $q^{-1}\log T$. If it occurs in the vicinity of the mid-point of the batch, then it will take an average number of $1.5q^{-1}\log T$ observations to detect the change. Therefore, we expect that the NP procedure will perform as well as the CUSUM or SR procedures (to a first order in T as $T \to \infty$) when the change takes place at the beginning or at the end of a batch, but it will perform 1.5 times worse if the change appears near the mid-point of a batch. However, the mid-point is not the worst point for the NP procedure at which the conditional average detection delay $\mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta)$ attains maximum.

A detailed analysis of the behavior of the conditional average detection delay $\mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta)$ of the NP procedure has been performed by Tartakovsky (1992) for the multi-channel detection problem. It follows from this work that, asymptotically as $T \to \infty$, the optimal batch size is $m \sim q^{-1}\log(qT)$ (for any point of change) and the worst point that maximizes the average detection delay is $\theta^* = \theta_T^* = q^{-1}(2\log T \log\log T)^{1/2}$ delivering

$$\sup_{\theta\in[0,m]} \mathsf{E}_\theta(\tau_{\mathrm{NP}} - \theta \mid \tau_{\mathrm{NP}} \geq \theta) \sim \mathsf{E}_{\theta^*}(\tau_{\mathrm{NP}} - \theta^* \mid \tau_{\mathrm{NP}} \geq \theta^*) \sim 2q^{-1}\log T, \quad T \to \infty.$$

Comparing the worst performance of the NP procedure to the optimum (which is attained by the CUSUM and SR procedures)

$$\inf_{\tau:\mathsf{E}_\infty\tau\geq T}\ \sup_{0\leq\theta<\infty} \mathsf{E}_\theta(\tau - \theta \mid \tau \geq \theta) = q^{-1}\log T + O(1), \quad T \to \infty,$$

we realize that the NP procedure exhibits a 100% increase in the worst average detection delay when compared to the optimal procedure. Similar asymptotic results hold in the discrete-time Gaussian case when detecting a change in the mean.

Improvement in performance can be obtained when the window of size m is advanced continuously (sliding window) with every new observation. In this case one achieves optimal performance to a first order in T (cf. Lai, 1998; Tartakovsky, 2005). This latter procedure presents an advantage over CUSUM and the SR procedures because it also allows for the efficient control of the local false alarm probability (in the window of size m) in addition to the usual ARL to false alarm.

2. MODIFICATIONS OF THE SHIRYAEV–ROBERTS PROCEDURE

When there is no prior knowledge about the change time, a natural way to tackle the change detection problem is through a minimax formulation. Prof. Shiryaev recalls the two approaches that are currently available in the literature (Variants C and D). As we mentioned previously, for the i.i.d. (and the BM) case, the CUSUM procedure is optimal in Variant D (i.e., with respect to Lorden's essential supremum average detection delay). Unfortunately, similar exact optimality results for Variant C (i.e., with respect to Pollak's supremum average detection delay) are still lacking. However, there are two simple modifications of the SR procedure that solve Variant C asymptotically in a very strong sense.

In Section 7 of the review article, Prof. Shiryaev is treating the BM case. The statistic ψ_t of the SR procedure is defined according to the following SDE: $d\psi_t = dt + (r/\sigma^2)\psi_t dX_t$, $\psi_0 = 0$. The equivalent formula for i.i.d. observations in discrete time has the form $R_n = (1 + R_{n-1})(f^0(X_n)/f^\infty(X_n))$, $R_0 = 0$. As we can see, the classical version of the SR statistic starts from 0. The two modifications that we introduce below simply suggest a different initialization strategy, while keeping the recursion formulas intact.

Pollak (1985) considered the discrete-time case and introduced a very specific *randomized* initialization of the SR statistic where R_0 is sampled from the *quasi-stationary* distribution of the SR statistic. By letting the false alarm constraint parameter $T \to \infty$ the performance of the proposed modification and the unknown optimum tend to ∞ at a rate $O(\log T)$, whereas their difference tends to 0. In other words, the performance of the modified SR procedure is equal to the optimum plus a term $o(1)$ that tends to 0 as $T \to \infty$. This form of asymptotic optimality is clearly very strong since it assures that the randomly initialized SR procedure is almost optimal.

For a number of years, this interesting result could not be practically appreciated due to lack of a proper technique that could compute the quasi-stationary distribution. Moustakides et al. (2011), for the first time, developed a simple numerical method capable of achieving this goal. In the same article, an alternative *deterministic* initialization strategy was proposed, i.e., starting the SR statistic R_n off $R_0 = r$, where $r > 0$ is a specially designed point depending on the ARL to false alarm constraint T. The numerical methods for computing the performance of the SR procedure with an arbitrary deterministic initialization have been also developed. This methodology was applied to several particular examples (mostly Gaussian and exponentially distributed observations). By using a proper deterministic initialization $R_0 = r > 0$ of the SR statistic, it was possible to obtain performance comparable and even (slightly) better than Pollak's randomized modification. This fact naturally raised the conjecture that there exist deterministic initializations of the SR procedure that enjoy the same strong asymptotic optimality property as Pollak's randomized counterpart. Recently, Tartakovsky et al. (2010) provided an analytic proof of this fact.

An extension of the aforementioned discrete-time results to the continuous-time BM case is currently under development. As far as the deterministic initialization is concerned, the goal is to provide an *analytic formula* for the appropriate initializing value $\psi_0 > 0$ of the SR statistic $\{\psi_t\}$ that will assure a similar strong asymptotic optimality property in Variant C, as in the discrete-time case.

3. GENERAL STOCHASTIC MODELS

In Section 10, Prof. Shiryaev tackles the case of non-i.i.d. observations. This is a rather delicate subject that is susceptible to different interpretations and approaches. We believe that the key point in analyzing the general case is the specification of the probability measure induced by a change occurring at a *deterministic* time instant θ. For simplicity, we will assume that all probability measures are mutually absolutely continuous and can be expressed through densities. Let $f^{\alpha}(X_1, \dots, X_n)$, $\alpha = \infty, 0$ be the pre- and post-change densities of the observed data $X_1, \dots, X_n$. To define the pdf $f^{\theta}(X_1, \dots, X_n)$ induced by a change occurring at time $\theta \geq 0$ note that when $n < \theta$ the observations are under the pre-change regime, so that $f^{\theta}(X_1, \dots, X_n) = f^{\infty}(X_1, \dots, X_n)$. When, however, $\theta \leq n$ there are different possibilities. One way is

$$f^{\theta}(X_1, \dots, X_n) = f^{\infty}(X_1, \dots, X_{\theta-1}) \times f^{0}(X_{\theta}, \dots, X_n \mid X_1, \dots, X_{\theta-1}), \tag{3.1}$$

according to which the observations before the change affect the observations after the change through the *conditional* post-change pdf. We stress that in (3.1) the form of the conditional post-change pdf is assumed to be *independent of the changepoint* θ.

The change model introduced in (3.1) is very appealing because conventional statistics, such as CUSUM and SR, can be updated recursively analogously to the i.i.d. case. Indeed, we only need to replace the likelihood ratio $f^{0}(X_n)/f^{\infty}(X_n)$ of the i.i.d. case with the conditional likelihood ratio

$$f^{0}(X_n \mid X_1, \dots, X_{n-1})/f^{\infty}(X_n \mid X_1, \dots, X_{n-1})$$

in order to cover the dependent data case. Characteristic examples are presented in Section 10 of the review article. Although it is not very apparent, Prof. Shiryaev actually adopts the change model defined in (3.1), which leads to the recursion formulas (10.13), (10.14), (10.15), and (10.18).

The model in (3.1) is undoubtedly intriguing but, unfortunately, fails to cover several classical data models. Consider, for example, a sequence $\{(X_n, Z_n)\}$ of random pairs where $\{X_n\}$ is the observed component and $\{Z_n\}$ is an unobservable *hidden* component. Let $f^{\alpha}(X_1, \dots, X_n, Z_1, \dots, Z_n), \alpha = \infty, 0,$ denote the pre- and post-change joint density. According to (3.1) the joint density induced by a change occurring at time θ is

$$\begin{aligned} f^{\theta}(X_1, \dots, X_n, Z_1, \dots, Z_n) &= f^{\infty}(X_1, \dots, X_{\theta-1}, Z_1, \dots, Z_{\theta-1}) \\ &\quad \times f^{0}(X_{\theta}, \dots, X_n, Z_{\theta}, \dots, Z_n \mid X_1, \dots, X_{\theta-1}, Z_1, \dots, Z_{\theta-1}). \end{aligned}$$

Using the joint pdf (3.2) we can now compute $f^{\theta}(X_1, \ldots, X_n)$, the pdf of the observations induced by the change, by simply integrating out the hidden variables $Z_1, \ldots, Z_n$. Unfortunately, the resulting pdf *does not* conform with (3.1) even though the original joint pdf was selected to satisfy this model. The reason is that after integration over hidden states the conditional post-change pdf $f_{\theta}^{0}(X_{\theta}, \ldots, X_n \mid X_1, \ldots, X_{\theta-1})$ *depends on the changepoint* θ. Note that in this case the pre-change model also affects the post-change distribution.

Clearly, in such cases the recursive updating formulas introduced in the review article are not valid. A classical example is the case of i.i.d. observations whose pdf is controlled by a HMM. Even if the transition matrix of the HMM stays the same before and after the change, the observation pdf $f^{\theta}(X_1, \ldots, X_n)$ does not satisfy (3.1), as can be easily verified.

In fact, the most general model that covers all imaginable scenarios, including hidden models (Markov and non-Markov) is

$$f^{\theta}(X_1, \ldots, X_n) = \prod_{i=1}^{\theta-1} f^{\infty}(X_i \mid X_1, \ldots, X_{i-1}) \times \prod_{i=\theta}^{n} f_{\theta}^{0}(X_i \mid X_1, \ldots, X_{i-1}),$$

where $f^{\infty}(X_i \mid X_1, \ldots, X_{i-1})$ and $f_{\theta}^{0}(X_i \mid X_1, \ldots, X_{i-1})$ are the conditional pre- and post-change densities. The latter may depend on the changepoint θ and may also be affected by the pre-change distribution, which is usually true for HMMs (cf., e.g., Tartakovsky, 2009; Tartakovsky and Veeravalli, 2005). In such cases implementation of standard detection procedures, such as CUSUM and SR, becomes burdensome due to the time-consuming computations that need to be performed in real time. Window-limited versions of the CUSUM and SR procedures with an optimally selected window length may constitute a practical alternative. Indeed, such variations are computationally feasible and at the same time enjoy certain asymptotic optimality properties (cf. Lai, 1998).

REFERENCES

Lai, T. L. (1998). Information Bounds and Quick Detection of Parameter Changes in Stochastic Systems, *IEEE Transactions on Information Theory* 44: 2917–2929.

Moustakides, G. V., Polunchenko, A. S., and Tartakovsky, A. G. (2011). A Numerical Approach to Comparative Efficiency Analysis of Quickest Change-Point Detection Procedures, *Statistica Sinica* 21. ArXiv: stat.CO/0907.3521v2 (8 Dec. 2009).

Pollak, M. (1985). Optimal Detection of a Change in Distribution, *Annals of Statistics* 13: 206–227.

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

Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures, *Technometrics* 8: 411–430.

Shiryaev, A. N. (1963). On Optimum Methods in Quickest Detection Problems, *Theory of Probability and Its Applications* 8: 22–46.

Tartakovsky, A. G. (1992). Efficiency of the Generalized Neyman–Pearson Test for Detecting Changes in a Multichannel System, *Problems of Information Transmission* 28: 341–350.

Tartakovsky, A. G. (2005). Asymptotic Performance of a Multichart CUSUM Test Under False Alarm Probability Constraint, in *Proceedings of 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05)*, December 12–15, 2005, pp. 320–325, Seville, Spain: Omnipress CD-ROM, ISBN 0-7803-9568-9.

Tartakovsky, A. G. (2009). Asymptotic Optimality in Bayesian Changepoint Detection Problems Under Global False Alarm Probability Constraint, *Theory of Probability and Its Applications* 53: 443–466.

Tartakovsky, A. G. and Veeravalli, V. V. (2005). General Asymptotic Bayesian Theory of Quickest Change Detection, *Theory of Probability and Its Applications* 49: 458–497.

Tartakovsky, A. G., Pollak, M., and Polunchenko, A. S. (2010). Third-Order Asymptotic Optimality of the Generalized Shiryaev–Roberts Changepoint Detection Procedures, *Bernoulli*, submitted. ArXiv: math.ST/1005.1129v1 (7 May 2010).