Reducing sequential change detection to sequential estimation

Shubhanshu Shekhar [1] **Aaditya Ramdas** [1 2]

Abstract

We consider the problem of sequential change detection under minimal assumptions on the distribution generating the stream of observations. Formally, our goal is to design a scheme for detecting any changes in a parameter or functional θ of the data stream distribution that has small detection delay, but guarantees control on the frequency of false alarms in the absence of changes. We describe a simple reduction from sequential change detection to sequential estimation using confidence sequences (CSs): begin a new level-$(1-\alpha)$ CS at each time step, and proclaim a change as soon as the intersection of all active CSs becomes empty. We prove that the average run length of our scheme is at least $1/\alpha$, resulting in a change detection scheme with minimal structural assumptions (thus allowing for possibly dependent observations, and nonparametric distribution classes), but strong guarantees. We also describe an interesting parallel with Lorden's reduction from change detection to sequential testing and connections to the recent "e-detector" framework.

1. Introduction

We consider the following problem of *sequential change detection* (SCD): for a general space $\mathcal{X}$, given a stream of $\mathcal{X}$-valued observations $X_1, X_2, \ldots$, our goal is to design a method to detect any changes in a prespecified parameter or functional θ (possibly infinite dimensional) associated with the source generating this stream. Let $\mathcal{P}$ denote a class of probability distributions on the infinite product space $\Omega = \mathcal{X}^\infty$, and let $\theta : \mathcal{P} \to \Theta$, denote a mapping from probability distributions to some (possibly infinite dimensional) parameter space Θ. The data distribution satisfies the following for some $T \in \mathbb{N} \cup \{\infty\}$:

- For $n \leq T$, the observations are the first T elements of a trajectory of $P_0 \in \mathcal{P}$, with $\theta(P_0) = \theta_0$.
- For $n > T$, the observations are drawn from another distribution $P_1 \in \mathcal{P}$, such that $\theta(P_1) = \theta_1 \neq \theta_0$.

In particular, P_0, P_1 are distributions over an infinite sequence of observations, and the data is not assumed to be i.i.d. (independent and identically distributed), meaning that we do not assume that P_0, P_1 take the form p^∞ for some distribution p over $\mathcal{X}$. Further, P_1 (and hence θ_1) is allowed to depend on $X_1, \ldots, X_T$. We refer to the term T as the *changepoint*, and to P_0 and P_1 as the *pre-* and *post-change* distributions respectively.

Under the above model, our goal is to design a data-driven method for detecting any change in the value of θ, without any knowledge of T, P_0, P_1. In technical terms, our objective is to define a stopping time τ, at which we stop collecting more data, and declare a changepoint has previously occurred.

There are two problem settings to consider: non-partitioned and partitioned. By default, define the filtration $\mathcal{F} \equiv (\mathcal{F}_n)_{n\geq 0}$, where $\mathcal{F}_n$ is by default the sigma algebra $\sigma(X_1, \ldots, X_n)$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

Problem 1.1 (Non-partitioned SCD)**.** *For some unknown triple (T, P_0, P_1), suppose $X_1, X_2, \ldots$ denote a data stream, such that $(X_n)_{n\leq T}$ are drawn from P_0, and $(X_n)_{n>T}$ are drawn from P_1, such that $\theta(P_1) = \theta_1 \neq \theta_0 = \theta(P_0)$. The goal is to define a stopping time τ, adapted to the filtration $\mathcal{F}$ satisfying:*

- *If $T = \infty$, and there is no changepoint, we require that $\mathbb{E}_\infty[\tau] \geq 1/\alpha$, for a prespecified $\alpha \in (0, 1)$. The term $\mathbb{E}_\infty[\tau]$ is called the average run length (ARL), and represents the frequency of false alarms.*
- *If $T < \infty$, and there is a changepoint at which the distribution changes from P_0 to P_1, we desire the detection delay, $\mathbb{E}_T[(\tau - T)^+]$, to be as small as possible.*

The non-partitioned SCD problem stated above, does not require any pre-specified partitioning of $\mathcal{P}$ into pre- and post change distribution classes. That is, it assumes that the data generating distribution changes from one unknown

[1]Department of Statistics and Data Science, Carnegie Mellon University [2]Machine Learning Department, Carnegie Mellon University. Correspondence to: Shubhanshu Shekhar <shubhan2@andrew.cmu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

distribution P_0 in $\mathcal{P}$ to another unknown distribution P_1 in $\mathcal{P}$. This is in contrast to another formulation of the SCD problem, where it is assumed that we know some partition of $\mathcal{P}$ into "pre-change" and "post-change" classes $\mathcal{P}_0, \mathcal{P}_1$ such that P_0, P_1 are known to respectively lie in $\mathcal{P}_0, \mathcal{P}_1$. In this setting, we are not interested in changes within $\mathcal{P}_0$, only changes from $\mathcal{P}_0$ to $\mathcal{P}_1$. Sometimes, $\mathcal{P}_0$ is even assumed to be a singleton, meaning the pre-change distribution is assumed to be known exactly; we will not make any such assumption. The additional knowledge about the two partitioned distribution classes, that is not available in the first formulation, is often critical in designing optimal SCD schemes, especially in parametric problems. The strategy we develop in this paper is also applicable to an intermediate variant of the SCD problem, that only assumes the knowledge of the pre-change parameter class, Θ_0.

Problem 1.2 (Partitioned SCD)**.** *For some unknown triple (T, P_0, P_1), suppose $X_1, X_2, \ldots$ denote a stream of $\mathcal{X}$-valued observations, satisfying the following assumptions: The observations $(X_n)_{n \leq T}$ are drawn according to a process $P_0 \in \mathcal{P}$ with parameter $\theta_0 \in \Theta_0 \subset \Theta$, where Θ_0 is known. The observations $(X_n)_{n>T}$ are drawn from P_1 with parameter θ_1, such that $\theta_1 \notin \Theta_0$, and θ_1 is unknown, and lies in a set $\Theta_1 \subset \Theta \backslash \Theta_0$. The goal is to design a stopping time τ, satisfying the bulleted criteria stated in Problem 1.1.*

Sequential changepoint detection is a very well-studied problem in the sequential analysis literature, going back to the early works by Shewhart (1925; 1930); Page (1954); Shiryaev (1963). These initial papers developed computationally efficient likelihood-based schemes for known pre- and post-change distributions, which have since then been extended to more general cases of composite, but parametric, pre- and post-change distribution classes. We refer the reader to the book by Tartakovsky et al. (2014) for a detailed discussion of the parametric SCD problem. Unlike the parametric case, there exist very few general principles (analogous to the likelihood and generalized likelihood based schemes) for designing SCD methods with nonparametric distribution classes. Two recent exceptions to this trend include the paper on e-detectors by Shin et al. (2024) for Problem 1.2 (partitioned), and the `BCS-Detector` scheme proposed by Shekhar & Ramdas (2023) for Problem 1.1 (non-partitioned). This paper focuses on the non-partitioned setting, and provides a simpler and theoretically stronger alternative to Shekhar & Ramdas (2023), while also being applicable to the partitioned setting, where it generalizes an old reduction by Lorden (1971).

Remark 1.3. Before proceeding, we clarify our use of the term "reduction" in this paper. Specifically, we use "reduction" from change detection to estimation to mean that if we have a scheme to construct confidence sequences (see Definition 1.4) for a parameter θ, then we can immediately employ it as a subroutine, along with some logically simple operations (such as checking for intersections), to develop a scheme for detecting changes in that parameter.

The `BCS-Detector` scheme. of Shekhar & Ramdas (2023), recalled in Definition A.5 in Appendix A, is also a reduction from changepoint detection to estimation, but in this paper we propose a different, and even simpler reduction. To elaborate, `BCS-Detector` uses a single confidence sequence (CS) in the forward direction, but with every new observation, it constructs a new CS in the backward direction (the so-called "backward CS" or BCS, constructed using observations with their time indices reversed). The scheme stops and declares a detection, as soon as the intersection of the single forward CS and the all active BCSs becomes empty. Since it is critical to our simplified scheme as well, we recall the definition of a CS below.

Definition 1.4 (Confidence Sequence (CS))**.** Given observations $X_1, X_2, \ldots$ drawn from a distribution P with associated parameter/functional $\theta \equiv \theta(P)$, a level-$(1-\alpha)$ CS for θ, is a sequence of sets $(C_n)_{n \geq 1}$, satisfying:

- For every $n \geq 1$, the set $C_n \subset \Theta$ is $\mathcal{F}_n$-measurable. In words, the set C_n is constructed using only the information contained in the first n observations.
- The sets satisfy a uniform coverage guarantee: $\mathbb{P}\left(\forall n \in \mathbb{N} : \theta(P) \in C_n\right) \geq 1 - \alpha$. Equivalently, a CS is a sequence of confidence intervals that is valid at any $\mathcal{F}$-stopping time τ: $\mathbb{P}\left(\theta(P) \in C_\tau\right) \geq 1 - \alpha$.

Remark 1.5. Due to the uniform coverage guarantee, if (C_n) is a CS, then so is $(\cap_{m \leq n} C_m)$. Thus, we can assume without loss of generality that the sets involved in a CS are nested; that is $C_n \subseteq C_{n'}$ for all $n' < n$.

Remark 1.6. Confidence sequences (CSs) are a fundamental tool in sequential and anytime-valid inference, and were first developed by Robbins and co-authors in the 1960s. Some early influential papers on this topic include the works of Darling & Robbins (1967), Lai (1976), Jennison & Turnbull (1984). More recently, there has been a renewed interest in confidence sequences, driven by their applications in multi-armed bandits (Jamieson et al., 2014) and A/B testing (Johari et al., 2015). Some important modern contributions in this area include the works of Howard et al. (2021); Howard & Ramdas (2022), Orabona & Jun (2023), Waudby-Smith & Ramdas (2023), Wang & Ramdas (2023), Chowdhury & Gopalan (2017); Chowdhury et al. (2023), and Kaufmann & Koolen (2021). Our primary objective in this paper is to develop a scheme to harness the significant recent progress on the topic of confidence sequences, in order to develop a general recipe for designing powerful change-detection schemes.

The `BCS-Detector` scheme of Shekhar & Ramdas (2023) satisfies several favorable properties: it can be instantiated

for a large class of parametric and nonparametric problems, it provides non-asymptotic control over the ARL, and has strong guarantees over the detection delay. However, a closer look at their scheme reveals that it implicitly makes a "bidirectional" assumption about the data generating process: at any $n \geq 1$, the `BCS-Detector` assumes the ability to construct a CS in the forward direction (based on $X_1, \ldots, X_n$), as well as in the backward direction (using $Y_1 = X_n, Y_2 = X_{n-1}, \ldots, Y_n = X_1$). Most methods for constructing CSs proceed by designing martingales or supermartingales adapted to the natural (forward) filtration of the observations. Hence, the `BCS-Detector` implicitly involves constructing martingales (or supermartingales) in both, the forward and reverse directions; this in turn is typically only possible if the observations are independent. This restriction limits the applicability of the `BCS-Detector` scheme, a weakness that we eliminate.

Contributions. Our main contributions are as follows:

- In Section 2, we propose a new SCD scheme, that we refer to as the `RCS-Detector`. This scheme proceeds by constructing a new forward CS with each observation, and stops as soon the intersection of all the active CSs (see Remark 2.2) becomes empty.
- Unlike the `BCS-Detector` of Shekhar & Ramdas (2023), our new scheme relies only on forward CSs, thus eliminating their independence requirement on the observations, and allowing for martingale dependence.
- Our scheme also achieves a tighter bound (by a factor 2) on the ARL, as compared to the `BCS-Detector`, while matching its detection delay guarantees.
- Finally, our reduction from change detection to sequential estimation for Problem 1.1 provides a satisfactory analog to the famous reduction by (Lorden, 1971) from Problem 1.2 to sequential testing (Section 3), and also significantly generalizes the latter for Problem 1.2.

2. Our proposed reduction

We now describe our scheme that proceeds by starting a new CS in the forward direction with each new observation.

Definition 2.1 (`RCS-Detector`)**.** Suppose we are given a stream of observations $X_1, X_2, \ldots$, and a functional θ associated with the source. For Problem 1.1 (non-partitioned), define the $C_n^{(0)} = \Theta$ for all $n \geq 1$, while for Problem 1.2 (partitioned) set $C_n^{(0)} = \Theta_0$, for all $n \geq 1$. Proceed as follows for $n = 1, 2, \ldots$:

1. Observe the next data-point X_n.
2. Using X_n, update all the previous CSs (that is, $\{C^{(m)} : 0 \leq m < n\}$) and also initialize a new level-$(1-\alpha)$ CS, denoted by $C^{(n)}$.
3. If the intersection of all initialized CSs becomes empty, meaning $\cap_{m=0}^{n} C_n^{(m)} = \emptyset$, then set $\tau \leftarrow n$, and declare a detection.

In the last step, we have implicitly used the nestedness discussed Remark 1.5, but if the CSs are not nested, we can use the stopping criterion $\cap_{m=0}^{n} \cap_{i=m}^{n} C_i^{(m)} = \emptyset$.

Remark 2.2. Note that at the end of round n, the `RCS-Detector` scheme has $n+1$ "active CSs": $C^{(0)}, C^{(1)}, \ldots, C^{(n)}$. At this time, each CS $C^{(m)}$ consists of the sets $\{C_m^{(m)}, \ldots, C_n^{(m)}\}$ constructed using the observations $X_m, \ldots, X_n$. Specifically, the most recent CS, denoted by $C^{(n)}$, consists of a single set $\{C_n^{(n)}\}$, constructed using only the observation X_n.

Remark 2.3. In general, the computation cost of our reduction (and the `BCS-Detector`) increases quadratically with n since at time n we need to update all CSs initialized so far using X_n, meaning that we form the sets $\{C_n^{(m)} : 1 \leq m \leq n\}$. One possible way of reducing this computational cost is by considering only the w most recent CSs, for some window-size $w > 0$. Selecting the appropriate value of w, either based on some prior information about the change magnitude, or by learning it in a data-driven manner is an interesting direction for future work.

Compared to the `BCS-Detector` of Shekhar & Ramdas (2023), the main change in the above scheme is that it creates a new forward CS with each new observation, instead of a new "backward CS" (a new concept defined by their paper, but this complexity is unnecessary with ours).

2.1. Analyzing the average run length

We now show that our `RCS-Detector` scheme admits a nonasymptotic lower bound on the average run length (ARL) when there is no change.

Theorem 2.4 (ARL control)**.** *The changepoint detection scheme described in Definition 2.1 controls the average run length (ARL) at level* $1/\alpha$*. That is, when* $T = \infty$*, our proposed stopping time* τ *satisfies* $\mathbb{E}_\infty[\tau] \geq 1/\alpha$*.*

The proof of this result is in Section 4.1. Note that for the `BCS-Detector` obtained a lower bound of $1/2\alpha - 3/2$ on the ARL. Thus, our `RCS-Detector` achieves an improved (approximately by a factor of 2) lower bound under weaker model assumptions on the data stream, while matching the detection delay guarantees of `BCS-Detector`, as we show in the next section. This improved performance guarantee is a consequence of more refined analysis (that we are unable to extend to `BCS-Detector`), and in practice, we observe

that the empirical performance of `RCS-Detector` and `BCS-Detector` are comparable to each other on independent data streams.

Remark 2.5. An alternative performance measure to ARL is the *probability of false alarms (PFA)*, which is equal to the probability that the stopping time τ is finite; that is $\mathbb{P}_\infty(\tau < \infty)$. If we modify our `RCS-Detector` to use a level-$(1 - 6\alpha/(n^2\pi^2))$ CS in each round, then the resulting scheme ensures

$$\mathbb{P}_\infty(\tau < \infty) \leq \sum_{n \geq 1} \mathbb{P}_\infty\left(\left\{(C_t^{(n)})_{t \geq n} \text{ miscovers } \theta_0\right\}\right) \leq \alpha \sum_{n \geq 1} \frac{6}{\pi^2 n^2} = \alpha.$$

This implies that the ARL of the above modified `RCS-Detector` scheme is infinity, since $\mathbb{E}_\infty[\tau] \geq (1 - \alpha) \times \infty = \infty$. This significantly stronger control over false alarms comes at the cost of an increase in detection delay. In particular, we can show that for most CSs, the detection delay of this modified scheme will have a logarithmic dependence on T. This means that the worst case (over all T values) detection delay of the PFA-controlling scheme is usually unbounded.

2.2. Analyzing the detection delay

We now state an assumption under which we will analyze the detection delay of our SCD scheme.

Assumption 2.6. *Letting d denote a metric on Θ, X^n be shorthand for $(X_1, \ldots, X_n)$, and $(C_n)_{n \geq 0}$ be a given confidence sequence, we assume that the width of the set $C_n \equiv C_n(X^n, \alpha)$ has a deterministic bound*

$$\sup_{\theta', \theta'' \in C_n} d(\theta', \theta'') \overset{a.s.}{\leq} w(n, P, \alpha),$$

with $\lim_{n \to \infty} w(n, P, \alpha) = 0$, for all $P \in \mathcal{P}$, $\alpha \in (0, 1]$.

The above assumption requires the existence of a deterministic envelope function for the diameter of the confidence sequence, which converges to zero pointwise for every (P, α), as n increases. This is a very weak assumption, and essentially all known CSs satisfy it.

We now analyze the detection delay of our SCD scheme for Problem 1.1 under Assumption 2.6.

Theorem 2.7. *Consider the SCD problem with observations $X_1, X_2, \ldots$ such that $(X_n)_{n \leq T}$ are drawn from a distribution P_0 (with parameter θ_0), while $(X_n)_{n > T}$ are drawn from a product distribution P_1 (with parameter $\theta_1 \neq \theta_0$), and are independent of the pre-change observations. Suppose the `RCS-Detector` from Definition 2.1 is applied to this problem, with the CSs satisfying Assumption 2.6. Let $\mathcal{E}_T := \{\theta_0 \in \cap_{n=1}^T C_n^{(1)}\}$ denote the "good event" (having at least $(1 - \alpha)$ probability) that the first CS covers the true parameter up to the changepoint. For Problem 1.1 (non-partitioned), if T is large enough to ensure that $w(T, P_0, \alpha) < d(\theta_0, \theta_1)$, then the detection delay of our proposed scheme satisfies*

$$\mathbb{E}_T[(\tau - T)^+ | \mathcal{E}_T] \leq \frac{3}{1 - \alpha} u(\theta_0, \theta_1, T),$$

where $u(\theta_0, \theta_1, T) := \min\{n \geq 1 : w(T, P_0, \alpha) + w(n, P_1, \alpha) < d(\theta_0, \theta_1)\}$. For Problem 1.2 (partitioned), the `RCS-Detector` satisfies

$$\mathbb{E}_T[(\tau - T)^+ | \mathcal{F}_T] \leq \frac{3}{1 - \alpha} u(\Theta_0, \theta_1), \tag{1}$$

where $u(\Theta_0, \theta_1) := \min\{n \geq 1 : w(n, P_1, \alpha) < \inf_{\theta' \in \Theta_0} d(\theta', \theta_1)\}$ for all values of $T < \infty$.

The proof of this result adapts the arguments developed by Shekhar & Ramdas (2023) for analyzing the `BCS-Detector`, and we present the details in Section 4.2.

Remark 2.8. The above detection delay bound *exactly* matches that obtained by the `BCS-Detector` of Shekhar & Ramdas (2023), which (as mentioned earlier) had a worse ARL guarantee of ARL $\geq 1/(2\alpha) - 3/2$. Recalling Theorem 2.4, our new scheme achieves an improved bound on the ARL, while matching its detection delay.

The previous result provides an explicit detection delay bound applicable to a large class of problems in terms of the CS width $w(n, P, \alpha)$. We now state a less explicit bound on the detection delay that is valid under much weaker conditions.

Proposition 2.9. *Consider an SCD problem in which the post change observations are (i) stationary, and (ii) independent of $\mathcal{F}_T = \sigma(X_1, \ldots, X_T)$. Then, the detection delay of the `RCS-Detector` on Problem 1.2 (partitioned) satisfies the following bound:*

$$\mathbb{E}_T[(\tau - T)^+ | \mathcal{F}_T] \leq \mathbb{E}_0[N_1], \quad \text{where}$$
$$N_m := \inf\{n - m : C_n^{(m)} \cap \theta_0 = \emptyset\}, \text{ for } m \geq 1.$$

An exactly analogous bound holds for $\mathbb{E}_T[(\tau - T)^+ | \mathcal{E}_T]$ for Problem 1.1 (non-partitioned), with the modification that Θ_0 is replaced with $\{\theta : d(\theta_0, \theta) \leq d(\theta_0, \theta_1)/2\}$ in the definition of N_m.

This result can be used as an intermediate step in obtaining sharp detection delay bounds for the `RCS-Detector` on problems with some additional structure. We demonstrate this next in Section 2.3, for the problem of detecting changes in mean of bounded data.

2.3. A nonparametric example: change in mean for bounded random variables

We now analyze the performance of our changepoint detection scheme the problem of detecting changes in the

mean of bounded real-valued random variables supported on $\mathcal{X} = [0,1]$. Note that despite the simple observation space, the class of distributions on this $\mathcal{X}$ is highly composite and nonparametric. In particular, there does not exist a common dominating measure for all distributions in this class, which renders likelihood based techniques inapplicable to this problem.

Formally, we consider the instances of Problem 1.1 and Problem 1.2 with $\mathcal{X} = [0,1]$, and the parameter space $\Theta = [0,1]$ with metric $d(\theta, \theta') = |\theta - \theta'|$. For Problem 1.2, we assume that the pre-change mean θ_0 lies in a known set $\Theta_0 \subset \Theta$. For an unknown value $T \in \mathbb{N} \cup \{\infty\}$, the distribution generating the observations changes from P_0, with mean θ_0, to another distribution P_1, with mean $\theta_1 \neq \theta_0$. The (unknown) change magnitude is denoted by $d(\theta_0, \theta_1) = \Delta = |\theta_1 - \theta_0|$.

For this problem, we employ our `RCS-Detector` strategy using an instance of the betting-based construction of CSs for the means of bounded random variables (details in Appendix B) proposed by Waudby-Smith & Ramdas (2023). Our next result analyzes its performance.

Proposition 2.10. *Consider the problem of detecting changes in mean with bounded observations, under these additional conditions: (i) the post-change observations are independent of the pre-change observations, and (ii) both, the pre- and post-change observations are i.i.d. (that is P_0, P_1 are infinite products of some distributions p_0, p_1 on $\mathcal{X}$). For Problem 1.1 (non-partitioned), if $T \geq 64\log(64/\Delta^2\alpha)/\Delta^2$, the `RCS-Detector` instantiated with the betting CS (details in Appendix B) satisfies:*

$$\mathbb{E}_\infty[\tau] \geq \frac{1}{\alpha}, \quad \textit{and} \quad \mathbb{E}_T[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\left(\frac{\log(1/\alpha K_1)}{K_1}\right), \tag{2}$$

where $K_1 = K_1(P_1, \theta_0) \coloneqq \inf_{P_\theta : |\theta - \theta_0| \leq \Delta/2} d_{KL}(p_1 \parallel p_\theta)$. In the display above, $\mathcal{E}$ is the "good event" in Theorem 2.7, having probability at least $1 - \alpha$.

For Problem 1.2 (partitioned), the `RCS-Detector` satisfies the following:

$$\mathbb{E}_\infty[\tau] \geq \frac{1}{\alpha}, \quad \textit{and} \quad \mathbb{E}_T[(\tau - T)^+] = \mathcal{O}\left(\frac{\log(1/\alpha K_2)}{K_2}\right), \tag{3}$$

where $K_2 \equiv K_2(P_1, \Theta_0) = \inf_{P_\theta : \theta \in \Theta_0} d_{KL}(p_1 \parallel p_\theta)$. In the statements above, $P_\theta = p_\theta^\infty$ denotes any product distribution on $\mathcal{X}^\infty$ with mean θ.

The proof of this result is in Appendix B, and relies on a careful analysis of the behavior of betting CSs. If the pre-change distribution P_0 is also i.i.d. (say $P_0 = p_0^\infty$), and is known, then K_2 in (3) reduces to $d_{\mathrm{KL}}(p_1 \parallel p_0)$. The resulting detection delay is order optimal, according to Lorden (1971)[Theorem 3], and furthermore, this optimality is achieved for an unknown P_1 lying in a nonparametric distribution class.

Remark 2.11. By an application of Pinsker's inequality (Fact A.3 in Appendix A), we know that both K_1 and K_2 are $\Omega(\Delta^2)$, which gives us the weaker upper bound on the detection delay, $\mathcal{O}\left(\log(1/\alpha\Delta)/\Delta^2\right)$. This is the upper bound on the detection delay derived by Shekhar & Ramdas (2023) for the change of mean detection problem, using the empirical-Bernstein CS of Waudby-Smith & Ramdas (2023), and a direct application of the general delay bound of Theorem 2.7.

2.4. Numerical experiments

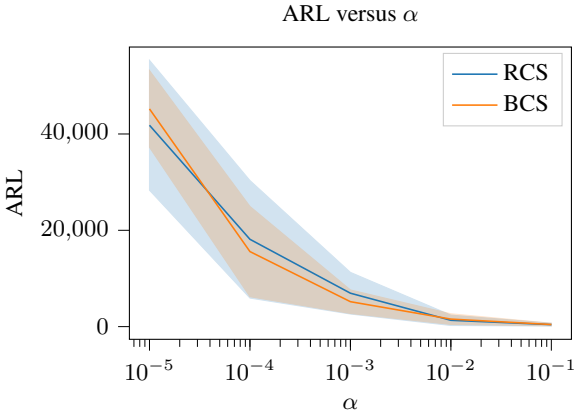

Figure 1: The figure plots the estimates of (lower bounds on) the ARL of the two schemes constructed based on 50 independent trials.

In this section, we empirically compare the performance of the our `RCS-Detector` with the `BCS-Detector` scheme of Shekhar & Ramdas (2023) on the problem of detecting changes in mean of bounded observations. For simplicity, we instantiate both these schemes using Hoeffding CS (details below) proposed by Waudby-Smith & Ramdas (2023), since it admits a closed form representation. Our main objective in this section is to illustrate how the `RCS-Detector` (and `BCS-Detector`) can be used to address non-partitioned change detection problems with minimal knowledge and assumptions on the distributions. We leave a more thorough empirical evaluation of our general scheme and its instantiations in different scenarios, as an interesting direction for future work.

Experiment setup. We consider the problem of detecting changes in the mean of bounded observations, where

- $X_1, X_2, \ldots,$ are independent observations, supported on $\mathcal{X} = [0,1]$.
- For $t \leq T$, each X_t is drawn according to a Beta

distribution with parameters $(2, 2(1-\mu_0)/\mu_0)$, where μ_0 denotes the pre-change mean.

- For $t > T$, each X_t is drawn from a Beta distribution with parameters $(2, 2(1-\mu_1)/\mu_1)$, where $\mu_1 \neq \mu_0$ is the post-change mean value.
- We denote the change magnitude by $\Delta = |\mu_1 - \mu_0|$.

We instantiate both the schemes using the following Hoeffding-type confidence sequence constructed by Waudby-Smith & Ramdas (2023):

$$C_t = \left[\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t \lambda_i} \right],$$

$$\text{where } \lambda_i = \sqrt{\frac{8\log(2/\alpha)}{i \times \log(i+1)}} \wedge 1, \text{ for all } i \geq 1.$$

ARL comparison. In the first experiment, we compare the average run length (ARL) of the two schemes. For this, we set $\Delta = 0$, and consider five values of α in the set $\{10^{-i} : 1 \leq i \leq 5\}$, and estimate the ARL as the average of the stopping times over 50 trials (capped at 50000). The results of this experiment, plotted in Figure 1, indicate that the both schemes provide the required control over ARL, but the `RCS-Detector` has a slightly higher variability, especially at smaller α values. See Appendix C for some additional details of this experiment.

Detection delay comparison. We now compare the detection delay of the two schemes. To do this, we set $\alpha = 0.001$, and consider six value of $\Delta \in \{0.05, 0.075, 0.10, 0.125, 0.15, 0.175\}$. For every Δ value, we ran 50 independent trials with the two schemes to estimate their average detection delay (clipped at a maximum of 24000). As shown in Figure 2, the detection delay of the two schemes are roughly comparable for independent data. However, the `RCS-Detector` has the additional benefit of being applicable to a more general class of problems with dependent data streams; we present one such example in Appendix C.

3. Connection to Lorden's reduction from SCD to testing

Using the duality between confidence sequences and sequential hypothesis tests, we now show that our `RCS-Detector` strategy is a generalization of a well-known result of Lorden (1971), that reduces the problem of SCD (with separated distribution classes) to that of repeated sequential tests.

Lorden's work built upon the interpretation of CuSum algorithm as repeated sequential probability ratio tests (SPRT) for known pre- and post-change distributions by Page (1954). In particular, Lorden (1971) considered a parametric SCD problem with a known pre-change distribution P_0, and a parametric composite class of post-change distributions $\{P_{\theta_1} : \theta_1 \in \Theta_1\}$. Then, given a sequential test, or equivalently, extended stopping time, $\{N(\alpha, \theta_0) : \alpha \in (0,1)\}$, satisfying $\mathbb{P}_{P_0}(N(\alpha, \theta_0) < \infty) \leq \alpha$, Lorden (1971) proposed the following SCD strategy:

- For every $m \geq 1$, define $N^{(m)}(\alpha, \theta_0)$ as the stopping rule $N(\alpha, \theta_0)$ applied to the observations $X_m, X_{m+1}, \ldots$.
- Using these, declare the changepoint at the time $\tau_L \equiv \tau_L(\alpha)$, defined as

$$\tau_L = \inf_{m \geq 1} \{N^{(m)}(\alpha, \theta_0) + m\}.$$

In words, this scheme can be summarized as: *initiate a new sequential level-α test with every new observation, and stop and declare a detection as soon as one of the active tests rejects the null*. For this scheme, Lorden (1971) established the ARL control; that is, $\mathbb{E}_{P_0}[\tau_L] \geq 1/\alpha$, for the specified $\alpha \in (0,1)$. Furthermore, under certain assumptions on the expected stopping time of the test $N(\alpha, \theta_0)$ under the alternative, Lorden (1971) also established the minimax optimality of the scheme in the regime of $\alpha \to 0$.

Our main result of this section establishes a connection between Lorden's reduction and `RCS-Detector`.

Proposition 3.1. *Consider an SCD problem with pre-change parameter set $\Theta_0 = \{\theta_0\}$, and a post-change parameter set Θ_1. Then, we have the following:*

- *For every Lorden-type scheme τ_L, there exists an `RCS-Detector` τ_R, such that $\tau_L = \tau_R$.*
- *For every `RCS-Detector` τ_R, there exists a Lorden-type scheme τ_L, such that $\tau_R = \tau_L$.*

Thus, there is a one-to-one correspondence between Lorden's reduction and `RCS-Detector` for such problems.

The proof of this statement is in Section 4.4, and it relies on the duality between CSs and power-one sequential tests. We end our discussion with two remarks.

Remark 3.2. While we focused on the case of a singleton null, $\{P_0\}$, a similar result holds for the case of a composite null $\{P_{\theta_0} : \theta_0 \in \Theta_0\}$, with $\Theta_0 \cap \Theta_1 = \emptyset$. The only modification needed is to update the stopping time $N(\alpha, \Theta_0)$ to be equal to $\inf\{n \geq 1 : \Theta_0 \cap C_n = \emptyset\}$. By Theorem 2.4, the resulting SCD scheme still controls the ARL at the required level $1/\alpha$.

Remark 3.3. Note that the e-detector framework, developed by Shin et al. (2024), also generalizes and strictly

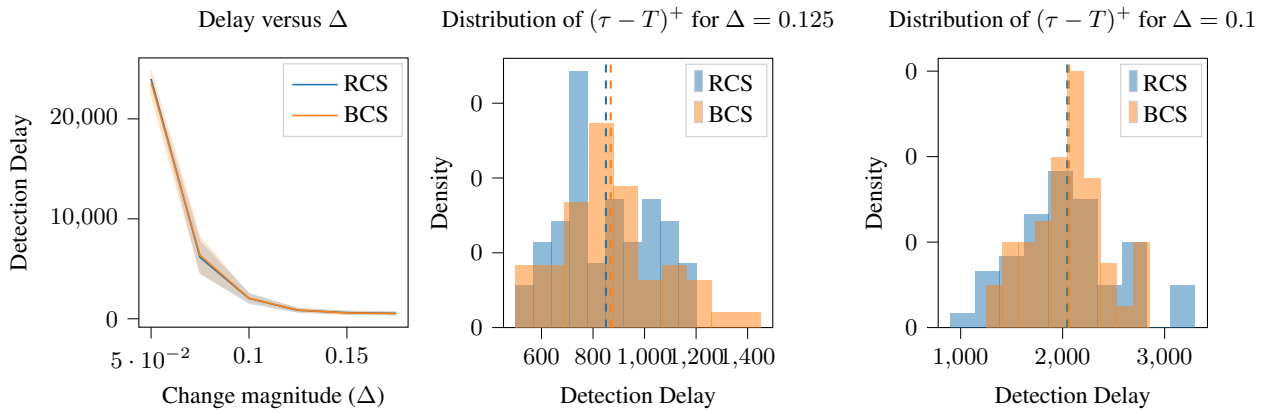

Figure 2: The first plot shows the variation of the detection delay (averaged over 50 trials) for the two schemes (both constructed using Hoeffding CS) with the change magnitude (Δ). As suggested by Theorem 2.7, the detection delay of the two schemes are comparable. In some problem instances (smaller Δ values), the performance of `RCS-Detector` is slightly better (middle plot), while in some others (larger Δ values), the `BCS-Detector` does better (right plot). The maximum detection delay in the experiments was capped at 24000, which is why the delay of the two schemes agree at $\Delta = 0.05$ (indicating that the true detection delay of both schemes is some quantity strictly larger than 24000).

improves upon Lorden's scheme to work for composite, and nonparametric pre- and post-change distribution classes ($\mathcal{P}_0$ and $\mathcal{P}_1$ respectively). However, the e-detectors were developed explicitly for Problem 1.2 (partitioned); that is for a known class of pre-change distributions $\mathcal{P}_0$ (although the general idea could be suitably adapted for the non-separated formulation in some cases). This is unlike our scheme that is applicable to both the partitioned and non-partitioned formulations of the SCD problem.

4. Deferred proofs

We now present the proofs of the two main technical results characterizing the performance of our `RCS-Detector` scheme, stated in Section 2 and Section 3.

4.1. Proof of Theorem 2.4

We prove this statement in three steps. First, we define an e-process (recalled in Definition 4.1) corresponding to every confidence sequence $(C_n^{(m)})_{n \geq m}$ involved in our scheme. Then, using these e-processes we introduce an e-detector $(M_n)_{n \geq 1}$, that is, a process adapted to the natural filtration $\mathcal{F}$ that satisfies $\mathbb{E}_\infty[M_{\tau'}] \leq \mathbb{E}_\infty[\tau']$ for all stopping times τ'. Finally, we show that our stopping time τ, introduced in Definition 2.1, is larger than $\tau' = \inf\{n \geq 1 : M_n \geq 1/\alpha\}$, defined using the e-detector. This allows us to leverage Shin et al. (2024, Proposition 2.4) to conclude that $\mathbb{E}_\infty[\tau'] \geq 1/\alpha$, which implies required statement about the ARL of τ.

Since we prove this result by attaching an e-process to every CS, we recall their definition below.

Definition 4.1 (e-processes)**.** Given a class of probability measures $\mathcal{P}$, and a filtration $\mathcal{F} \equiv (\mathcal{F}_n)_{n \geq 1}$ defined on some measurable space, an e-process for $\mathcal{P}$ is a collection of nonnegative random variables $(E_n)_{n \geq 1}$ adapted to $\mathcal{F}$, satisfying $\mathbb{E}_P[E_{\tau'}] \leq 1$ for all $P \in \mathcal{P}$, and for all stopping times τ' (adapted to the same filtration).

Step 1. Construct an e-process for every CS. For every CS starting with the m^{th} observation, denoted by $(C_n^{(m)})_{n \geq m}$, we associate a process defined as

$$E_n^{(m)} = \begin{cases} 0, & \text{if } n < m, \text{ OR if } n \geq m, \text{ and } \theta_0 \in C_n^{(m)}, \\ \frac{1}{\alpha}, & \text{if } n \geq m, \text{ and } \theta_0 \notin C_n^{(m)}. \end{cases}$$

It is easy to verify that for every $m \geq 1$, the process $\{E_n^{(m)} : n \geq 1\}$ is an e-process:

- For every $n \geq 1$, the value of $E_n^{(m)}$ is $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ measurable.
- For any stopping time τ', adapted to the filtration $\mathcal{F}$, we have

$$\begin{aligned} \mathbb{E}_\infty[E_{\tau'}^{(m)}] &= \mathbb{E}_\infty\left[0 \times \mathbf{1}_{\tau' < m} + \frac{1}{\alpha} \times \mathbf{1}_{\tau' \geq m} \mathbf{1}_{\theta_0 \notin C_{\tau'}^{(m)}}\right] \\ &= \frac{1}{\alpha} \times \mathbb{E}_\infty\left[\mathbf{1}_{\tau' \geq m} \mathbf{1}_{\theta_0 \notin C_{\tau'}^{(m)}}\right] \\ &\leq \frac{1}{\alpha} \times \mathbb{E}_\infty\left[\mathbf{1}_{\theta_0 \notin C_{\tau'}^{(m)}}\right] \\ &= \frac{1}{\alpha} \times \mathbb{P}_\infty\left(\theta_0 \notin C_{\tau'}^{(m)}\right) \leq \frac{1}{\alpha} \times \alpha = 1. \end{aligned}$$

The last inequality uses the fact that $(C_n^{(m)})_{m\geq n}$ is a level-$(1-\alpha)$ CS for θ_0. Thus, for every $m \geq 1$, the process $(E_n^{(m)})_{n\geq 1}$ is a valid e-process.

Step 2. Construct an e-detector. For every $n \geq 1$, we define M_n to be equal to $\sum_{m=1}^n E_n^{(m)}$, and observe that the process $(M_n)_{n\geq 1}$ is an *e-detector*, as defined by Shin et al. (2024, Definition 2.2) because it satisfies the following two properties:

- $(M_n)_{n\geq 1}$ is adapted to $(\mathcal{F}_n)_{n\geq 1}$: since for any $n \geq 1$, all the $E_n^{(m)}$ are $\mathcal{F}_n$-measurable by construction.
- For any stopping time τ', we have $\mathbb{E}_\infty[M_{\tau'}] \leq \mathbb{E}_\infty[\tau']$, as noted by Shin et al. (2024, Definition 2.6).

Step 3. Bound the ARL using the e-detector. Finally, we translate the stopping criterion of our proposed scheme (stated as the non-intersection of the confidence sequences) in terms of the e-detector $(M_n)_{n\geq 1}$. In particular, we have

$$\begin{aligned}\{\tau \leq n\} &= \{\cap_{m=0}^n C_n^{(m)} = \emptyset\} \subset \{\exists m \in [n] : \theta_0 \notin C_n^{(m)}\} \\ &= \{\exists m \in [n] : E_n^{(m)} = 1/\alpha\} \\ &= \{M_n \geq 1/\alpha\}. \end{aligned} \tag{4}$$

In words, when $T = \infty$, if the intersection of the CSs is empty prior to some time n, it means that at least of of the CSs constructed prior to n must miscover. This in turn implies that the value of at least one of the e-processes at n is equal to $1/\alpha$; or the value of the e-detector M_n is at least $1/\alpha$. Recall, that in the first equality above, we have assumed that the sets in the confidence sequences are nested; that is, $C_n^{(m)} \subset C_{n'}^{(m)}$ for every $m \leq n' < n$. This allows us to look only at the intersection of the most recent sets to define the stopping condition. We now define a new stopping time $\tau' = \inf\{n \geq 1 : M_n \geq 1/\alpha\}$, and observe that it is stochastically dominated by τ; that is, (4) implies

$$\{\tau' > n\} = \{M_n < 1/\alpha\} \subset \{\tau > n\},$$

which allows us to conclude that $\mathbb{E}_\infty[\tau'] \leq \mathbb{E}_\infty[\tau]$. From Shin et al. (2024, Proposition 2.4), we know that $\mathbb{E}_\infty[\tau'] \geq 1/\alpha$, and we conclude the result by noting that $\mathbb{E}_\infty[\tau] \geq \mathbb{E}_\infty[\tau']$ since τ stochastically dominates τ'. □

4.2. Proof of Theorem 2.7

The proof of this result follows the general argument developed by Shekhar & Ramdas (2023) for analyzing their `BCS-Detector` strategy, with some modifications due to the use of forward CSs (instead of backward CSs used in the `BCS-Detector`).

In particular, we consider blocks of the post-change observations, each of length $u \equiv u(\theta_0, \theta_1, T)$, starting at time $T_j = T + ju$ for $j \geq 0$. Note that all these blocks are independent of each other (since P_1 is a product distribution), and also independent of the event $\mathcal{E} = \{\theta_0 \in C_T^{(1)}\}$. Now, observe that for $k \geq 1$, we have $\{\tau > T_k\} = \cap_{j=1}^k \{\tau > T_j\}$, which furthermore implies

$$\begin{aligned}\{\tau > T_k\} \cap \mathcal{E} &\subset \cap_{j=1}^k \{C_{T_j}^{(T_{j-1})} \cap C_T^{(1)} \neq \emptyset\} \cap \mathcal{E} \\ &\subset \cap_{j=1}^k \{C_{T_j}^{(T_{j-1})} \text{ miscovers } \theta_1\} \cap \mathcal{E}. \end{aligned} \tag{5}$$

The last inclusion follows from the definition of $u \equiv u(\theta_0, \theta_1, T)$, and the event $\mathcal{E}$. Introducing $D_k = \sum_{t=T_k+1}^{T_{k+1}} \mathbb{P}_T(\tau \geq t|\mathcal{E})$, and using (5), we obtain:

$$\begin{aligned} D_k &\leq \sum_{t=T_k+1}^{T_{k+1}} \mathbb{P}_T(\tau \geq T_k|\mathcal{E}) = u\mathbb{P}_T(\tau > T_k|\mathcal{E}) \\ &\leq u \times \mathbb{P}_T\left(\cap_{j=1}^k \{C_{T_j}^{(T_{j-1})} \text{ miscovers } \theta_1\} \cap \mathcal{E}|\mathcal{E}\right) \\ &\overset{(i)}{\leq} \frac{u}{\mathbb{P}_T(\mathcal{E})} \times \prod_{j=1}^k \mathbb{P}_T\left(\{C_{T_j}^{(T_{j-1})} \text{ miscovers } \theta_1\} \cap \mathcal{E}\right) \\ &\leq \frac{u\alpha^k}{1-\alpha}. \end{aligned}$$

The inequality (i) uses the fact that $\mathcal{E}$ only depends on the pre-change observations, and hence is independent of the post-change CSs. For any $k_0 > 1$, observe that

$$\mathbb{E}_T[(\tau - T)^+] \leq k_0 u + \sum_{k=k_0}^\infty D_k = u\left(k_0 + \frac{\alpha^{k_0}}{(1-\alpha)^2}\right).$$

By setting $k_0 = \lceil 2\log(1/1-\alpha)/\log(1/\alpha)\rceil$, we get the required statement for Problem 1.1.

To prove the second part of Theorem 2.7, we proceed as above, considering blocks of post-change observations of length $u \equiv u(\Theta_0, \theta_1)$ as defined in (1). We then define $T_j = T + ju$ for $j \geq 0$, and note that

$$\begin{aligned}\{\tau > T_k\} &\subset \cap_{j=1}^k \{C_{T_j}^{(T_{j-1})} \cap \Theta_0 \neq \emptyset\} \\ &\subset \cap_{j=1}^k \{C_{T_j}^{(T_{j-1})} \text{ miscovers } \theta_1\}. \end{aligned}$$

The rest of the argument, then proceeds exactly as before.

4.3. Proof of Proposition 2.9

The first step is to show that when $T < \infty$, we have

$$(\tau - T)^+ \leq N_{T+1}, \quad \text{almost surely.}$$

This inequality is true trivially on the event $\{\tau \leq T\}$, since each N_m is non-negative. Now, observe that

$$\begin{aligned}(\tau - T)\mathbf{1}_{\tau > T} &= \inf\{n - T : \cap_{m\leq n} C_n^{(m)} = \emptyset,\ n > T\} \\ &\leq \inf\{n - T : C_n^{(0)} \cap C_n^{(T+1)} = \emptyset,\ n > T\} \\ &= \inf\{n - T : \Theta_0 \cap C_n^{(T+1)} = \emptyset,\ n > T\} \\ &= N_{T+1} \end{aligned}$$

The first inequality uses the fact that we are taking fewer intersections (hence will take longer to stop), while the second equality uses the fact that $C_n^{(0)} = \Theta_0$ for Problem 1.2. To conclude the proof, we observe that

$$\mathbb{E}_T[(\tau - T)^+ | \mathcal{F}_T] \leq \mathbb{E}_T[N_{T+1} | \mathcal{F}_T] \overset{(i)}{=} \mathbb{E}_T[N_{T+1}] \overset{(ii)}{=} \mathbb{E}_0[N_1].$$

The equality (i) follows from the independence of post-change data to $\mathcal{F}_T$, and (ii) follows from the stationarity of the post-change data.

4.4. Proof of Proposition Proposition 3.1

We prove this statement in three steps: in the first two steps, we show show how to construct a τ_R from τ_L, and a τ_L from τ_R respectively; and then establish their equivalence in the third step.

Step 1. Consider a Lorden-type stopping time $\tau_L = \inf\{m + N^{(m)}(\alpha, \theta_0) : m \geq 1\}$, and use its underlying test $N(\alpha, \cdot)$ to construct CSs as follows:

$$C_n^{(m)} = \{\theta \in \Theta_0 \cup \Theta_1 : N^{(m)}(\alpha, \theta) > n\}.$$

Note that $\{N^{(m)}(\alpha, \theta) > n\} = \{N^{(m)}(\alpha, \theta) \leq n\}^c$ is an $\mathcal{F}_n$ measurable event as required. These CSs can now be used to define an `RCS-Detector` $\tau_R = \inf\{n \geq 1 : \cap_{m=0}^n C_n^{(m)} = \emptyset\}$.

Step 2. Consider an `RCS-Detector` τ_R, and let $\mathcal{C}$ denote the method for constructing confidence sequences used by τ_L. Then, we can define a sequential test for $\{P_0\}$ as

$$N(\alpha, \theta_0) = \inf\{n \geq 1 : \theta_0 \notin C_n\},$$

where $C_n = \mathcal{C}(X_1, \ldots, X_n; \alpha)$. By the uniform coverage guarantee of confidence sequences, we have $\mathbb{P}_{P_0}(N(\alpha, \theta_0) < \infty) = \mathbb{P}_{P_0}(\exists n \in \mathbb{N} : \theta_0 \notin C_n) \leq \alpha$. Thus, $N(\alpha, \theta_0)$ is a valid level-α sequential test for the simple null $\{P_0\}$. Similarly, $N^{(m)}$ for $m \geq 1$, can be defined as the stopping time $N(\alpha, \theta_0)$ constructed using observations $(X_n)_{n \geq m}$ starting at time m. More specifically, we have

$$N^{(m)}(\alpha) = \inf\{n - m : \theta_0 \notin C_n^{(m)},\ n \geq m\},$$
$$\text{where} \quad C_n^{(m)} = \mathcal{C}(X_m, X_{m+1} \ldots, X_n; \alpha).$$

Using this, we can define a Lorden-type change detector $\tau_L = \inf_{m \geq 1}\{N^{(m)}(\alpha, \theta_0) + m\}$.

Step 3. We conclude the proof by noting that in both the cases above, we have $\tau_L = \tau_R$. In particular, observe the following chain of equalities:

$$\begin{aligned}
\{\tau_L \leq n\} &= \{\exists n' \leq n, \exists m \leq n' : N^{(m)}(\alpha) = n' - m\} \\
&= \{\exists n' \leq n, \exists m \leq n' : \theta_0 \notin C_{n'}^{(m)}\} \\
&= \{\exists n' \leq n : \left(\cap_{m=1}^{n'} C_{n'}^{(m)}\right) \cap \{\theta_0\} = \emptyset\} \\
&= \{\cap_{n'=1}^{n} \cap_{m=0}^{n'} C_{n'}^{(m)} = \emptyset\} \\
&= \{\cap_{m=0}^{n} \cap_{n'=m}^{n} C_{n'}^{(m)} = \emptyset\} \\
&= \{\tau_R \leq n\}.
\end{aligned}$$

In the last two equalities, we have used the fact that $C_n^{(0)}$ for all $n \geq 0$ is equal to $\Theta_0 = \{\theta_0\}$.

5. Conclusion

In this paper, we proposed a new and simple reduction from sequential changepoint detection to sequential estimation: our scheme that constructs a new CS with every observation, and declares a detection as soon as the intersection of the active CSs becomes empty. The design of our scheme improves on the `BCS-Detector` of Shekhar & Ramdas (2023), which proceeds by initializing new "backward CSs" with each new observation. Indeed, we showed that our new scheme matches the detection delay performance of `BCS-Detector`, while improving the ARL lower bound by a factor of 2. Furthermore, our scheme achieves this improvement under weaker dependence assumptions (i.e., without needing the ability to construct CSs in both forward and backward directions). Interestingly, our proposed scheme can be seen as a nonparametric generalization of Lorden's reduction from SCD to repeated sequential testing, due to the duality between sequential testing and CSs.

As a consequence of our proposed reduction, the large and rapidly growing literature on CSs can now immediately be brought to bear on change detection problems. While our method does involve per-step computation that grows linearly with sample size, it at least provides an immediate statistically optimal baseline method for new (even nonparametric) problems, and we argue that this versatility will result in its broad use, even if it is superseded by other computationally efficient methods for specific problems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Chowdhury, S. R., Saux, P., Maillard, O., and Gopalan, A. Bregman deviations of generic exponential families. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 394–449. PMLR, 2023.

Darling, D. A. and Robbins, H. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

Durrett, R. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2019.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pp. 13–30, 1963.

Honda, J. and Takemura, A. An asymptotically optimal bandit algorithm for bounded support models. In *The Twenty Third Annual Conference on Learning Theory*, pp. 67–79. PMLR, 2010.

Howard, S. R. and Ramdas, A. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439. PMLR, 2014.

Jennison, C. and Turnbull, B. W. Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, 5(1):33–45, 1984.

Johari, R., Pekelis, L., and Walsh, D. J. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*, 2015.

Kaufmann, E. and Koolen, W. M. Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research*, 22 (1):11140–11183, 2021.

Lai, T. L. On confidence sequences. *The Annals of Statistics*, pp. 265–280, 1976.

Lorden, G. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.

Orabona, F. and Jun, K.-S. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 2023.

Page, E. S. Continuous inspection schemes. *Biometrika*, 41 (1/2):100–115, 1954.

Shekhar, S. and Ramdas, A. Sequential changepoint detection via backward confidence sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30908–30930. PMLR, 23–29 Jul 2023.

Shewhart, W. A. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152):546–548, 1925.

Shewhart, W. A. Economic quality control of manufactured product. *Bell System Technical Journal*, 9(2):364–389, 1930.

Shin, J., Ramdas, A., and Rinaldo, A. E-detectors: a nonparametric framework for online changepoint detection. *New England Journal of Statistics and Data Science*, 2024.

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

Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics, 2009.

Wang, H. and Ramdas, A. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and their Applications*, 163:168–202, 2023.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society B (with discussion)*, 2023.

A. Background

In this section, we recall some facts from probability theory that were used in proving our main results, and then present the details of the `BCS-Detector` scheme of Shekhar & Ramdas (2023).

Facts from probability. We begin by recalling the following statement about the expectation of a random sum of random variables (Durrett, 2019, Theorem 2.6.2).

Fact A.1 (Wald's equation)**.** *Suppose $X_1, X_2, \ldots$ denote a sequence of i.i.d. random variables with $\mathbb{E}[X_i] < \infty$, and let S_n denote the sum $\sum_{i=1}^n X_i$ for all $i \geq 1$. Then, for any random stopping time N with $\mathbb{E}[N] < \infty$, we have $\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X_1]$.*

Next, we recall a standard concentration inequality for bounded random variables (Hoeffding, 1963, Theorem 1).

Fact A.2 (Hoeffding's inequality)**.** *Suppose $X_1, X_2, \ldots$ denote independent random variables supported on the interval $[a, b]$. Then, we have*

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| > u\right) \leq \exp\left(-\frac{u^2}{2n(b-a)^2}\right)$$

for any $u > 0$, where $S_n = \sum_{i=1}^n X_i$.

Finally, we state a result connecting the KL divergence and the total variation distance between two probability distribution (Tsybakov, 2009, Lemma 2.5).

Fact A.3 (Pinsker's inequality)**.** *Suppose P and Q denote two probability distributions. Then, we have*

$$d_{TV}(P, Q) \leq \sqrt{\frac{d_{KL}(P \parallel Q)}{2}},$$

where d_{TV} and d_{KL} denote the total variation distance and the KL divergence respectively.

Details of `BCS-Detector`. First, we recall the definition of backward CSs that are crucial to the design of `BCS-Detector`.

Definition A.4 (Backward CS)**.** Suppose $X_1, X_2, \ldots, X_n$ denote observations drawn from a distribution P_θ, parametrized by $\theta \in \Theta$. Then, a level-$(1-\alpha)$ backward CS for θ with n observations is a collection of sets $\{B_t^{(n)} : 1 \leq t \leq n\}$ satisfying the following:

- For any $t \in [n]$, the set $B_t^{(n)}$ is $\sigma(X_t, \ldots, X_n)$ measurable.
- The sets satisfy the uniform coverage guarantee: $\mathbb{P}\left(\forall t \in [n] : \theta \in B_t^{(n)}\right) \geq 1 - \alpha$.

Having introduced this notion of backward CSs, the `BCS-Detector` strategy of Shekhar & Ramdas (2023) proceeds as follows:

Definition A.5 (`BCS-Detector`)**.** Given a stream of observations $X_1, X_2, \ldots$, proceed as follows:

- Construct one level-$(1-\alpha)$ forward CS, denoted by $\{C_t : t \geq 1\}$
- With each new observation, construct a new backward CS $\{B_s^{(t)} : 1 \leq s \leq t\}$.
- Stop as soon as $\cap_{s=1}^t C_s \cap B_s^{(t)}$ becomes empty.

B. Proof of Proposition 2.10

Before presenting the proof of Proposition 2.10, we first recall some of details of the betting CS first proposed by Waudby-Smith & Ramdas (2023).

Background on betting CS. Given observations $X_1, X_2, \ldots$ drawn from an independent process with mean θ, the betting CS is defined as

$$C_n = \{s \in [0,1] : W_n(s) < 1/\alpha\}, \quad \text{with}$$
$$W_n(s) \coloneqq \prod_{i=1}^{n}(1 + \lambda_t(s)(X_t - s)), \quad \text{for all } s \in [0,1],$$

where $\{\lambda_t(s) : t \geq 1, s \in [0,1]\}$ are predictable bets, taking values in $[-1/(1-s), 1/s]$. For certain betting strategies, such as the *mixture method* (Hazan, 2016, § 4.3), the *regret* is logarithmic for all s. In particular, this implies that $\sup_{\lambda \in \left[\frac{-1}{1-s}, \frac{1}{s}\right]} \sum_{t=1}^{n} \log\left(1 + \lambda(X_t - s)\right) - \sum_{t=1}^{n} \log\left(1 + \lambda_t(s)(X_t - s)\right) \leq 2\log n$, for all $n \geq 13$. Note that this idea of using the mixture method with known regret guarantees, for the specific context of betting CS, was first considered by Orabona & Jun (2023). We now present the details of the proof of Proposition 2.10. First we show that under the condition that $T \geq 64\log(64/\Delta^2\alpha)/\Delta^2$, the analysis of the first setting (i.e., with unknown Θ_0) can be reduced to the second case (with known Θ_0). Then, we present the details of the proof of the second setting.

Proof of (2). Using the fact that $\log(1+x) \geq x - x^2/2$ for $x > -1$, we can further lower bound $\log W_n(s)$ with

$$\log W_n(s) \geq \sup_{\lambda \in \left[\frac{-1}{1-s}, \frac{1}{s}\right]} \left(\sum_{t=1}^{n} \lambda(X_t - s) - \frac{\lambda^2}{2}(X_t - s)^2 - \log(n^2) \right).$$

By setting the value of λ to $\frac{1}{n}\sum_{t=1}^{n} X_t - s$, and on simplifying, we can show that the betting CS after n observation satisfies $|C_n^{(1)}| \leq 4\sqrt{\log(n/\alpha)/n}$. This implies that for $T \geq 64\log(64/\Delta^2\alpha)/\Delta^2$, the width of the CS starting at time 1 must be smaller than $\Delta/2 = |\theta_1 - \theta_0|/2$. If the event $\mathcal{E} = \{\theta_0 \in C_T^{(1)}\}$ happens (recall that this is a probability $1-\alpha$ event), then we know that $\theta_0 \in \widetilde{\Theta}_0 \coloneqq \{\theta : |\theta - \theta_0| \leq \Delta/2\}$. This set $\widetilde{\Theta}_0$ plays the role of the known pre-change parameter class in the analysis. Hence the rest of the proof to obtain the upper bound stated in (2) proceeds exactly as in the case when the pre-change distribution class is known, and we present the details for the latter case next.

Proof of (3). Since the proof of this result is long, we break it down into four simpler steps.

Step 1: Bound $\mathbb{E}_T[(\tau - T)^+|\mathcal{F}_T]$ with the maximum expectation of a class of stopping times $(N_\theta)_{\theta \in \Theta_0}$. Introduce the stopping times $N_m = \inf\{n - m : C_n^{(m)} \cap \Theta_0 = \emptyset\}$, and recall from Proposition 2.9 that

$$\mathbb{E}_T[(\tau - T)^+|\mathcal{F}_T] \leq \mathbb{E}_T[N_{T+1}|\mathcal{F}_T] = \mathbb{E}_0[N_1].$$

The equality uses the fact that the post-change observations are independent of $\mathcal{F}_T$, and are drawn i.i.d. (hence stationary).

To simplify the ensuing argument, we will use C_n to denote $C_n^{(1)}$ for $n \geq 1$. Furthermore, we also assume that $\theta_1 < \theta$ for all $\theta \in \Theta_0$, and $\inf_{\theta \in \Theta_0} \theta = \theta_1 + \Delta$. The other case, can be handled in an exactly analogous manner.

By the definition of betting CS, the stopping time N_1 can be written as the supremum of a collection of stopping times:

$$N_1 = \sup_{\theta \in \Theta_0} N_\theta,$$

where $N_\theta = \inf\{n \geq 1 : W_n(\theta) \geq 1/\alpha\}$.

Step 2: Bound $(N_\theta)_{\theta \in \Theta_0}$ with a monotonic class of stopping times. Next, we will upper bound each N_θ with another stopping time γ_θ, which have the property that $\gamma_{\theta'} < \gamma_\theta$ for $\theta' > \theta$. In particular, using the regret guarantee of the betting strategy, observe the following:

$$\begin{aligned}
\log(W_n(\theta)) &\geq \sup_{\lambda \in \left[\frac{-1}{1-\theta}, \frac{1}{\theta}\right]} \sum_{t=1}^{n} \log(1 + \lambda(X_t - \theta)) - 2\log n \\
&\geq \sup_{\lambda \in \left[0, \frac{1}{1-\theta}\right]} \sum_{t=1}^{n} \log(1 + \lambda(\theta - X_t) - 2\log n \\
&\coloneqq Z_n(\theta) - 2\log n.
\end{aligned}$$

Define a new stopping time $\gamma_\theta = \inf\{n \geq 1 : Z_n(\theta) - 2\log n \geq \log(1/\alpha)\}$, and note that the above display implies $\gamma_\theta \geq N_\theta$, and thus we have $N_{T+1} - T \leq \sup_{\theta\in\Theta_0} \gamma_\theta$. We now show the monotonicity of γ_θ.

For any $\theta' > \theta$, we have $\lambda(\theta' - X_t) \geq \lambda(\theta - X_t)$ for any $\lambda > 0$, which implies that $\sum_{t=1}^n \log(1 + \lambda(\theta' - X_t)) \geq \sum_{t=1}^n \log(1 + \lambda(\theta - X_t))$. Thus, we have the following relation (for $\theta' > \theta$):

$$\begin{aligned} Z_n(\theta') &= \sup_{\lambda\in\left[0,\frac{1}{1-\theta'}\right]} \sum_{t=1}^n \log(1 + \lambda(\theta' - X_t)) \\ &\geq \sup_{\lambda\in\left[0,\frac{1}{1-\theta}\right]} \sum_{t=1}^n \log(1 + \lambda(\theta' - X_t)) \\ &\geq \sup_{\lambda\in\left[0,\frac{1}{1-\theta}\right]} \sum_{t=1}^n \log(1 + \lambda(\theta - X_t)) = Z_n(\theta). \end{aligned}$$

Thus, $Z_n(\theta') \geq Z_n(\theta)$, which implies that $\gamma_{\theta'} \leq \gamma_\theta$, and in particular, $\gamma_\theta \leq \gamma_{\theta_1+\Delta}$ for all $\theta \in \Theta_0$. This leads to the required conclusion

$$N_{T+1} - T \leq \sup_{\theta\in\Theta_0} N_\theta \leq \sup_{\theta\in\Theta_0} \gamma_\theta \leq \gamma_{\theta_1+\Delta}.$$

This is a crucial step, as it reduces the task of analyzing the supremum of a large collection of stopping times, into that of analyzing a single stopping time $\gamma_{\theta_1+\Delta}$.

Step 3: Bound $\gamma_{\theta_1+\Delta}$ with the 'oracle' stopping time ρ^*. Let $\lambda^* \equiv \lambda^*(\theta_1 + \Delta)$ denote the log-optimal betting fraction, defined as $\operatorname{argmax}_{\lambda\in[0,1/(1-\theta_1-\Delta)} \mathbb{E}[\log(1 + \lambda(\theta_1 + \Delta - X))]$, where X is drawn from the post-change distribution. By definition then, we have

$$\begin{aligned} Z_n(\theta_1 + \Delta) &\geq \sum_{t=1}^n \log(1 + \lambda^*(\theta_1 + \Delta - X_t)) \\ &\coloneqq Z_n^*(\theta_1 + \Delta), \end{aligned}$$

which immediately implies

$$\gamma_{\theta_1+\Delta} \leq \rho^* \coloneqq \inf\{n \geq 1 : Z_n^*(\theta_1 + \Delta) \geq \log(n^2/\alpha)\}.$$

The stopping time ρ^* is much easier to analyze as it is the first crossing of the boundary $\log(n^2/\alpha)$ by the random walk $Z_n^*(\theta_1 + \Delta)$ with i.i.d. increments.

Step 4: Evaluate the expectation of ρ^*. Observe that $Z_n^* \equiv Z_n^*(\theta_1 + \Delta) = \sum_{t=1}^n V_t$, with $V_t = \log(1 + \lambda^*(\theta_1 + \Delta - X_t))$. Without loss of generality, we can assume that $\lambda^* < 1/(1 - \theta_1 - \Delta)$ (if not, we simply repeat the argument with $\lambda^* - \epsilon$ for an arbitrarily small $\epsilon > 0$), and hence $(V_t)_{t\geq 1}$ are i.i.d. and bounded increments, which means that $\mathbb{E}[V_t] < \infty$. In fact, by the dual definition of the information projections (Honda & Takemura, 2010), we have $\mathbb{E}[V_t] = K_2 \equiv K_2(P_1, \Theta_0)$. Next, with $n_0 \coloneqq \inf\{n \geq 1 : \log(n^2/\alpha)/n < K_2/2\}$, we have for $n \geq n_0$ by an application of Hoeffding's inequality (Fact A.2 in Appendix A):

$$\mathbb{P}\left(\rho^* > n\right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^n V_t - K_2 \leq -\frac{K_2}{2}\right) \leq \exp\left(-c''n\right),$$

for some $c'' > 0$. Hence, the expectation of ρ^* satisfies

$$\begin{aligned} \mathbb{E}[\rho^*] &= \sum_{n\geq 0} \mathbb{P}\left(\rho^* > n\right) \leq n_0 + \sum_{n\geq n_0} \exp\left(-c''n\right) \\ &= n_0 + \frac{e^{-c''n_0}}{1 - e^{-c''}} < \infty. \end{aligned}$$

Thus, both ρ^* and $(V_t)_{t\geq 1}$ have bounded expectations, and we can appeal to Wald's lemma (Fact A.1 in Appendix A) to obtain $\mathbb{E}[Z^*_{\rho^*}] = \mathbb{E}[\rho^*]K_2$. Furthermore, by the definition of ρ^*, and the boundedness of $(V_t)_{t\geq 1}$, we can upper bound $\mathbb{E}[Z^*_{\rho^*}]$ with $\log(1/\alpha) + 2\log(\mathbb{E}_T[\rho^*]) + c'$, where $c' = \max\{\log(1+\lambda^*_{\theta_1+\Delta}), \log(1-\lambda^*_{\theta_1+\Delta})\}$. In other words, we have

$$\mathbb{E}[\rho^*] \leq \frac{\log(1/\alpha) + 2\log(\mathbb{E}[\rho^*]) + c'}{K_2},$$

which on further simplification, gives us $\mathbb{E}[\rho^*] = \mathcal{O}\left(\frac{\log(1/\alpha K_2)}{K_2}\right)$. This completes the proof.

C. Details of Experiments

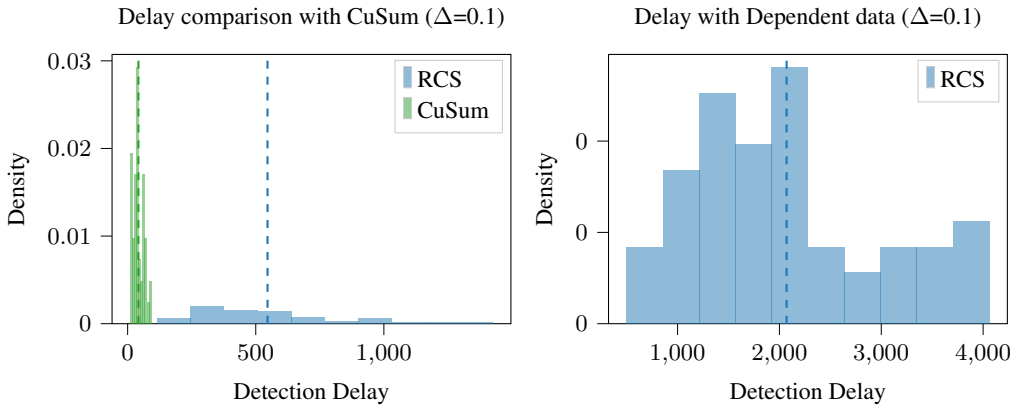

Figure 3: The plot on the left compares the distribution of the detection delay of CuSum (with exact knowledge of pre- and post-change distributions) and `RCS-Detector`. As expected the performance of CuSum is significantly better than `RCS-Detector` in problems where precise information of the distributions is available. The plot on the right shows the detection delay performance of `RCS-Detector` on a problem with dependent data: a situation in which the `BCS-Detector` becomes inapplicable.

Confidence Sequence. Recall that in our experiments with bounded observations, we employed the following Hoeffding CS developed by Waudby-Smith & Ramdas (2023):

$$C_t = [\tilde{\mu}_t \pm w_t], \quad \text{where}$$
$$\tilde{\mu}_t = \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i}, \quad w_t = \frac{\log(2/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t \lambda_i}, \quad \text{and} \quad \lambda_i = \sqrt{\frac{8\log(2/\alpha)}{i(i+1)}} \wedge 1.$$

The main reason for using this CS is that it has a closed-form expression, which makes the `RCS-Detector` and `BCS-Detector` implementations based on this CS computationally feasible.

Heuristics. In many cases, the Hoeffding CS can be wider than state-of-the-art methods, such as the betting CSs of Waudby-Smith & Ramdas (2023), which are computationally more expensive. As a result, the `RCS-Detector` and `BCS-Detector` instances based on Hoeffding CS can be very conservative when there is no change (i.e., their actual ARL can be much larger than $1/\alpha$). To address this, in our ARL experiments, we shrunk the width of the CS by a multiplicative factor less than one; we made the same changes in both `RCS-Detector` and `BCS-Detector` to allow for comparison.

To further reduce the computational cost of estimating their ARLs, we also checked the stopping conditions of `RCS-Detector` and `BCS-Detector` at intervals (i.e., every 10 steps, or 20 steps, etc.), instead of checking it every round. This allowed us to run the ARL experiments for longer horizons.

Cusum. Both the `RCS-Detector` and `BCS-Detector` schemes require minimal information about the pre- and post-change data distributions. This generality, however, comes at the cost of weaker detection-delay performance in

situations where additional information is known about the distributions. As an extreme case, if the pre- and post-change distributions are known exactly, then a simpler change detection scheme, such as CuSum, is more appropriate. Recall that CuSum strategy proceeds as follows:

$$\tau_C = \inf\{n \geq 1 : W_n \geq c_\alpha\}, \quad \text{where} \quad W_0 = 0, \quad W_n = \max\left\{0, W_{n-1} + \log\left(\frac{dP_1}{dP_0}(X_n)\right)\right\},$$

and c_α is an appropriately chosen threshold to control the ARL at $1/\alpha$ if $T = \infty$. This scheme requires the precise knowledge of the likelihood ratio of the post- and pre-change distributions, and hence its performance is significantly better than `RCS-Detector` when such information is available, as illustrated in Figure 3.

Non-independent data. Since the `RCS-Detector` only uses forward CSs, it can work with certain types of dependent data-streams, where `BCS-Detector` cannot be applied. As a simple example, let P_{0a}, P_{0b} denote two distributions on $[0, 1]$ with mean μ_0, and (P_{1a}, P_{1b}) be two distributions with mean $\mu_1 \neq \mu_0$. Let $X_1 \sim P_{0a}$, and for $n \geq 2$:

- If $X_{n-1} < 0.5$ and $n \leq T$, we have $X_n \sim P_{0a}$.
- If $X_{n-1} \geq 0.5$ and $n \leq T$, we have $X_n \sim P_{0b}$.
- If $X_{n-1} < 0.5$ and $n > T$, we have $X_n \sim P_{1a}$.
- If $X_{n-1} \geq 0.5$ and $n > T$, we have $X_n \sim P_{1b}$.

Because of this dependence structure, we cannot construct Backward Confidence Sequences for this data stream, and thus `BCS-Detector` is not applicable to this problem. However, the `RCS-Detector` is still applicable, and its detection performance is plotted in Figure 3.

- Fix Figure 3 + caption + color of RCS + title of the figure