Sequential Changepoint Detection via Backward Confidence Sequences

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

Abstract

We present a simple reduction from sequential estimation to sequential changepoint detection (SCD). In short, suppose we are interested in detecting changepoints in some parameter or functional θ of the underlying distribution. We demonstrate that if we can construct a confidence sequence (CS) for θ, then we can also successfully perform SCD for θ. This is accomplished by checking whether two CSs — one forwards and the other backwards — ever fail to intersect. Since the literature on CSs has been rapidly evolving recently, the reduction provided in this paper immediately solves several old and new change detection problems. Further, our "backward CS", constructed by reversing time, is new and potentially of independent interest. We provide strong nonasymptotic guarantees on the frequency of false alarms and detection delay, and demonstrate numerical effectiveness on several problems.

1. Introduction

We study the problem of sequential changepoint detection (SCD), where the goal is to quickly detect any changes in the distribution generating a stream of observations, while controlling the false alarm rate at a specified level. Formally, for some (possibly infinite-dimensional) index set Θ, let $\{P_\theta\}_{\theta \in \Theta}$ denote a class of distributions on some observation space $\mathcal{X}$. Suppose that for some $T \geq 1$, the observations $\{X_t : 1 \leq t \leq T\}$ are drawn i.i.d. from P_{θ_0} with $\theta_0 \in \Theta$, and $\{X_t : t > T\}$ are drawn i.i.d. from P_{θ_1} for some $\theta_1 \in \Theta$, with $\theta_1 \neq \theta_0$. Then, the SCD problem involves deciding between the null $H_0 : \{T = \infty\}$, meaning no change occurred, and the alternative $H_1 = \cup_{i \in \mathbb{N}} \{T = i\}$.

Since the observations arrive sequentially, our task is to design a random stopping time, τ, adapted to the natural filtration $\{\mathcal{F}_t : t \geq 1\}$ with $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$, at which we reject the null. A good stopping rule τ takes large values under the the null (i.e., when $T = \infty$), while minimizing the time required to detect the change under the alternative (i.e., when $T < \infty$). Formally, when $T = \infty$, we require the *average run length (ARL)*, $\mathbb{E}_\infty[\tau]$, to be lower bounded by $1/\alpha$, for a given $\alpha \in (0, 1]$, while ensuring a small *detection delay*, $\mathbb{E}_T[(\tau - T)^+]$, when $T < \infty$. Informally, this means that we will have a false alarm roughly every $1/\alpha$ steps, so the reader may use $\alpha = 10^{-3}$ as a rough guideline. (We also briefly discuss how to keep the probability of even a single false alarm below α, but there is a tradeoff between the false alarm guarantee and detection delay; detecting true changes quickly necessitates tolerating infrequent false alarms.)

The literature on the topic of sequential changepoint detection is vast, as this problem arises in several important real-world applications, such as quality control (Shewhart, 1930), monitoring power networks (Chen et al., 2015), analysis of genomes (Chen et al., 2011; Shen and Zhang, 2012), and epidemic detection (Baron et al., 2004; Yu et al., 2013). Some of the earliest works in this topic (Shewhart, 1925; Page, 1954; Shiryaev, 1963) assume that the pre- and post-change distributions admit known densities f_0 and f_1 (w.r.t. some common reference measure). These methods use statistics involving likelihood ratios, that can be computed efficiently in an incremental manner, and have also been shown to admit strong optimality properties. The ideas underlying these likelihood-based schemes have also been extended to the case of (finite-dimensional) parametric families of distributions, such as the exponential family; see Tartakovsky et al. (2014) for a detailed discussion. However, these parametric assumptions are often too stringent to be applicable to many practical applications, where the data distributions may lie in much larger, nonparametric, classes. Most of the ideas developed for the parametric setting, and in particular the likelihood-based schemes, fail to be applicable in the nonparametric case. With some exceptions discussed later, there are very few general principles for constructing nonparametric changepoint detection schemes. Our work in this paper addresses this issue, by developing a conceptually simple 'meta-algorithm' for transforming any confidence sequence construction into a powerful changepoint detection method. As a consequence, we can immediately build upon

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

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

the recent progress in constructing confidence sequences to instantiate new changepoint detection methods.

Remark 1. We note that the SCD problem is usually studied in two settings, that differ from each other is a very subtle manner. In the first (and the more common) setting, the pre- and post change distributions (P_{θ_0} and P_{θ_1}) are assumed to lie in two different, and usually well-separated, classes of distributions. Using our notation, this is equivalent to assuming that there exist two known disjoint sets Θ_0 and Θ_1, such that $\theta_i \in \Theta_i$, for $i = 0, 1$. The second setting, that is the subject of our paper, assumes less information. That is, both θ_0 and θ_1 are assumed to lie in some common index set (Θ in our notation), and the only condition is that $\theta_0 \neq \theta_1$. Hence, the first setting is in some sense "easier", as the additional knowledge about Θ_0 and Θ_1 (their size, and their separation) can be exploited to design appropriate SCD schemes. While there exist some works that develop methods for SCD in the second setting, those schemes often rely on the specific structure of the problems considered. In this paper, we address this issue by developing a general principle for designing SCD schemes in the second setting.

2. Preliminaries

The primary technical tool we use in our strategy are time-uniform version of confidence sets, called *confidence sequences* (CSs), that were first introduced in statistics literature by Darling and Robbins (1967). We present a definition adapted to our problem below.

Definition 2 (`Confidence Sequences`)**.** Suppose $X_1, X_2, \ldots$ are drawn i.i.d. from P_θ, for some $\theta \in \Theta$. Then, for any $\alpha \in (0, 1)$, a level-$(1-\alpha)$ CS, denoted by $\{C_t : t \geq 1\}$, is a collection of subsets $C_t \subset \Theta$, such that **(i)** C_t is $\sigma(X_1, \ldots, X_t)$-measurable and **(ii)** $\mathbb{P}\big(\forall t \geq 1 : \theta \in C_t\big) \geq 1 - \alpha$.

Remark 3. Due to the time-uniformity in the definition of CSs, we can replace the confidence set C_t with the smaller set $\widetilde{C}_t \coloneqq \cap_{s \leq t} C_s$. The new CS, $\{\widetilde{C}_t : t \geq 1\}$, consists of *nested* confidence sets; that is, $\widetilde{C}_t \subset \widetilde{C}_s$ for $s < t$.

Remark 4. The data do not need to be i.i.d. for defining CSs. The above definition can be easily generalized to the case of independent random variables, with $X_t \sim P_{\vartheta_t}$, with $\vartheta_t \in \Theta$ (see Appendix A). This, however, requires that Θ is endowed with the notions of addition and scalar multiplication (a sufficient condition is that Θ is a vector space), which we implicitly assume when needed.

We now introduce a notion of the 'size' of the confidence set C_t, that reflects the amount of uncertainty.

Definition 5 (`CS width`)**.** Let Θ be endowed with a distance metric d, and let $\{C_t : t \geq 1\}$ denote a level-$(1-\alpha)$ CS constructed on observations $X_1, X_2, \ldots$ drawn i.i.d. from P_θ. A function $w(t, \theta, \alpha)$ denotes the pointwise width (bound) of the CS, if for all $t \in \mathbb{N}$ and $\theta \in \Theta$, we have $\sup_{\theta', \theta'' \in C_t} d(\theta', \theta'') \leq w(t, \theta, \alpha)$. We define the uniform width over Θ, as $w(t, \Theta, \alpha) = \sup_{\theta \in \Theta} w(t, \theta, \alpha)$.

As we will see later in Section 5, most of the non-trivial CSs have their pointwise widths (and often, the uniform widths as well) converging to 0 with the number of observations.

Example 6. Consider independent random variables $\{X_t : t \geq 1\}$, with $X_t \sim N(\theta, 1)$ and $\theta \in \mathbb{R}$ for all $t \geq 1$. In this case, the parameter set is $\Theta = \mathbb{R}$. For this process, we can define the CS $\{C_t : t \geq 1\}$ as follows: $C_t = [\bar{X}_t - w_t/2, \bar{X}_t + w_t/2]$, where $\bar{X}_t = (1/t) \sum_{i=1}^t X_i$, and $w_t = 3.4\sqrt{(\log\log(2t) + 0.72\log(10.4/\alpha))/t}$. Thus, if we endow the parameter space Θ with the metric $d(\theta, \theta') = |\theta - \theta'|$, we observe that the uniform width of C_t, denoted by $w(t, \Theta, \alpha) = w_t$, converges to 0.

Related Work. As we mentioned earlier, a large part of the existing SCD literature focuses on the parametric setting. We refer the reader to some recent surveys, such as those by Veeravalli and Banerjee (2014); Xie et al. (2021), and the textbook by Tartakovsky et al. (2014) for details. In this section, we discuss some results on nonparametric SCD methods that are more relevant to our work.

Shin et al. (2022) developed a novel framework for changepoint detection by introducing *e-detectors*; obtained by combining a sequence of e-processes defined uniformly over the class of pre-change distribution. They showed that their resulting strategy using e-detectors controls the ARL under very general conditions, and also proved the optimality of their schemes (in terms of worst-case detection delays) in some cases. However, as we mentioned in Remark 1, their framework is applicable mainly in cases where the pre- and post-change distributions are known to lie in different classes. Our techniques, described in Section 3 and Section 4, addresses this issue.

Maillard (2019b) considered the task of detecting a change in the mean of a sequence of independent, univariate, sub-Gaussian random variables; and proposed an SCD method by deriving a new, doubly time-uniform confidence sequence for the scan statistics associated with a generalized likelihood ratio scheme (Lai and Xing, 2010). The original proof of this concentration inequality (Maillard, 2019b, Theorem 4) was incomplete, and a corrected version (with an additional $\log\log$ term) was obtained by the author in (Maillard, 2019a, Chapter 3, § 4.1). For the resulting scheme, Maillard (2019a) obtained bounds on the probability of false alarm and on the detection delay, and also established the optimality of this scheme under certain scenarios (such as for Gaussian observations). Unlike Maillard (2019a), our SCD framework is applicable to a much wider class of problems beyond univariate mean testing. Nevertheless, when specialized to the case of univariate Gaussian observations, our scheme matches the optimal detection delay bound, while also providing control over the ARL (instead of the probability of false alarm).

Puchkin and Shcherbakova (2023) considered the SCD problem under the assumption that both the pre- and post-change distributions admit densities w.r.t. a common reference measure, and proposed a strategy based on learning a discriminator to estimate the density ratio. They showed that their strategy can control ARL at the required level, and also obtained high probability upper bound on the detection delay in terms of the Jensen Shannon (JS) divergence and the L_2 norm of the difference of densities. Unlike them, our framework does not require the existence of densities, and it also works for distributions that are separated in terms of a large class of metrics, and not just the JS divergence.

Another class of nonparametric schemes for SCD are based on the kernel-MMD metric, first employed by Gretton et al. (2012) for designing powerful nonparametric two-sample tests. Li et al. (2019) proposed a SCD scheme based on a variant of the block-MMD statistic (Zaremba et al., 2013) computed using the observations, and a block of pre-change data. More recently, Flynn and Yoo (2019) and Wei and Xie (2022) proposed new SCD schemes that use linear and block-MMD statistics to define nonparametric analogues of the CuSum test of Page (1954). However, these schemes suffer from weak theoretical guarantees on the detection delay. Furthermore, similar to the case of Puchkin and Shcherbakova (2023), the strategies for designing these SCD schemes are specific to the kernel-MMD metric; and there is no obvious way to extend them to other popular metrics, such as the Kolmogorov-Smirnov metric. Our work addresses these issues.

Similarly, most other existing works in SCD are geared towards specific problem settings. Hence, both the design of the scheme as well as their analysis are strongly tied to the details of the problem being studied. Examples include empirical likelihood based methods for distributions on finite alphabets (Lau et al., 2018), nearest-neighbor techniques for multivariate or non-euclidean data (Chen, 2019), and spectral scan statistics for graph valued data (Sharpnack et al., 2013). However, our objective in this paper is different: instead of developing a powerful SCD scheme for a specific task, we develop an abstract unifying template for designing SCD schemes, that can then be instantiated for a large range of (old and new) SCD problems.

Our Contributions. In Section 3, we first present (as a warmup) a changepoint detection scheme that uses a single level-α forward CS. Our strategy is to stop as soon as the CS becomes 'inconsistent'; that is, it includes a point that it had previously discarded. We show in Proposition 8, that this simple strategy controls the probability of false alarm at level α, and we also obtain a high probability upper bound on its detection delay. However, this scheme is too conservative as its ARL is infinite, and in practice this might result in large detection delays, especially when T is large.

In Section 4, we present our main strategy that proceeds by checking at each time t, whether a forward CS and a backward CS (a new notion) are consistent, and stops whenever an inconsistency is detected. In Theorem 13, we show that the ARL of this scheme is at least $1/(2\alpha)$, and we characterize its expected detection delay under general conditions.

Finally, in Section 5, we demonstrate the power and generality of our proposed scheme by instantiating it with five different confidence sequences. The general bound on the detection delay obtained in Theorem 13 easily translate into problem-specific upper bounds in all these cases, and we also empirically verify the theoretical predictions through some simple numerical simulations.

3. Warmup: change detection via a forward CS

Before presenting our general scheme in the next section, we first introduce a simpler SCD method that only uses a single forward CS. We refer to this scheme as the `FCS-Detector`. The idea underlying this scheme is that if there is a change in the distribution generating the observations $\{X_t : t \geq 1\}$, then the intersection of the CS will eventually end up being empty. Formally, we proceed by constructing a level-$(1-\alpha)$ confidence sequence (CS) for the unknown θ_0, denoted by $\{C_t : t \geq 1\}$, as introduced in Definition 2. When there is no changepoint, then the CS satisfies $\mathbb{P}_\infty(\forall t \in \mathbb{N} : \theta_0 \in C_t) \geq 1-\alpha$. However, if there is a changepoint at some time T, we expect that the confidence sets, C_t, deviate away from the confidence set C_T, for $t > T$. Eventually, after sufficiently many post-change observations, the confidence sequence will be inconsistent and self-contradictory. That is, at some time t such that $t - T$ is large enough, we expect that $\cap_{s=1}^t C_s = \emptyset$. We thus define the stopping time, τ, as the smallest t at which the above inconsistency is observed.

Definition 7 (`FCS-Detector`)**.** Given observations $X_1, X_2, \ldots$, we construct a confidence sequence (CS), denoted by $\{C_t : t \geq 1\}$ for the pre-change parameter θ_0. We stop at time $\tau := \min\{n \geq 1 : \exists t < n,\, C_t \cap C_n = \emptyset\}$.

This strategy satisfies the following properties.

Proposition 8. *Consider a change point detection problem with observations $X_1, X_2, \ldots$ drawn i.i.d. from P_{θ_0} for $t \leq T$ and from P_{θ_1} for $t > T$, with T lying in $\mathbb{N} \cup \{\infty\}$ and $\theta_0, \theta_1 \in \Theta$. Suppose for any $\theta \in \Theta$, we can construct confidence sequences $\{C_t : t \geq 1\}$, with uniform width $w(t) \equiv w(\cdot, \Theta, \alpha)$. Then, we have the following:*

(i) *When $T = \infty$, the `FCS-Detector` controls the probability of false alarm (PFA) at level α. That is, $\mathbb{P}_\infty(\tau < \infty) \leq \alpha$.*

(ii) *Suppose $T < \infty$ is large enough to ensure that $w(T) < d(\theta_0, \theta_1)$. For $t > T$, define $\lambda_t := T/t$, and $\widetilde{\theta}_t := \lambda_t \theta_0 + (1 - \lambda_t)\theta_1$. Then, we have $\tau - T \leq \min\left\{t - T : w(t) + w(T) \leq d(\theta_0, \widetilde{\theta}_t)\right\}$, with probability at least $1 - \alpha$.*

Remark 9. If d is induced by a norm on Θ, we can bound the detection delay under the event $\mathcal{E}$ by $(\tau - T)^+ \leq \min\left\{t - T : w(t) + w(T) \leq \frac{t-T}{t} d(\theta_0, \theta_1)\right\}$. In many instances, we have $w(t) \approx 1/\sqrt{t}$ suppressing logarithmic factors. Then, the above expression implies that from 'small' T, the delay is linear in T, while for large T, the delay behaves approximately like $\sqrt{T}$. The details of these calculations, along with plots of delay versus T are in Appendix B.

Remark 10. The condition on T used above for obtaining the bound on detection delay is necessary, because we do not assume that the pre-change distribution (i.e., the parameter θ_0) is known to us. Thus, to be able to detect the change in distribution, we must have enough observations from the pre-change distribution to estimate θ_0 accurately enough, in comparison the magnitude of change, $d(\theta_0, \theta_1)$. As an extreme example, if $T = 1$, no method can realistically detect that a change occurred (since it is statistically plausible that all the data are simply i.i.d. and no change occurred at all). Said differently, $T = 0$ and $T = \infty$ are information theoretically equivalent, and we need to be far enough away from those extremes for practical detectability.

When $T = \infty$, the `FCS-Detector` continues sampling without stopping w.p. at least $1 - \alpha$. Hence, its ARL is infinite. This makes it is too conservative in detecting the changepoint — we cannot provide an upper bound on the expected detection delay when $T < \infty$, and can only characterize the delay under an event of probability $1 - \alpha$ (see Figure 5 in Appendix B). We address this next, by proposing a scheme that augments the forward CS with a series of backward CSs.

4. Change detection via a backward CS

We now introduce our main SCD strategy that addresses the two drawbacks of the simpler SCD strategy discussed in the previous section. Informally, the idea underlying our strategy is as follows: in each round $t \geq 2$, we construct the usual forward CS, and a new backward CS (using reversed observations, see Definition 11). We refer to this scheme as the `BCS-Detector`. If there has been a changepoint, we expect the forward and backward CSs to concentrate on different regions of Θ (i.e., around θ_0 and θ_1 resp.). Hence, we stop as soon as they become inconsistent. See **??** in Appendix A for a visual illustration, and Appendix F for an interpretation of our scheme in terms of repeated sequential tests.

We now present our definition of backward CSs.

Definition 11 (Backward Confidence Sequences)**.** Let $X_1, X_2, \ldots$ be drawn i.i.d. from P_θ, for some $\theta \in \Theta$. For any $n \geq 1$, we say that a sequence of sets $\{B_t^{(n)}\}_{1 \leq t \leq n} \subseteq \Theta$ is a backward CS, if **(i)** $B_t^{(n)}$ is $\sigma(X_t, \ldots, X_n)$ measurable, and **(ii)** $\mathbb{P}\left(\forall t \in [n] : \theta \in B_t^{(n)}\right) \geq 1 - \alpha$.

Note that for $n > 1$, a forward CS C_t does not satisfy the first condition, since C_t is built using $X_1, \ldots, X_t$, but $B_t^{(n)}$ can only use $X_t, \ldots, X_n$. But, a backward CS at any n can be interpreted as the usual forward CS, introduced in Definition 2, constructed on observations seen in a reverse order from n to 1; see Appendix A for details.

Without loss of generality, we assume that any CS consists of a nested sequence of sets, as discussed in Remark 3. In other words, at a given time n, C_n is the smallest set among $\{C_t : t \in [n]\}$, while $B_1^{(n)}$ is the smallest among $\{B_t^{(n)} : t \in [n]\}$. We say that 'the two confidence sequences $\{C_t : t \geq 1\}$ and $\{B_t^{(n)} : t \in [n]\}$ 'do not intersect at time n' if $C_r \cap B_s^{(n)} = \emptyset$ for some $1 \leq r, s \leq n$. It is easy to check that two nested CSs do not intersect if and only if their smallest sets (i.e., C_n and $B_1^{(n)}$) do not intersect.

When $T = \infty$, at every time n, both CSs $\{C_t : t \geq 1\}$ and $\{B_t^{(n)} : t \in [n]\}$ will contain θ_0 with probability at least $1 - 2\alpha$, and thus they will intersect with the same probability. This motivates us to stop and declare a changepoint at the first time n at which they do not intersect.

Definition 12 (`BCS-Detector`)**.** Given observations $X_1, X_2, \ldots$, suppose we construct a forward CS $\{C_t : t \geq 1\}$ and new backward CSs $\{B_t^{(n)} : t \in [n]\}$ for every $n \geq 1$. Assume that all the constructed CSs are nested. Then, we define the stopping time, τ, as the first time at which the forward and backward CSs do not intersect: $\tau := \inf\{n \geq 1 : C_n \cap B_1^{(n)} = \emptyset\}$.

Illustration of the `BCS-Detector` strategy. In Figure 1, we illustrate the intuition underlying our general `BCS-Detector` strategy, introduced above, using the task of detecting change in means of bounded observations (details in Section 5.2). The three plots in Figure 1 highlight the following aspects of our scheme:

- Prior to the changepoint (or if there is no changepoint), both the forward and backward CSs concentrate around the same region in parameter space Θ (in this case, $[0, 1]$). In particular, note that in this case, for all values of t, one of the confidence intervals (CI) is entirely contained in the corresponding CI of the other CS.

- As observations from the post change distribution start arriving, we expect the forward and backward CSs to start drifting away from each other. This is illustrated in the second plot of Figure 1.

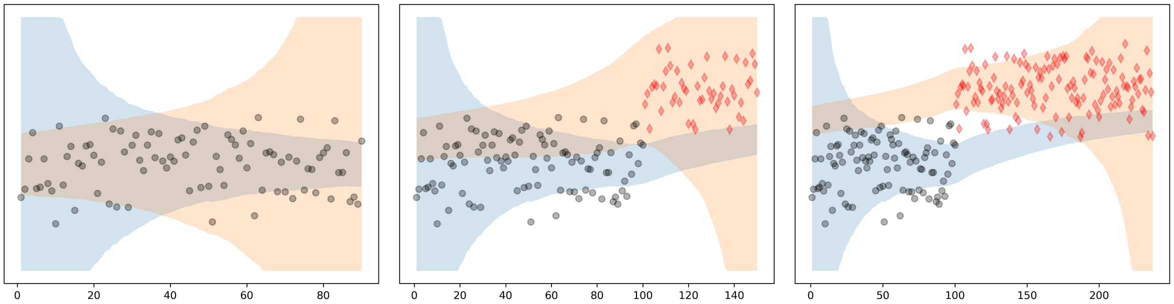

Figure 1. The plots illustrate the general ideas underlying our `BCS-Detector` strategy. Prior to the changepoint (first plot), both the forward and backward CSs have significant overlap. After getting some post-change observations (middle plot), the backward CS starts to drift away from the forward CS, although the deviation is not enough for the two CSs to become inconsistent. Finally, the last plot shows the scenario, where a sufficiently large number of post-change observations have arrived, which causes the forward and backward CSs to disagree. When this occurs, our scheme stops and rejects the null.

- Finally, after sufficiently many post-change observations, the backward CS starts concentrating around the post-change parameter θ_1, as a result of which some of the backward CIs become disjoint with the forward CIs. Our `BCS-Detector` scheme uses this occurrence as a signal to stop, and declare a changepoint.

We now present the main result of this section.

Theorem 13. *Consider a SCD problem with observations $X_1, X_2, \ldots$ drawn i.i.d. from P_{θ_0} for $t \leq T$ and from P_{θ_1} for $t > T$, with T lying in $\mathbb{N} \cup \{\infty\}$ and $\theta_0, \theta_1 \in \Theta$. Suppose for any $\theta \in \Theta$, we can construct confidence sequences $\{C_t : t \geq 1\}$ with pointwise width $w(\cdot, \theta, \alpha)$. Then, we have the following:* ***(i)*** *When, there is no changepoint ($T = \infty$), the `BCS-Detector` satisfies $\mathbb{E}_\infty[\tau] \geq \frac{1}{2\alpha} - \frac{3}{2}$.*
(ii) *Suppose $T < \infty$, and the pre-change parameter θ_0 is not known. Introduce the event $\mathcal{E} = \{\theta_0 \in C_t : 1 \leq t < T\}$, and note that $\mathbb{P}_T(\mathcal{E}) \geq 1 - \alpha$ by construction. Then, for $\alpha \in (0, 0.5)$, we have $\mathbb{E}_T[(\tau - T)^+ | \mathcal{E}] \leq 3\frac{u_0(\theta_0, \theta_1, T)}{1-\alpha}$, where $u_0 := \min\{t \geq 1 : w(t, \theta_1, \alpha) + w(T, \theta_0, \alpha) < d(\theta_1, \theta_0)\}$.*

Recall from Definition 5 that $w(t, \theta_1, \alpha)$ denotes the width of the level-α confidence set with t observations, when the true parameter is θ_1. The proof of this theorem is given in Appendix C. In many problem instances, the pre-change parameter θ_0 is known as it represents the 'natural state' of the process being observed. We can specialize the above result stated to this case as follows.

Corollary 14. *Suppose the pre-change parameter θ_0 is known, and $T < \infty$. Then, we have $\mathbb{E}_T[(\tau - T)^+] \leq (3/(1 - \alpha))t_0(\theta_0, \theta_1, T)$, where $t_0 \equiv t_0(\theta_0, \theta_1) := \min\{t \geq 1 : w(t, \theta_1, \alpha) < d(\theta_1, \theta_0)\}$.*

Remark 15. These results demonstrate how the drawbacks of the `FCS-Detector` (introduced in Section 3) are addressed by carefully incorporating the idea of backward CSs in the design strategy. In particular, this new scheme, called the `BCS-Detector`, is less conservative, and has a finite lower bound on the ARL under the null. More importantly, when $T < \infty$, the expected detection delay of `BCS-Detector` is also finite, and furthermore, it is also independent of the value of the changepoint T. This is in contrast to the (high probability) bound on the detection delay for `FCS-Detector`, in which the detection delay increases approximately as $\sqrt{T}$, as the changepoint $T \to \infty$.

Remark 16 (`Other estimates`)**.** We can also construct an estimate of the changepoint, denoted by $\widehat{T}$, as the time at which C_t and $B_t^{(\tau)}$ are most separated: $\widehat{T} \equiv \widehat{T}(\tau) := \max\ \mathrm{argmax}_{1 \leq t \leq \tau}\ d\big(C_t, B_t^{(\tau)}\big) = \max\{s \leq \tau : d(C_s, B_s^{(\tau)}) = \max_{1 \leq t \leq \tau} d(C_t, B_t^{(\tau)})\}$. Additionally, we define an estimate of the magnitude of the change $\epsilon := d(\theta_0, \theta_1)$ as the separation between $C_{\widehat{T}}$ and $B_{\widehat{T}}^{(\tau)}$, as measured by the distance metric d: $\widehat{\epsilon} \equiv \widehat{\epsilon}(\tau) := \max_{\theta \in C_{\widehat{T}}, \theta' \in B_{\widehat{T}}^{(\tau)}}\ d(\theta, \theta')$. While we do not obtain theoretical guarantees, some empirical results in the next section indicate that these estimates are accurate for several instantiations of the `BCS-Detector`.

5. Instantiations of BCS-Detector

In the previous section, we introduced a conceptually simple device that allows us to transform any confidence sequence construction into a powerful, sequential changepoint detector. This allows us to instantiate our general changepoint detection meta-algorithm to various scenarios, by leveraging the recent progress in constructing confidence sequences. We illustrate this, by presenting a variety of parametric and nonparametric SCD problems in this section. The code for reproducing the empirical results is available here.

5.1. Parametric Change of Mean Detection

We begin by considering the simple case of univariate Gaussian mean changepoint detection. In this problem, we observe $\{X_t : t \geq 1\}$, drawn i.i.d. according to the distribution $N(\mu_t, 1)$, with $\mu_t = \theta_0$ for $t \leq T$ and $\mu_t = \theta_1$ for $t > T$. Note that in this problem, we have $\Theta = \mathbb{R}$ and we can set the distance metric, d, to be the absolute value of the difference. We will use the CS for Gaussian means, recently derived by Howard et al. (2021), that we had introduced earlier in Example 6. Furthermore, we can use the same expression for constructing the backward CS at any time n, denoted by $\{B_t^{(n)} : t \geq 1\}$, but with the order of observations reversed, as described in Section 4. The following result shows that in this parametric setting, our changepoint detection scheme achieves an order-optimal detection delay (i.e., optimal modulo poly-logarithmic factors):

Corollary 17. *Suppose $\{X_t : t \geq 1\}$ are drawn i.i.d. according to $P_{\theta_0} = N(\theta_0, 1)$ for $t \leq T$, and $P_{\theta_1} = N(\theta_1, 1)$ for $t > T$. Note that in this case, we have $d_{KL}(P_{\theta_1}, P_{\theta_0}) = (\theta_1 - \theta_0)^2/2$. Then, if $T = \Omega\left(\log(1/d_{KL}(P_{\theta_1}, P_{\theta_0}))/d_{KL}(P_{\theta_1}, P_{\theta_0})\right)$, then we have*

$$\mathbb{E}_T[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\left(\frac{\log\log(\frac{1}{d_{KL}(P_{\theta_1}, P_{\theta_0})}) + \log(\frac{1}{\alpha})}{d_{KL}(P_{\theta_1}, P_{\theta_0})}\right),$$

where $\mathcal{E}$ is the $(1-\alpha)$ probability event in Theorem 13.

Remark 18. If the pre-change mean (θ_0) is known, the above upper bound holds for the worst-case detection delay without the conditioning, defined as $J_L(\tau) := \sup_{T>0} \operatorname{esssup} \mathbb{E}[(\tau - T)^+ | \mathcal{F}_T]$. Under the assumption that θ_1 is also known, Lorden (1971) showed the following universal lower bound on this quantity: $\inf_{\tau'} J_L(\tau') = \frac{\log(1/\alpha)}{d_{\mathrm{KL}}(P_{\theta_1}, P_{\theta_0})}(1 + o(1))$, as $\alpha \to 0$. Thus, `BCS-Detector` matches this optimal performance, modulo logarithmic factors, without the knowledge of the post-change parameter. This is unlike some of the existing schemes for Gaussian mean change detection, such as Pollak and Siegmund (1991), which achieve the same order optimal detection delay, but with additional assumptions (known lower bound on change, and in the limit of $T \to \infty$).

Empirical Verification. We now verify the theoretical claims of our proposed changepoint detection scheme using observations drawn from a unit-variance normal distribution with the pre-change mean $\theta_0 = 0$, and post-change mean $\theta_1 = \Delta$. In Figure 2, we consider the case of $\Delta = 0.4$ with the change occurring at $T = 800$. The plots show the forward and backward CSs at the time at which the change is detected in a trial, as well as the distributions of the detection delay, estimated changepoint, and estimated change magnitude over 250 trials. The plots indicate that `BCS-Detector` detects changes quickly and accurately. In Figure 3, we study the variation of the average detection delay of our changepoint detection scheme as the change magnitude Δ is varied. The empirical results verify the expected proportionality to $1/\Delta^2$ of the average detection delay, as claimed by our theoretical results.

5.2. Nonparametric Change of Mean Detection

We now consider a nonparametric analog of the change of mean detection problem from the previous section. Here we assume that $X_1, X_2, \ldots$ are independent random variables taking values in a bounded interval $\mathcal{X} \subset \mathbb{R}$, which we set to $[0,1]$ without loss of generality. Prior to the changepoint T, we assume that the observations have a mean $\theta_0 \in \Theta = [0,1]$, while it changes to $\theta_1 \neq \theta_0$ after time T. In this case, we can use the empirical Bernstein (EB) confidence intervals developed by Waudby-Smith and Ramdas (2023). To state the closed-form expression of the EB confidence sequence, we first need to introduce the following terms: $\widehat{\mu}_t = \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}$, $\widehat{\sigma}_t^2 = \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \widehat{\mu}_t)^2}{t+1}$, $\lambda_t = \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_t^2 t \log(t+1)}} \wedge \frac{1}{2}$, $\widehat{\theta}_t = \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i}$, $v_t = (4/\log(2/\alpha))(X_t - \widehat{\mu}_{t-1})^2$, and $\Psi_E(x) = \frac{-\log(1-x) - x}{4}$, for all $x \in [0,1)$. Using these terms, we can now state the EB-CS derived by Waudby-Smith and Ramdas (2023) as follows: $C_t = [\widehat{\theta}_t + w_t/2, \widehat{\theta}_t - w_t/2]$, with $w_t = \frac{\log(2/\alpha) + \sum_{i=1}^t v_i \Psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i}$. Again, by an application of the general result, Theorem 13, we can get the following bound on the expected detection delay of the SCD scheme, that uses the above EB-CS.

Proposition 19. *Suppose $X_1, X_2, \ldots, X_T$ are drawn i.i.d. from a distribution on $\mathcal{X} = [0,1]$ with mean $\theta_0 \in \Theta = [0,1]$, while $X_{T+1}, \ldots$ are drawn from P_{θ_1} with mean $\theta_1 \neq \theta_0$. Then, assuming θ_0 is known, this instance of `BCS-Detector` (Definition 12) satisfies the following (with $\Delta := |\theta_0 - \theta_1|$, and $\sigma_1^2 = \mathbb{E}_{P_{\theta_1}}[(X - \theta_1)^2]$):*

$$\mathbb{E}_T[(\tau - T)^+] = \mathcal{O}\left(\sigma_1^2 \frac{\log(1/\Delta) + \log(1/\alpha)}{\Delta^2}\right).$$

Remark 20. While we stated the above result under the assumption that the pre- and post-change observations are i.i.d. from P_{θ_0} and P_{θ_1} respectively, we note that similar results can be obtained when the random variables are only independent, with fixed (pre- and post-change) means. Further, the assumption that θ_0 is known can also be waived, at the cost of conditioning on the 'good' event $\mathcal{E}$.

Remark 21. In this subsection, we have considered the task of change-of-mean detection in perhaps the simplest (but nontrivial) nonparametric setting. The same ideas developed here, can however, we extended easily to other interesting cases, such as the sub-Gaussian family using the CS derived by Howard et al. (2021), or for heavy-tailed distributions (Wang and Ramdas, 2022).

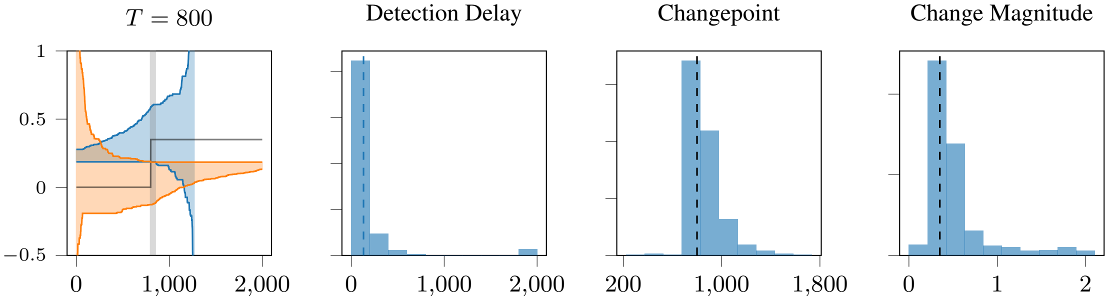

Figure 2. The figures show the performance of our changepoint detection scheme, `BCS-Detector`, with univariate Gaussian observations whose mean changes from 0 to 0.4 at the time $T = 800$. The first plot shows the forward and backward CSs at time of detection ($\tau = 1264$) in one of the trials, with the shaded gray region being the points at which the two CSs disagree. The next three plots show the empirical distribution of the detection delay, the estimated changepoint location, and the estimated changepoint magnitude over 250 repeated trials.

Empirical Verification. To verify our theoretical claims, we consider the case of an independent stream of observations supported on $\mathcal{X} = [0, 1]$, with pre- and post-change parameters, θ_0 and θ_1 respectively. We define distributions with specified means by taking approprirate mixtures of uniform distributions, as described in Appendix E. We plot the performance of our changepoint detection scheme for a fixed problem instance with $(\theta_0, \theta_1, T) = (0.4, 0.6, 800)$ in Figure 6 (Appendix E). The predicted inverse quadratic dependence of the detection delay is verified in Figure 3.

5.3. Detecting Changes in CDFs

Staying with real-valued observations (or more generally, observations on totally ordered spaces), we now consider a more general question of detecting whether there have been any changes in distribution generating the observations. Since real valued random-variables are completely characterized by their cumulative distribution functions (CDFs), this task can be framed in terms of detecting changes in the CDFs. More formally, we assume that we are given a stream of observations, $X_1, X_2, \ldots$, that are drawn according to a distribution $\theta_0 = F_0$ for $t < T$, and according to a distribution $\theta_1 = F_1$ for $t \geq T$. Thus, in this case, Θ is the infinite-dimensional space of all feasible CDFs on $\mathbb{R}$, and we endow it with the Kolmogorov-Smirnov (KS) metric, d_{KS}, defined as $d_{\mathrm{KS}}(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|$.

To instantiate our SCD scheme, we will employ the following level-α confidence sequence for the CDF in terms of the KS metric, recently derived by Howard and Ramdas (2022): $C_t = \{\theta \in \Theta : d_{\mathrm{KS}}(\theta, \widehat{\theta}_t) \leq w_t/2\}$, where $w_t = 1.7\sqrt{\log\log(et) + 0.8\log(1612/\alpha)/t}$. As a consequence of the general result in Theorem 13, we can obtain the following performance guarantee for this scheme.

Corollary 22. *Suppose, for some $T < \infty$, the observations $X_1, \ldots, X_T$ are drawn from a known distribution F_0, while for $t \geq 1$, the observations $X_{T+1}, X_{T+2}, \ldots$ are drawn from an unknown F_1. Then, our SCD scheme instantiated with the CS stated above satisfies (with $\Delta := d_{KS}(F_1, F_0)$):*

$$\mathbb{E}_T[(\tau - T)^+] = \mathcal{O}\left(\frac{\log\log(1/\Delta) + \log(1/\alpha)}{\Delta^2}\right).$$

Empirical verification. We test the performance of our proposed scheme for t-distributions with 3 degrees of freedom. In Figure 7 in Appendix E, we show the performance of our proposed scheme for a fixed problem instance where the pre- and post-change CDFs satisfy $\Delta = d_{\mathrm{KS}}(F_0, F_1) \approx 0.4$. The variation of the average detection delay with changing values of Δ is plotted in Figure 3, and it displays the expected inverse quadratic dependence.

5.4. Detecting Change in Homogeneity of Two Streams

Suppose we have a stream of observations in a product space $\mathcal{X} = \mathcal{U} \times \mathcal{U}$, and the parameter set consists of all product distributions on $\mathcal{U} \times \mathcal{U}$; that is, $\Theta = \{P \times Q : P, Q \in \mathcal{P}(\mathcal{U})\}$. Prior to the changepoint, we assume that the observations $X_1, X_2, \ldots, X_T$, with $X_t = (U_t, V_t) \in \mathcal{U} \times \mathcal{U}$, are drawn from $\theta_0 = P_U \times P_V$; while the post-change observations $X_{T+1}, X_{T+2}, \ldots$ are assumed to be drawn from some other product distribution $Q_U \times Q_V$. Given some statistical distance measure, $\rho : \mathcal{P}(\mathcal{U}) \times \mathcal{P}(\mathcal{U}) \to \mathbb{R}$, we assume that the $\rho(P_U, P_V) \neq \rho(Q_U, Q_V)$. An interesting special case of this problem, motivated by the two-sample testing problem, is when $P_U = P_V$, and $Q_U \neq Q_V$.

If ρ is a probability metric, we can use it induce a distance metric, d, on the parameter space Θ as follows: $d(\theta_0, \theta_1) = |\rho(P_U, P_V) - \rho(Q_U, Q_V)$, where $\theta_0 = P_U \times P_V$ and $\theta_1 = Q_U \times Q_V$. Then, to obtain a changepoint detection scheme we can employ CSs for the statistical distance ρ. We instantiate this strategy with the kernel-MMD metric (defined in Appendix A) associated with a kernel k, denoted by $d_{\mathrm{MMD}}(\cdot, \cdot)$. We use the following CS derived by Manole and Ramdas (2021) for the kernel-MMD distance between two distributions (assuming that $\sup_{u,u'} k(u, u') \leq 1$): $C_t = \{(P, Q) : d_{\mathrm{MMD}}(\widehat{P}_t, \widehat{Q}_t) -$

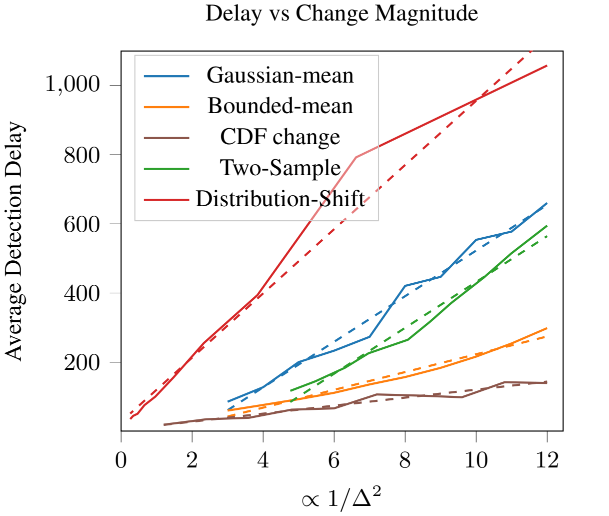

Figure 3. In all the instantiations of the `BCS-Detector` within Section 5, the width of the CS at a $\sqrt{(\log\log(t) + \log(1/\alpha)/t}$ rate. Hence, by Theorem 13, we expect the detection delay to have an inverse quadratic dependence on the magnitude of change (Δ). The figure above verifies this claim empirically. In particular, the solid lines plot the average detection delay computed using 250 trails, against $1/\Delta^2$, where Δ is the change magnitude (i.e., $d(\theta_1, \theta_0)$ for each problem. The dashed lines of the same color represent the best linear fit between the observed detection delay and $1/\Delta^2$. The good agreement between these two lines for each problem validates the prediction of Theorem 13.

$\gamma_t \leq d_{\text{MMD}}(P, Q) \leq d_{\text{MMD}}(\widehat{P}_t, \widehat{Q}_t) + 2\kappa_t\}$ where $\kappa_t = \sqrt{\left(\log((1 \vee \log_2(t))^2\pi^2/6) + \log(4/\alpha)\right)/t}$, and $\gamma_t = \left(4\sqrt{2}/\sqrt{t}\right)\left(1 + \sqrt{\log\left(3.54e(1 \vee \log_2 t)^3\right) + \log(2/\alpha)}\right)$. For this instantiation, Theorem 13 implies the following.

Corollary 23. *With X_t denoting the pair (U_t, V_t) on $\mathcal{X} \times \mathcal{X}$, suppose that $X_1, \ldots, X_T$ are drawn i.i.d. from $\theta_0 = P_U \times P_V$, and $X_{T+1}, X_{T+2}, \ldots$ are drawn i.i.d. from a distribution $\theta_1 = Q_U \times Q_V$. Then, with $\Delta > 0$ denoting $d(\theta_0, \theta_1) = |d_{MMD}(P_U, P_V) - d_{MMD}(Q_U, Q_V)|$, we have the following upper bound on the expected detection delay of our BCS-detector based on the CS described above:*

$$\mathbb{E}\left[(\tau - T)^+ | \mathcal{E}\right] = \mathcal{O}\left(\frac{\log(1/\alpha) + \log\log(1/\Delta)}{\Delta^2}\right),$$

where $\mathcal{E}$ denotes the 'good' event $\{d_{MMD}(P_U, P_V) \in C_t : t \leq T-1\}$ associated with the forward CS $\{C_t : t \geq 1\}$.

Remark 24. Consider the special case mentioned earlier, where $P_U = P_V = Q_U = P$ for some distribution P, and $Q_Y = Q \neq P$ for some other distribution. Furthermore, assume that it is known that prior to changepoint $P_U = P_V$ (that is, the event $\mathcal{E}$ is a probability one event). Then, the above result in this case implies an upper bound on the expected detection delay of $\mathcal{O}\left(\frac{\log(1/\alpha)+\log\log(1/\Delta)}{\Delta^2}\right)$, with $\Delta = d_{\text{MMD}}(P, Q)$. This matches existing results, such as the kernel CuSum scheme of Wei and Xie (2022).

Remark 25. We focused on the case of the kernel-MMD metric mainly due to its generality (it is applicable to distributions over arbitrary spaces on which positive definite kernels can be defined). However, the same ideas are applicable to any statistical distance measure that is convex in its arguments, by using the reverse submartingale based confidence sequence construction of Manole and Ramdas (2021). This family includes all popular statistical distances, such as Wasserstein metrics, f-divergences and general integral probability metrics.

Remark 26. The overall computational cost of our scheme is $\mathcal{O}(\tau^3)$, as our scheme involves constructing a new backward CS, with $\mathcal{O}(t^2)$ cost, every round. In practice, this complexity can be reduced, either by using linear or block-MMD statistics, and/or by computing a new backward CS less frequently (instead of doing so every round).

Empirical Verification. We study the performance of our scheme on a stream of paired multivariate Gaussian observations in $p = 5$ dimensions. The pre-change distributions, P_U and P_V both have zero mean and identity covariance; while for the post change distributions we have $Q_U = P_U$, and Q_Y has a mean $\delta\mathbf{1}$, and a diagonal covariance matrix with randomly chosen values. In Figure 8 in Appendix E, we plot the performance of our changepoint detection scheme for a fixed problem instance with $\Delta = d_{\text{MMD}}(Q_U, Q_V) \approx 0.33$, and $T = 800$, while the inverse quadratic dependence of the average detection delay with Δ is verified in Figure 3.

5.5. Detecting Harmful Distribution Shifts

As a final application, consider the task of detecting 'harmful' changes between train and test distributions of a machine learning (ML) model. Following Podkopaev and Ramdas (2021), we are interested in detecting only those distribution changes that lead to a sufficiently large increase in the risk (i.e., expected loss) of the trained ML model.

Formally, suppose a machine learning model, denoted by h, is trained on a dataset drawn i.i.d. from a *source* distribution P_S taking values on some space $\mathcal{X}$. For some bounded loss function, ϕ, we let θ_0 denote the expected training loss of this model; $\theta_0 = \mathbb{E}_{P_S}[\phi(X, h)]$. Next, we assume that the model h is deployed on a stream of test data, denoted by $X_1, X_2, \ldots$, drawn from the source (or training) distribution P_S for $t < T$; and from some other distribution $P_T \neq P_S$ with $P_T \neq P_S$. Our goal is to detect post-change distributions P_T that are 'harmful' to the trained model; that is, they result in an increase in expected loss: $\theta_1 := \mathbb{E}_{P_T}[\phi(h, X)] > \theta_0$ (see Figure 10 in Appendix E).

For bounded loss functions ϕ, this problem fits into the nonparametric change of mean detection framework of Section 5.2. Since we are only interested in one-sided changes, we can modify the strategy of Section 5.2 to use only upper CS in the forward direction, and lower CSs in the backward direction. As in Proposition 19, for this strategy, we can show that the expected detection delay of the scheme will depend inversely on how 'harmful' the target distribution is (i.e., the gap $\theta_1 - \theta_0$).

Empirical Verification. To illustrate the ideas discussed above, we consider a simple binary classification problem with linear classifiers and 2-dimensional features (see Appendix E for details). We plot the performance of our scheme on a specific problem with $\Delta \approx 0.16$ in Figure 9 in Appendix E, and also verify the inverse quadratic dependence of average detection delay on Δ in Figure 3.

5.6. Other change detection tasks

We have illustrated the generality of our `BCS-Detector` strategy by instantiating it for five different scenarios in this section. For simplicity, we focused mainly on univariate observations (with the exception of Section 5.4). However, we note that the same ideas used in the previous instantiations also carry over easily to more general observations, or under additional robustness or privacy constraints. We list some such examples here, without going into the details of analysis or practical implementations:
(i) Exponential family. In this case, the observations $X_1, X_2, \ldots$ lie in $\mathcal{X} = \mathbb{R}^p$, and the pre- and post-change distributions (P_{θ_0} and P_{θ_1}) are chosen from a finite-dimensional exponential family with $\Theta = \mathbb{R}^m$ for some $m < \infty$. Here, we can use the `BCS-Detector` with the CSs derived by Chowdhury et al. (2022).
(ii) Covariance matrix. Again, we assume that $\mathcal{X} = \mathbb{R}^p$, but now we assume that at the change point T, the covariance matrix of the observations changes from $\theta_0 \in \mathbb{R}^{p\times p}$ to some $\theta_1 \neq \theta_0$. For this problem, we can instantiate the `BCS-Detector` with the CS for covariance matrices derived by Howard et al. (2021, § 4.3).
(iii) Nonparametric regression. Suppose $U_1, U_2, \ldots$ denote i.i.d. uniform draws from $\mathcal{U} = [0,1]^p$, for some $p \geq 1$. Let Θ denote an RKHS (with kernel k) of functions from $\mathcal{U}$ to $\mathbb{R}$. Then, for any $\theta \in \Theta$, define the random variable $Y_t \equiv Y_t(\theta) = \theta(U_t) + \eta_t$, where $\{\eta_t : t \geq 1\}$ are an i.i.d. sequence of 1-sub-Gaussian noise. Clearly, the joint distribution of (U_t, Y_t) is parametrized by θ. Consider the SCD problem, where $\theta = \theta_0$ prior to changepoint T, and $\theta = \theta_1$ after that, with $\|\theta_0 - \theta_1\|_k > 0$. Our `BCS-Detector` strategy is easily applicable to this scenario, with an infinite-dimensional index set Θ, using the CS constructed by Chowdhury and Gopalan (2017).
(iv) Robust SCD. An interesting variant of the SCD problem involves detecting changepoints under adversarial contamination (Li and Yu, 2021). Our `BCS-Detector` strategy readily extends to such scenarios, by exploiting recent robust confidence sequence constructions, such as those by Wang and Ramdas (2023).
(v) Private SCD. Privacy is an important concern in many applications, especially involving personal data, and is often ensured by revealing only randomized versions of the actual data to the analyst. This adds another layer of complexity to the usual SCD task (Cummings et al., 2018). However, our `BCS-Detector` framework can easily handle this, building upon the recent advances in private CS construction (Waudby-Smith et al., 2022).

6. Conclusion

We proposed a general strategy (`BCS-Detector`) for designing sequential changepoint detection (SCD) schemes by carefully combining confidence sequences (CSs), and backward CSs — a novel variant of CSs, that we introduced in this paper. Under very mild, and natural requirements on the CSs, we showed that `BCS-Detector` provides tight control over the ARL and the detection delay. Leveraging the recent progress in constructing CSs, we instantiated our strategy for a wide range of SCD problems (both parametric, and nonparametric), and empirically verified the theoretical claims via some small-scale numerical experiments.

Our work opens up several directions for future work:
(i) Constructing a new backward CS in every round can be computationally costly, and in most cases results in an overall quadratic (or even cubic) complexity. An interesting direction to pursue is to investigate if we can achieve the same performance by updating the backward CS fewer times. **(ii)** In Remark 16 we defined estimators of the changepoint (T), and the change magnitude (Δ), which performed well empirically as shown in Figure 2 and Appendix E. Establishing theoretical guarantees for them is another interesting question for future work.

References

M. Baron, V. Antonov, C. Huber, M. Nikulin, and V. Polischook. Early detection of epidemics as a sequential change-point problem. *Longevity, aging and degradation models in reliability, public health, medicine and biology, LAD*, pages 7–9, 2004.

H. Chen. Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407, 2019.

J. Chen, A. Yiğiter, and K.-C. Chang. A Bayesian approach to inference about a change point model with application to DNA copy number experimental data. *Journal of Applied Statistics*, 38(9):1899–1913, 2011.

Y. C. Chen, T. Banerjee, A. D. Dominguez-Garcia, and V. V. Veeravalli. Quickest line outage detection and identification. *IEEE Transactions on Power Systems*, 31(1): 749–758, 2015.

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

S. R. Chowdhury, P. Saux, O.-A. Maillard, and A. Gopalan. Bregman deviations of generic exponential families. *arXiv preprint arXiv:2201.07306*, 2022.

R. Cummings, S. Krehbiel, Y. Mei, R. Tuo, and W. Zhang. Differentially private change-point detection. *Advances in neural information processing systems*, 31, 2018.

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

T. Flynn and S. Yoo. Change detection with the kernel cumulative sum algorithm. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6092–6099. IEEE, 2019.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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

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

T. L. Lai and H. Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

T. S. Lau, W. P. Tay, and V. V. Veeravalli. A binning approach to quickest change detection with unknown post-change distribution. *IEEE Transactions on Signal Processing*, 67(3):609–621, 2018.

M. Li and Y. Yu. Adversarially robust change point detection. *Advances in Neural Information Processing Systems*, 34:22955–22967, 2021.

S. Li, Y. Xie, H. Dai, and L. Song. Scan B-statistic for kernel change-point detection. *Sequential Analysis*, 38 (4):503–544, 2019.

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

O.-A. Maillard. Mathematics of statistical sequential decision making. *Habilitation a Diriger des Recherches*, 2019a.

O.-A. Maillard. Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In *Algorithmic Learning Theory*, pages 610–632. PMLR, 2019b.

T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *arXiv preprint arXiv:2103.09267*, 2021.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

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

A. Podkopaev and A. Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2021.

M. Pollak and D. Siegmund. Sequential detection of a change in a normal mean when the initial value is unknown. *The Annals of Statistics*, 19(1):394–416, 1991.

N. Puchkin and V. Shcherbakova. A contrastive approach to online change point detection. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5686–5713. PMLR, 25–27 Apr 2023.

J. Sharpnack, A. Singh, and A. Rinaldo. Changepoint detection over graphs with the spectral scan statistic. In *Artificial Intelligence and Statistics*, pages 545–553. PMLR, 2013.

J. J. Shen and N. R. Zhang. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *The Annals of Applied Statistics*, 6(2):476–496, 2012.

W. A. Shewhart. 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.

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

J. Shin, A. Ramdas, and A. Rinaldo. E-detectors: a nonparametric framework for online changepoint detection. *arXiv preprint arXiv:2203.03532*, 2022.

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

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

V. V. Veeravalli and T. Banerjee. Quickest change detection. In *Academic press library in signal processing*, volume 3, pages 209–255. Elsevier, 2014.

H. Wang and A. Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *arXiv preprint arXiv:2202.01250*, 2022.

H. Wang and A. Ramdas. Huber-robust confidence sequences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9662–9679. PMLR, 25–27 Apr 2023.

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

I. Waudby-Smith, Z. S. Wu, and A. Ramdas. Locally private nonparametric confidence intervals and sequences. *arXiv preprint arXiv:2202.08728*, 2022.

S. Wei and Y. Xie. Online kernel CUSUM for change-point detection. *arXiv preprint arXiv:2211.15070*, 2022.

L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

X. Yu, M. Baron, and P. K. Choudhary. Change-point detection in binomial thinning processes, with applications in epidemiology. *Sequential Analysis*, 32(3):350–367, 2013.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A nonparametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26, 2013.

A. Additional Background

CS for non-i.i.d. observations. As mentioned in Remark 4, we can also define CSs for non-i.i.d. observations, as follows.

Definition 27. Suppose $X_1, X_2, \ldots,$ denote an independent stream of observations on $\mathcal{X}$, with $X_t \sim P_{\vartheta_t}$ for $\vartheta_t \in \Theta$. Assuming Θ is a vector space, we say that $\{C_t \subset \Theta : t \geq 1\}$ is a level $(1-\alpha)$ CS for the running average of parameters if $\mathbb{P}\left(\{\forall t \geq 1 : (1/t)\sum i = 1^t \vartheta_t \in C_t\}\right) \geq 1 - \alpha$.

The same definition of pointwise width introduced in Definition 5 is still applicable to the above CS. However, for defining the uniform width, we need to take the supremum over the sequence of parameters, instead of a fixed parameter (θ). More specifically, the pointwise and uniform widths for the above CS is defined as

$$w(t, \theta_1^t, \alpha) = \sup_{\theta', \theta'' \in C_t} d(\theta', \theta''), \quad \text{and} \quad w(t, \Theta, \alpha) = \sup_{\theta_1^t \in \Theta^{\otimes t}} w(t, \theta_1^t, \alpha).$$

We show a simple illustration of the CS introduced in Example 6 with time varying parameters in Figure 4.

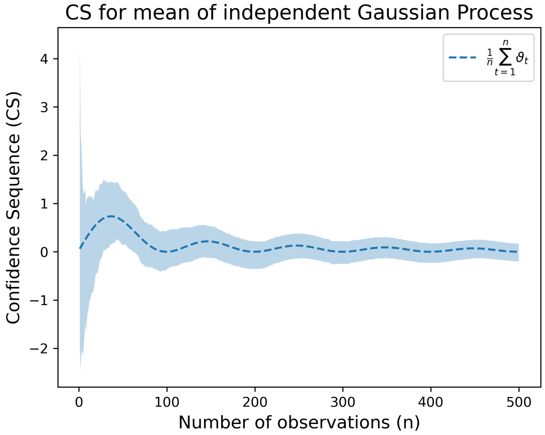

Figure 4. An example of the CS introduced in Example 6 for the running (conditional) mean of independent Gaussian processes with a time-varying mean function, variance fixed at 1.

Implementing backward CSs. If we know how to construct forward CSs, we can use that directly to construct backward CSs in the following steps:

- At the end of round n, we have observed $X_1, \ldots, X_n$. Introduce the time-reversed version of the observations $Y_s = X_{n+1-s}$ for $s \in [n] := \{1, \ldots, n\}$.
- Construct a new level-$(1-\alpha)$ CS using $\{Y_s : s \in [n]\}$, denoted by $\{\bar{B}_s^{(n)} : s \in [n]\}$. Note that $\bar{B}_s^{(n)}$ is $\sigma(Y_1, \ldots, Y_s) = \sigma(X_n, X_{n-1}, \ldots, X_{n-s+1})$ measurable for all $s \in [n]$.
- Finally, we again reverse the index of the CS, to obtain $\{B_t^{(n)} : t \in [n]\}$, where $B_t^{(n)} = \bar{B}_{n+1-s}^{(n)}$. Note that by virtue of being a CS, we have $\mathbb{P}_\infty\left(\forall t \in [n] : \theta_0 \in B_t^{(n)}\right) \geq 1 - \alpha$. The superscript (n) serves as a reminder that there is only one forward CS, but there is a different backward CS constructed afresh at each time n.

The kernel-MMD metric. In Section 5.4, we constructed a scheme for detecting changes the pairwise kernel-MMD distance between the distributions generating a stream of paired observations. Here, we recall the its definition.

We assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denotes a uniformly-bounded positive-definite kernel, and let $\mathcal{H}_k$ denote the reproducing kernel Hilbert space (RKHS) associated with k.

Definition 28. Given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the kernel-MMD distance between two-distributions P and Q on $\mathcal{X}$ is defined as

$$d_{\text{MMD}}(P, Q) = \sup_{g \in \mathcal{H}_k : \|g\|_k \leq 1} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(Y)].$$

The kernel-MMD distance defined above is an instance of a class of statistical distances called *integral probability metrics (IPMs)*. For a class of kernels, called characteristic kernels, it is known that d_{MMD} is a distance metric on the space of probability distributions.

B. Proof of Proposition 8

Proof of the bound on probability of false alarm. Recall that the stopping time is defined as the first time, τ, at which we have the condition $\cap_{t=1}^{\tau} C_t = \emptyset$. Now, consider the 'good' event of the CS under the null: $\mathcal{E} = \cap_{t=1}^{\infty} \{\theta_0 \in C_t\}$, which satisfies $\mathbb{P}_\infty(\mathcal{E}) \geq 1 - \alpha$ by definition. Hence, under this event $\{\theta_0\} \subset \cap_{t=1}^{\infty} \neq \emptyset$, which in turn implies that $\mathbb{P}_\infty(\tau = \infty) \geq 1 - \alpha$, as required.

Proof of the bound on detection delay. Let $w(t) \equiv w(t, \Theta, \alpha)$ denote the width of the confidence C_t after t observations. By assumption, T is large enough to ensure that under the 'good' event $\mathcal{E} = \cap_{t=1}^{\infty} \left\{ \frac{1}{t} \vartheta_t \in C_t \right\}$, we have that $\theta_1 \notin C_T$ at the changepoint. Note that in the definition of $\mathcal{E}$, we have $\vartheta_t = \theta_0 \mathbf{1}_{t \leq T} + \theta_1 \mathbf{1}_{t > T}$.

Under the event $\mathcal{E}$, we know that $\theta_0 \in C_T$. For any $t > T$, introduce the terms $\lambda_t = T/t$ and $\bar{\lambda}_t = 1 - \lambda_t$. Then, by definition of confidence sequences, we have $\lambda_t \theta_0 + \bar{\lambda}_t \theta_1 \in C_t$ for all $t > T$ under the event $\mathcal{E}$. The width of the set C_t at $t > T$ is no larger than $w(t) \equiv w(t, \Theta, \alpha)$. Hence, a sufficient condition for stopping prior to $t > T$ is if the sum of the widths of C_t and C_T, that is $w(t) + w(T)$, is smaller than $d\left(\theta_0, \lambda_t \theta_0 + \bar{\lambda}_t \theta_1\right)$.

Informal calculations for Remark 9. As mentioned in Remark 9, in many cases, the stopping time τ satisfies: $\tau \approx \min\{t \geq T : 1/\sqrt{t} + 1/\sqrt{T} \leq ((t-T)/t)\Delta\}$, where $\Delta = d(\theta_0, \theta_1)$. We now consider the behavior of the delay, $\tau - T$, in two different regimes of the changepoint T.

First we consider the case where T is 'small'; that is $T \approx 1/\Delta^2$. For concreteness, assume that $T = 9/\Delta^2$. Then, it is easy to check that with $\tau \leq 4T$, since for $t = 4\tau$

$$\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{T}} = \frac{\Delta}{4} + \frac{\Delta}{8} \leq \frac{3\Delta}{4} = \frac{t-T}{t}\Delta.$$

Next, we consider the case where Δ is fixed, but $T \to \infty$. In this case, we have $\tau - T = \mathcal{O}\left(\sqrt{T}\right)$. To see this, consider $t = T + u$, with $u = o(T)$. Then, we have

$$\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{T}} \approx \frac{2}{\sqrt{T}}, \quad \text{and} \quad \frac{t-T}{t}\Delta \approx \frac{u}{T}\Delta.$$

Thus, this implies that the appropriate order of growth of the detection delay is $u = \mathcal{O}(\sqrt{T}/\Delta)$, as $T \to \infty$ with Δ fixed.

C. Proof of Theorem 13

Proof of the ARL control. To prove this result, note that for us to stop under the null at some time τ, either the forward or the backward CS (or both) must be miscovering and failing to contain θ_0. In other words,

$$\{\tau = N, T = \infty\} \implies \{C_N \text{ miscovers}\} \vee \{B_1^{(N)} \text{ miscovers}\}.$$

$$\{\tau \leq N, T = \infty\} \implies \{(C_t)_{t=1}^N \text{ miscovers}\} \vee \bigcup_{t=1}^{N} \{B_1^{(t)} \text{ miscovers}\}.$$

In both the above implications, we have used the property of nested CSs to define the miscovering events. Now, using the fact that the forward CS and each backward CS is a $(1-\alpha)$-CS, we have by a simple union bound that for any fixed time N,

$$\text{Pr}_\infty(\tau \leq N) \leq \min\{(N+1)\alpha, 1\}.$$

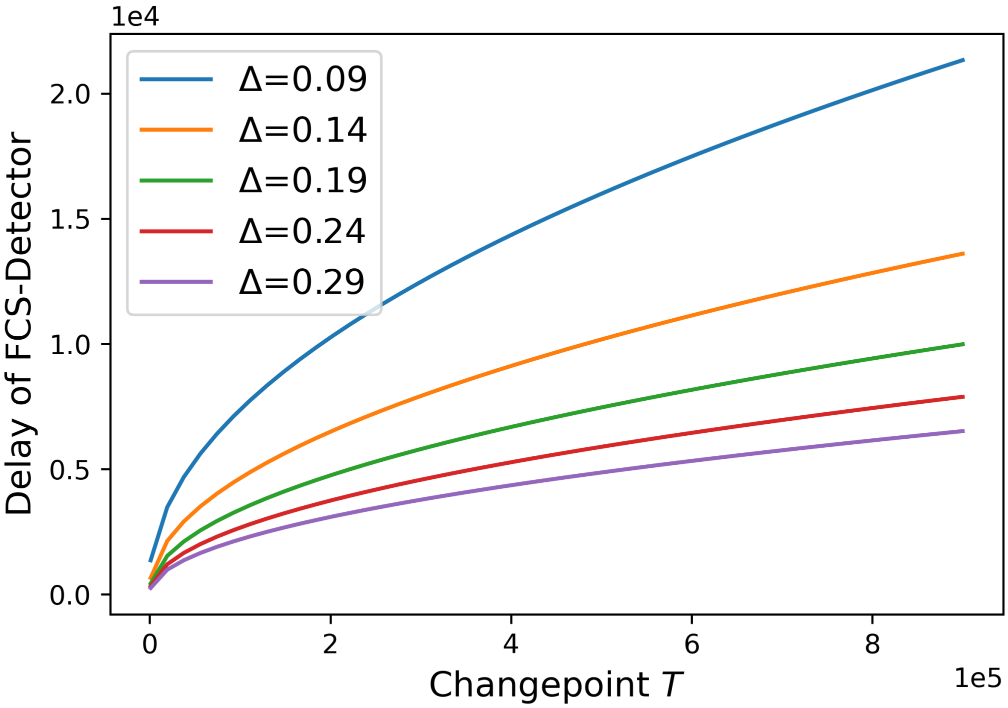

Figure 5. The plots show the variation of the delay of SCD scheme introduced in Section 3, under the conditions of Remark 9. When the changepoint $T \approx 1/\Delta^2$, the delay has a linear dependence on T, while in the regime where $T \to \infty$ with Δ fixed, the delay behaves $\approx \sqrt{T}$.

Rephrasing, we have $\mathbb{P}_\infty(\tau > N) \geq (1 - \alpha(N+1)) \vee 0$, and thus

$$\begin{aligned}
\mathbb{E}_\infty \tau &= \sum_{N=1}^{\infty} \text{Pr}_\infty(\tau > N) \geq \sum_{N=1}^{1/\alpha - 1} (1 - \alpha - \alpha N) \\
&= (1/\alpha - 1)(1 - \alpha) - \alpha \sum_{N=1}^{1/\alpha - 1} N \\
&= 1/\alpha - 2 + \alpha - \alpha \frac{(1/\alpha - 1)(1/\alpha)}{2} \\
&= \frac{1}{2\alpha} - 3/2 + \alpha.
\end{aligned}$$

As we alluded to earlier in Remark 1, we did not need to know θ_0 or θ_1 or the ‘direction’ of the changepoint. The aforementioned argument goes through for any indexed family of distributions. Furthermore, the above guarantee does not require the width of the CSs (forward of backward) to decay to zero; it only uses the coverage property of CSs.

Proof of the detection delay bound. We next obtain the upper bound on the average detection delay conditioned on the event $\mathcal{E} := \{\theta_0 \in C_t : 1 \leq t \leq T\}$, stated in Theorem 13. Since, $(\tau - T)^+ = \max\{0, \tau - T\}$ is a non-negative random variable, we have

$$\begin{aligned}\mathbb{E}_T[(\tau - T)^+|\mathcal{E}] &= \sum_{t=0}^{\infty} \mathbb{P}_T\left((\tau - T)^+ \geq t|\mathcal{E}\right) \\ &\leq \sum_{t=0}^{t'-1} 1 + \sum_{t \geq t'} \mathbb{P}_T\left((\tau - T)^+ \geq t'|\mathcal{E}\right) \\ &= \sum_{t=0}^{t'-1} 1 + \sum_{t \geq t'} \mathbb{P}_T\left(\tau > T + t'|\mathcal{E}\right) \\ &= t' + \sum_{t \geq t'} \mathbb{P}_T\left(\tau > T + t'|\mathcal{E}\right), \quad \text{for any } t' \geq 1. \qquad (1)\end{aligned}$$

Recall that $u_0 \equiv u_0(\theta_0, \theta_1, T) \coloneqq \min\{t - T : w(T, \theta_0, \alpha) + w(t - T, \theta_1, \alpha) < d(\theta_0, \theta_1)\}$ was introduced in the statement of Theorem 13, and it represents an upper bound on the smallest time after T at which the backward CS must stop intersecting with the forward CS (assuming none of the CSs miscover). For any integer $i \geq 1$, consider the event $\{\tau > T + iu_0\}$, and note that it satisfies the following inclusion:

$$\begin{aligned}\{\tau > T + iu_0\} \cap \mathcal{E} &\subset \left(\cap_{j=1}^{i} \left\{C_T \cap B_T^{(T+ju_0)} \neq \emptyset\right\}\right) \cap \mathcal{E} \\ &\subset \left(\cap_{j=1}^{i} \left\{C_T \cap B_{T+(j-1)u_0}^{(T+ju_0)} \neq \emptyset\right\}\right) \cap \mathcal{E} \qquad (2) \\ &\subset \left(\cap_{j=1}^{i} \left\{\theta_1 \not\in B_{T+(j-1)u_0}^{(T+ju_0)}\right\}\right) \cap \mathcal{E}. \qquad (3)\end{aligned}$$

In the display above, (2) uses the fact that $B_s^{(n)} \subset B_{s'}^{(n)}$ for any $s' > s$. The inclusion (3) is the crucial observation for our proof. It relies on the fact that if for some $j \geq 1$, the BCS $\{B_s^{(T+ju_0)} s \in [T + ju_0]\}$ does not miscover, then the $B_s^{(T+ju_0)}$ contains θ_1 for all $s \in \{T, \ldots, T + ju_0\}$, and in particular at $s = T + (j-1)u_0$. Also note that the diameter of C_T is smaller than $w(T, \theta_0, \alpha)$, and that of $B_{T+(j-1)u_0}$ is smaller than $w(u_0, \theta_1, \alpha)$. Finally, the definition of u_0 implies that $w(T, \theta_0, \alpha) + w(u_0, \theta_1, \alpha) < d(\theta_0, \theta_1)$ — implying that C_T and $B_{T+(j-1)u_0}^{(T+ju_0)}$ are contained in two disjoint balls, and hence are disjoint. Thus, if $C_T \cap B_{T+(j-1)u_0}^{(T+ju_0)} \neq \emptyset$, then the backward CS at $T + ju_0$ must miscover.

Next, we make the following two observations:

(I) For $j \neq j'$, the events $E_j \coloneqq \{\theta_1 \not\in B_{T+(j-1)u_0}^{(T+ju_0)}\}$ and $E_{j'} \coloneqq \{\theta_1 \not\in B_{T+(j'-1)u_0}^{(T+j'u_0)}\}$ are independent.

(II) For all $1 \leq j \leq i$, the event E_j (introduced above) is independent of $\mathcal{E}$.

The statement **(I)** follows from the observation that the event E_j lies in the sigma-algebra $\sigma\left(\{X_k : T + (j-1)u_0 \leq k < T + ju_0\}\right)$, while $E_{j'}$ lies in $\sigma\left(\{X_k : T + (j'-1)u_0 \leq k < T + j'u_0\}\right)$; which are independent. Similarly, the second statement **(II)** uses the fact that $\mathcal{E}$ lies in $\sigma\left(\{X_k : 1 \leq k < T\}\right)$, which is independent of $\sigma\left(\{X_k : T + (j-1)u_0 \leq k < T + ju_0\}\right)$ for all $1 \leq j \leq i$.

Based on the above observations, we conclude that

$$\begin{aligned}\mathbb{P}_T\left(\tau > T + iu_0|\mathcal{E}\right) &= \mathbb{P}_T\left(\{\tau > T + iu_0\} \cap \mathcal{E}\right)/\mathbb{P}_T(\mathcal{E}) \\ &\leq \mathbb{P}_T\left(\left(\cap_{j=1}^{i}\{\theta_1 \not\in B_{T+(j-1)u_0}^{(T+ju_0)}\}\right) \cap \mathcal{E}\right)/\mathbb{P}_T(\mathcal{E}) \\ &= \mathbb{P}_T\left(\cap_{j=1}^{i}\{\theta_1 \not\in B_{T+(j-1)u_0}^{(T+ju_0)}\}\right) \qquad (4) \\ &\leq \alpha^i, \quad \text{for all } i \geq 1. \qquad (5)\end{aligned}$$

In the above display, (4) follows from **(II)**, and (5) uses **(I)**. Now, we return to (1), and set t' to $i_0 \times u_0$ for some integer i_0

to be specified later. We then note that

$$\begin{aligned}
\mathbb{E}_T[(\tau - T)^+|\mathcal{E}] &\leq i_0 u_0 + \sum_{i=i_0}^{\infty} \sum_{t=iu_0}^{(i+1)u_0-1} \mathbb{P}_T(\tau > T + t|\mathcal{E}) \\
&\leq i_0 u_0 + \sum_{i=i_0}^{\infty} u_0 \mathbb{P}_T\left(\tau > T + iu_0|\mathcal{E}\right) \\
&\leq i_0 u_0 + u_0 \frac{\alpha^{i_0}}{1-\alpha} = u_0 \left(i_0 + \frac{\alpha^{i_0}}{1-\alpha} \right). \qquad (6)
\end{aligned}$$

The inequality in (6) uses the fact that $\mathbb{P}_T(\tau > T + iu_0) \leq \alpha^i$ derived in (5). The final result, as stated in Theorem 13, then follows by selecting $i_0 = \lceil \log(1/1-\alpha)/\log(1/\alpha) \rceil$. Note that when $\alpha < 0.5$, we have $i_0 = 1$.

D. Deferred proofs from Section 5

D.1. Proofs of Corollary 17, Corollary 22, Corollary 23

All these three results can be obtained as a direct consequence of the following proposition.

Proposition 29. *For some $\Delta > 0$, define the time t_0 as*

$$t_0 = \min \left\{ t \geq 1 : c\sqrt{\frac{\log\log t + \log(1/\alpha)}{t}} \leq \frac{\Delta}{2} \right\},$$

where $c > 0$ is some constant. Then, we have

$$t_0 = \mathcal{O}\left(c^2 \, \frac{\log\log(c/\Delta) + \log(1/\alpha)}{\Delta^2} \right).$$

Proof. Without loss of generality, we assume that $c = 1$; or equivalently, we can replace Δ with Δ/c. Now, note that we can upper bound $t_0 \leq t_1 + t_2$, where

$$t_1 = \min\{t \geq 1 : \sqrt{\log\log(t)/t} \leq \Delta/4\}, \quad \text{and} \quad t_2 = \min\{t \geq 1 : \sqrt{\log(1/\alpha)/t} \leq \Delta/4\}.$$

By a simple calculation, we can obtain $t_2 = \mathcal{O}\left(\log(1/\alpha)/\Delta^2\right)$. Hence to complete the proof, we will show that $t_1 = \mathcal{O}\left(\log\log(1/\Delta)/\Delta^2\right)$. We proceed in two steps: (i) first we show that $t_1 \leq 32/\Delta^3$, and (ii) using this, we refine the result to show that $t_1 = \mathcal{O}\left(\log\log(1/\Delta)/\Delta^2\right)$.

Let $t_3 = 32/\Delta^3$ for $\Delta \leq 1$. Then, observe that

$$\left(\frac{4}{\Delta} \sqrt{\frac{\log\log(t_3)}{t_3}} \right)^2 = \frac{16}{\Delta^2} \times \frac{\Delta^3 \, \log\log(32/\Delta^3)}{32} = \frac{1}{2} \frac{\log\log(32/\Delta^2)}{1/\Delta} \leq 0.63 < 1.$$

The last inequality, along with the definition of t_1 implies that $t_1 \leq t_3$.

Hence, $\log\log(t_1) \leq \log\log(t_3) = \log\log(32/\Delta^3) \leq 2.35 + \log\log(1/\Delta)$. Thus, we have

$$\frac{\log\log(t_1)}{t_1} \leq \frac{2.35 + \log\log(1/\Delta)}{t_1},$$

which implies that $t_1 \leq \frac{16}{\Delta^2}\left(2.35 + \log\log(1/\Delta)\right) = \mathcal{O}\left(\log\log(1/\Delta)/\Delta^2\right)$. Combining this bound on t_1, with the previously obtained upper bound on t_2; and using the fact that $t_0 \leq t_1 + t_2$, we get the required result. □

D.2. Proof of Proposition 19

To prove the variance adaptive bound on the expected detection delay, we first show the following result for the width of the CS derived by Waudby-Smith and Ramdas (2023).

Proposition 30. *We can modify the backward CS used instantiating the* `BCS-Detector` *in Section 5.2, to obtian a level $(1-2\alpha)$-backward CS (for every $n \geq T$), denoted by $\{B_t^{(n)} : 1 \leq t \leq n\}$, such that the width of $B_t^{(n)}$ satisfies*

$$w(t, \theta_1, \alpha) = \mathcal{O}\left(\frac{\sigma_1}{\sqrt{t}} \left(\sqrt{\log(1/\alpha)} + \sqrt{\log t} \right) \right), \quad \textit{for all } T \leq t \leq n.$$

Remark 31. The main benefit of this result is that it characterizes the width of the backward CS (for $n \geq T$) explicitly in terms of the standard deviation (σ_1) of the post-change distribution P_{θ_1}, unlike the original CS described in Section 5.2, whose width depends on empirical estimates of σ_1. As a consequence of this result, we obtain Proposition 19 by first appealing to Corollary 14, and then repeating the calculations used to obtain Proposition 29.

Proof. Since we are only interested in characterizing the order with which the width of the CS decays (and not the exact constants), we will not track the constants in our argument for this proof. In particular, we will use $A \lesssim B$ to indicate that by $A/B = \mathcal{O}(1)$, and $A \approx B$ to indicate that $A \lesssim B$ and $B \lesssim A$.

We proceed in the following steps:

- First, we show that we can construct a level-$(1-\alpha)$ confidence sequence for the empirical variance based on samples from the post-change distribution. In particular, let $\widehat{\sigma}_n^2 = (\frac{1}{4} + \sum_{t=1}^n (X_t - \widehat{\mu}_t)^2)/(n+1)$, with $X_1, X_2, \ldots \sim P_{\theta_1}$ i.i.d.. Then, we have the following:

$$\mathbb{P}\left(\mathcal{E}_1\right) \geq 1 - \alpha, \quad \text{where} \quad \mathcal{E}_1 \coloneqq \cap_{n \geq 1} \left\{ |\widehat{\sigma}_n^2 - \sigma_1^2| = \mathcal{O}\left(\sqrt{\frac{\log\log n + \log(1/\alpha)}{n}} \right) \right\}. \tag{7}$$

  Thus, using this event, for any $n \geq T$, we can modify the backward CS to get a level-$(1-2\alpha)$ Backward CS, in which $\widehat{\sigma}$ is replaced by σ_1 (plus a small approximation error term) for $T \leq t \leq n$. In the next two steps, we characterize the width of these level-$(1-2\alpha)$ CSs.

- Next, we show that under the event $\mathcal{E}_1$, we have

$$\sum_{i=1}^n v_i \psi_E(\lambda_i) \approx \log(\sigma_1^2 n). \tag{8}$$

- Under the same event, we then show that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n \lambda_i \approx \sqrt{\frac{2\log(2/\alpha)}{\sigma_1^2}}. \tag{9}$$

Combining these results, we get that the width of the CS is of the order

$$w_n \approx \frac{1}{\sqrt{n}} \frac{\log(2/\alpha) + \sum_{t=1}^n v_t \psi_E(t)}{\frac{1}{\sqrt{n}} \sum_{t=1}^n \lambda_t} \approx \frac{\sigma_1}{\sqrt{n}} \left(\sqrt{\log(2/\alpha)} + \sqrt{\log(n)} \right),$$

as required. Thus, it remains to prove (7), (8), and (9).

Proof of (7). To prove this, we first introduce the usual unbiased estimate of the variance: $\widetilde{\sigma}_n^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{(X_i - X_j)^2}{2}$. Since the observations $X_1, X_2, \ldots \sim P_{\theta_1}$ are bounded, and lie in $[0,1]$, it is easy to verify that $|\widehat{\sigma}_n^2 - \widetilde{\sigma}_n^2| = \mathcal{O}(1/n)$, and hence $\widehat{\sigma}_n^2 \approx \widetilde{\sigma}_n^2$.

Since, $\widetilde{\sigma}_n^2$ is an instance of a U-statistic, and hence the process $\{\widetilde{\sigma}^2 : n \geq 1\}$ is a reverse-martingale, adapted to the exchangeable filtration. Using this fact, along with the boundedness (and hence sub-Gaussianity) of the random variable $\widetilde{\sigma}^2$ for all $n \geq 1$, we can use Manole and Ramdas (2021, Corollary 8), to conclude the time-uniform concentration result: $\mathbb{P}\left(\forall n \geq 1 : |\widetilde{\sigma}^2 - \sigma_1^2| = \mathcal{O}(r_n)\right) \geq 1 - \alpha$, where $r_n = \sqrt{(\log\log n + \log(1/\alpha))/n}$.

Proof of (8). To show this, we recall the fact that $\psi_E(\lambda)/(\lambda^2/8) \to 1$ as $\lambda \to 0$. Hence, we have the following:

$$\begin{aligned}
\sum_{i=1}^n v_i \psi_E(\lambda_i) &= \frac{4}{\log(2/\alpha)} \sum_{i=1}^n (X_i - \widehat{\mu}_i)^2 \psi_E(\lambda_i) \approx \frac{4}{\log(2/\alpha)} \sum_{i=1}^n (X_i - \widehat{\mu}_i)^2 \frac{\lambda_i^2}{8} \\
&\lesssim \frac{1}{\log(2/\alpha)} \sum_{i=1}^n (X_i - \widehat{\mu}_i)^2 \frac{\log(2/\alpha)}{i\widehat{\sigma}_{i-1}^2} \lesssim \frac{1}{\log(2/\alpha)} \sum_{i=1}^n (X_i - \widehat{\mu}_i)^2 \frac{\log(2/\alpha)}{i\widehat{\sigma}_{i-1}^2} \\
&\approx \sum_{i=1}^n \frac{(X_i - \widehat{\mu}_{i-1})^2}{\sum_{j=1}^i (X_j - \widehat{\mu}_{j-1})^2} \\
&\leq \log\left(\sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 \right) \qquad (10) \\
&\approx \log\left(n\widetilde{\sigma}_n^2\right) \approx \log\left(n\sigma_1^2\right).
\end{aligned}$$

In the above display, (10) follows by an application of the following lemma with $f(x) = 1/x$.

Lemma 32 (Orabona (2019), Lemma 4.13). *Let $a_i \geq 0$ for all i, and $f : [0, \infty) \to [0, \infty)$ be an increasing function. Then*

$$\sum_{t=1}^T a_t f(a_0 + \sum_{i=1}^t a_i) \leq \int_{a_0}^{\sum_{t=0}^T a_t} f(x) dx.$$

This concludes the proof of (8).

Proof of (9). We proceed as follows with $r_i = \sqrt{\left(\log\log i + \log(1/\alpha)\right)/i}$

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_i^2\, i}} &\gtrsim \sqrt{\frac{2\log(2/\alpha)}{n}} \sum_{i=1}^n \frac{1}{\sqrt{i\sigma_1^2(1+r_i)}} \\
&\gtrsim \sqrt{\frac{2\log(2/\alpha)}{\sigma_1^2\, n}} \left(\sum_{i=1}^n \frac{1}{\sqrt{i}} - \sum_{i=1}^n \frac{r_i}{\sqrt{i}} \right) \\
&\approx \sqrt{\frac{2\log(2/\alpha)}{\sigma_1^2\, n}} \times \sqrt{n} = \sqrt{\frac{2\log(2/\alpha)}{\sigma_1^2}}.
\end{aligned}$$

This completes the proof of Proposition 30. □

E. Details of Experiments

Details of the bounded source with a specified mean (Section 5.2). For testing the performance of the SCD scheme described in Section 5.2, we constructed a probability distribution, P_θ, over $\mathcal{X} = [0, 1]$ with a specified mean $\theta \in [0, 1]$ by appropriately mixing two uniform distributions. In particular, we define $P_\theta = (1-\theta)U_1 + \theta U_2$, where $U_1 \sim \text{Uniform}([0, \theta])$ and $U_2 \sim \text{Uniform}([\theta, 1])$.

Details of the CDF change detection experiment. (Section 5.3). For this experiment, we used univariate t-distributions with 3 degrees-of-freedom. For the pre-change distribution, we set the mean to 0, and the scale parameter to 1. For the post-change distribution, we set the mean to some value $\Delta > 0$, and the scale parameter to 2.

Details of the Binary classification source (Section 5.5) We consider feature-label pairs $(Z_i, L_i) \in \mathbb{R}^2 \times \{0, 1\}$, with a source distribution $P_S = P_L \times P_{Z|L}$. We assume that the label L is drawn uniformly on the set $\{0, 1\}$, and the features are drawn from a bivariate normal conditioned on the labels: $P_{Z|L} = N(\mu_L, I_2)$, with $\mu_L = (2L-1)[1, 0]^T \in \mathbb{R}^2$. For this problem we will consider linear classifiers parameterized by a weight vector $w \in \mathbb{R}^2$, of the form $h_w(z) = \mathbf{1}_{\langle w, z\rangle \geq 0}$. For the 0-1 loss function, $\phi(z, l, h_w) = \mathbf{1}_{h_w(z) \neq l}$, it is easy to check that the Bayes-optimal classifier for the source distribution is $h^* \equiv h_{w^*}$, with $w^* \propto [1, 0]^T$. For the post change distributions, we rotate the mean of the features by an angle γ; that is, $\mu_L = (2L-1)[\cos\gamma, \sin\gamma]^T$.

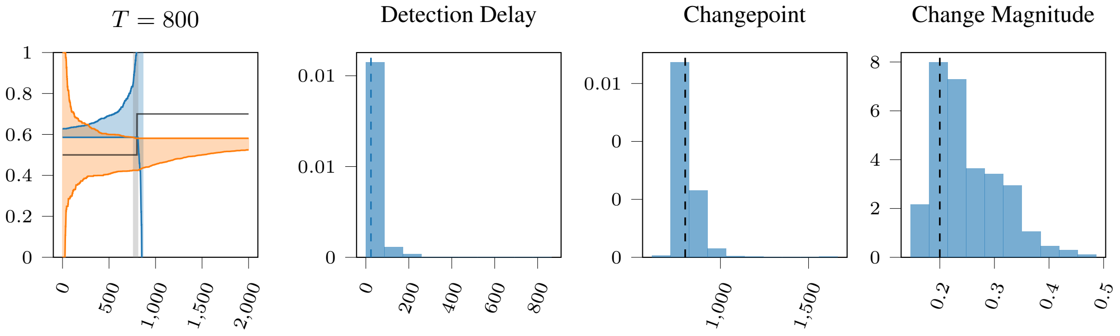

Figure 6. The figures show the performance of our changepoint detection scheme with independent bounded observations whose mean changes from $p_0 = 0.4$ to $p_1 = 0.6$ at the time $T = 800$. The first plot shows the forward and backward CSs at time of detection ($\tau = 863$) in one of the trials, with the shaded gray region being the points at which the two CSs disagree. The next three plots show the empirical distribution of the detection delay, the estimated changepoint location, and the estimated changepoint magnitude over 250 trials of the experiment with the same value of $\Delta = 0.2$ and $\alpha = 0.01$.

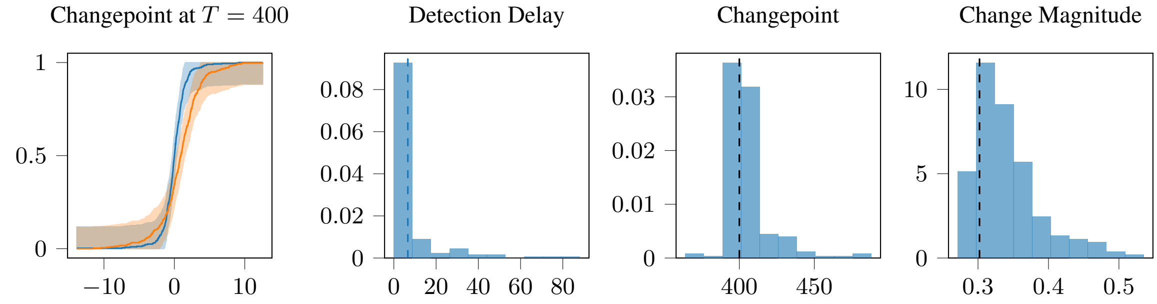

Figure 7. The figures show the performance of our changepoint detection scheme with observations drawn from univariate t-distributions (3 degrees of freedom) whose mean changes from 0 to 1.0 at the time $T = 800$. The first plot shows the forward and backward CSs around the empirical CDFs, at time of detection in one of the trials. The next three plots show the empirical distribution of the detection delay, the estimated changepoint location, and the estimated changepoint magnitude over 250 trials of the experiment.

F. Repeated sequential test interpretation

Our scheme as repeated sequential tests. Due to the equivalence between sequential hypothesis tests and confidence sequences, we can also motivate our general strategy (Definition 12) in the language of sequential hypothesis testing. In particular, our approach can be informally described as follows, due to the time-uniform coverage guarantees of CSs: in each round $t \geq 2$, we run a new sequential hypothesis test for every $1 \leq s \leq t-1$ to decide whether $(X_1, \ldots, X_s)$ and $(X_{s+1}, \ldots, X_t)$ are drawn from the same distribution, or not. As soon as we find a t for which a partition of the observations are sufficiently distinct, we can stop and declare the existence of a changepoint. As we saw in Section 4, this idea can be implemented in an elegant manner by combining a single forward CS (similar to Section 3) with a succession of CSs constructed on reversed versions of the data, that we called 'backward CSs'.

Connections to CuSum. We now demonstrate that we can also interpret the popular parametric SCD scheme, CuSum, as also performing repeated sequential tests. Let f_0 and f_1 denote two density functions on some observation space $\mathcal{X}$. Let $\{X_t : t \geq 1\}$ denote a sequence of independent observations, and consider a changepoint detection problem with f_0 and f_1 as the pre- and post-change distributions respectively.

Definition 33 (CuSum)**.** The cumulative sum (CuSum) method proceeds as follows:

$$\tau_c = \min\{n \geq 1 : W_n \geq b_\alpha\}, \quad \text{where}$$

$$W_1 = 0, \quad \text{and} \quad W_n = \max_{1 \leq t \leq n} \prod_{i=t+1}^{n} \frac{f_1(X_i)}{f_0(X_i)}, \text{ for } t \geq 2.$$

The term b_α is selected to ensure that the ARL is at least $1/\alpha$ for a given $\alpha \in (0, 1)$.

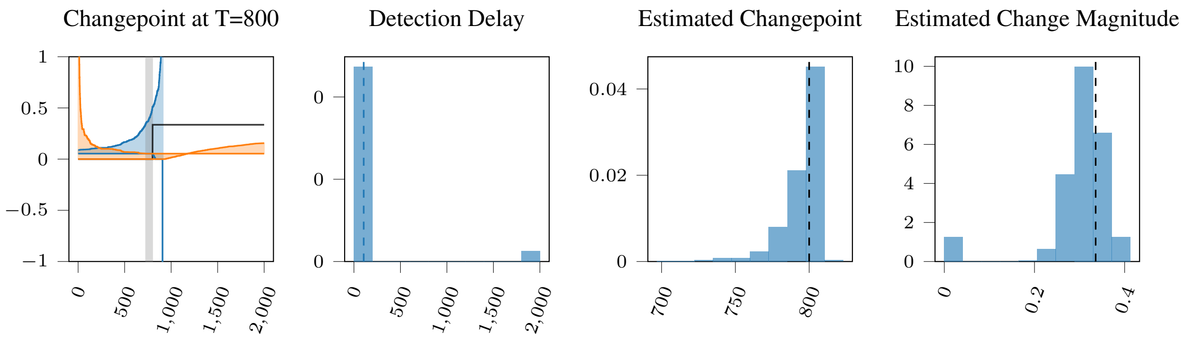

Figure 8. The figures show the performance of our changepoint detection scheme with independent paired multivariate-Gaussian observations whose kernel-MMD distance changes from a pre-change value of 0 to $\Delta \approx 0.33$ at the time $T = 800$.

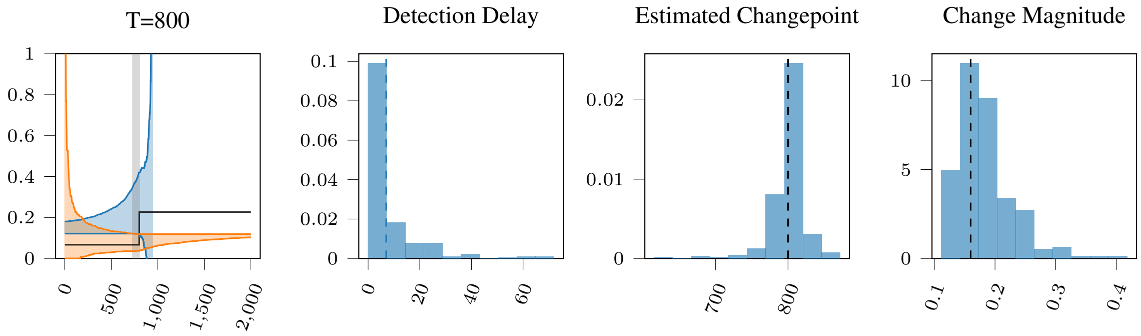

Figure 9. The figures show the performance of our scheme for detecting harmful changes in test distribution for two-dimensional feature vectors, as described in Section 5.5. In these plots, there is a change with magnitude $\Delta \approx 0.16$ at the time $T = 800$.

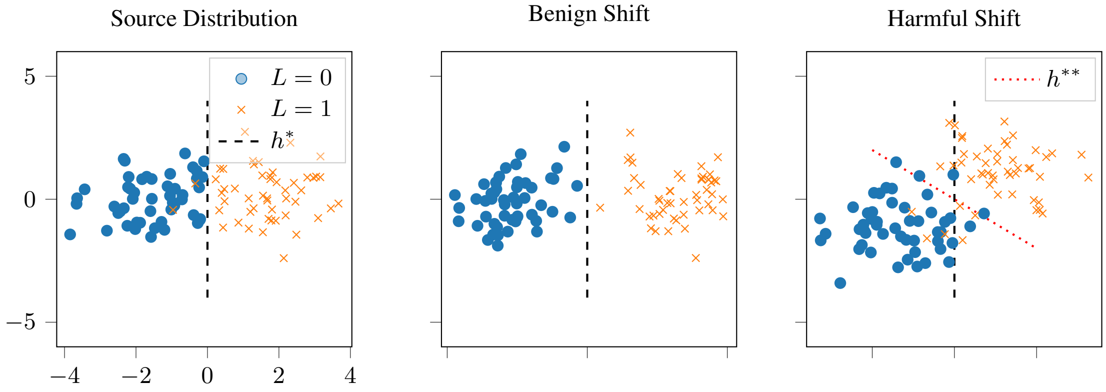

Figure 10. The first plot shows the samples corresponding to the two labels ($L = 0$ and $L = 1$), as well as the optimal linear classifier h^* for this problem (the dashed black line). In the second plot, the distributions are more separated, and the same classifier h^* is also Bayes-optimal for this problem, with a smaller risk. Finally, in the third figure, we have an example of a harmful distribution shift. In this case, the feature distributions for the two labels are rotated anti-clockwise by 45 degrees, which makes h^* a suboptimal classifier for this problem. The new Bayes-optimal classifier is shown by the red dotted line.

Observe that the CuSum procedure can also be interpreted as a sequence of repeated sequential (power-one) tests in the backward direction, testing the null f_0 against the alternative f_1. These are power-one tests (as opposed to the usual SPRT) because they only stop when the likelihood ratio process is large, and not when it is small.

We now introduce an alternative version of CuSum. In this definition, we use the convention that the infimum or minimum over an empty set is infinity, that is, $\inf\{x : x \in \emptyset\} = \infty$.

Definition 34 (CuSum-II)**.** For any $n \geq 1$, and $1 \leq t \leq n$, define the backward sequential test, τ_n^{back}, as follows:

$$\tau_n^{\text{back}} = \min\{1 \leq t \leq n : L_t^n \geq b_\alpha\}, \quad \text{where} \quad L_t^n = \prod_{s=n-t+1}^{n} \frac{f_1(X_s)}{f_0(X_s)}.$$

Then, we can define the modified CuSum stopping time as

$$\tau_c' \coloneqq \inf\{n : \tau_n^{\text{back}} < \infty\}.$$

Proposition 35. *The two tests defined in Definition 33 and Definition 34 are the same.*

Proof. We show that for any $N \in \mathbb{N}$, the sets $\{\tau_c = N\}$ and $\{\tau_c' = N\}$ are equal.

$$\begin{aligned}\{\tau_c = N\} &= \{W_N \geq b_\alpha\} \cap \left(\cap_{n=1}^{N-1}\{W_n < b_\alpha\}\right) \\ &= \{\exists t \in [N] : L_t^N \geq b_\alpha\} \cap \left(\cap_{n=1}^{N-1}\{L_t^n < b_\alpha,\ \forall t \in [n]\}\right) \\ &= \{\tau_N^{\text{back}} < \infty\} \cap \left(\cap_{n=1}^{N-1}\{\tau_n^{\text{back}} = \infty\}\right) \\ &= \{\tau_c' = N\}.\end{aligned}$$

□

F.1. CuSum as an instance of the `BCS-Detector`

We now discuss how the CuSum test for simple pre- and post-change distributions can be considered an instance of our general `BCS-Detector` method. To do this, we need to identify a quantity for which we can construct forward and backward CSs. Define θ_0 and θ_1 as follows:

$$\theta_0 = \mathbb{E}_{f_0}\left[\frac{f_1(X)}{f_0(X)}\right] = 1, \quad \text{and} \quad \theta_1 = \mathbb{E}_{f_1}\left[\frac{f_1(X)}{f_0(X)}\right] = \mathbb{E}_{f_0}\left[\left(\frac{f_1(X)}{f_0(X)}\right)^2\right] = 1 + d_{\chi^2}(f_0, f_1) \geq 1.$$

- Since the pre-change distribution is known, we set the forward CS simply equal to θ_0. That is, we have $C_t = \{1\}$ for all $t \geq 1$.
- For any $n \geq 1$, we can construct a betting based backward CS consisting of subsets of $\Theta = \{\theta_0, \theta_1\} = \{1, \theta_1\}$. To do this, we introduce the following terms, with $\varrho : \mathbb{R} \to [-1, 1]$ denoting an odd sigmoid function:

$$W_t^{(n)}(\theta_1) = \prod_{i=n-t}^{n}\left(1 + \varrho\left(\frac{f_1(X_i)}{f_0(X_i)} - \theta_1\right)\right), \quad \text{and} \quad W_t^{(n)}(1) = \prod_{i=n-t}^{n}\frac{f_1(X_i)}{f_0(X_i)}.$$
$$B_t^{(n)} = \{a \in \Theta : W_t(a) < 1/\alpha\}$$

- Since $C_t = \{1\}$ for all $t \geq 1$, the `BCS-Detector` stops for the first time n, at which the backward CS rejects the point 1. In other words, we can define the stopping time, τ, as follows:

$$\begin{aligned}\tau &\coloneqq \min\{n \geq 1 : \exists t \in [n], W_t^{(n)}(1) \geq 1/\alpha\} \\ &= \min\left\{n \geq 1 : \max_{t \in [n]} W_t^{(n)}(1) \geq 1/\alpha\right\}\end{aligned}$$

  The above stopping time is the same as the original CuSum with $b_\alpha = 1/\alpha$.

G. Details for the Gaussian mean change detection problem

Setup. Recall that in this problem, we are given a stream of real-valued observations, $X_1, X_2, \ldots$, drawn independently according to the distribution $N(\mu_t, 1)$, with $\mu_t = \theta_0$ for $t < T$, and $\mu_t = \theta_1$ for $t \geq T$. Here $\theta_0 \neq \theta_1$ are two unknown parameters, belonging to the parameter set $\Theta = \mathbb{R}$ endowed with the distance metric $d(\theta, \theta') = |\theta - \theta'|$.

Confidence Sequences. To instantiate both of our schemes (the FCS-Detector of Section 3, and the BCS-Detector of Section 4), we need to construct confidence sequences. A suitable closed-form CS for the mean of an independent Gaussian process was derived by Howard et al. (2021), and can be directly employed as the 'forward CS' in both of our schemes.

$$C_t = [\bar{X}_t \pm w_t], \quad \text{where} \quad w_t = \sqrt{\frac{3.4 \log\log(2t) + 0.72 \log(10.4/\alpha)}{t}}.$$

The same CS can also be used in the definition of the new backward CS in every round n, as follows:

$$B_t^{(n)} = \left[\left(\frac{1}{n-t+1} \sum_{i=1}^{t} X_{n-i+1} \right) \pm w_{n-t+1} \right].$$

Note that the width of the (forward) CS decays uniformly to 0 as the number of observations grows, and thus the conditions of both, Proposition 8 and Theorem 13, are satisfied.

Performance Guarantees for `FCS-Detector`. For the SCD scheme obtained by using the above CS in the FCS-Detector method, we can claim the following as a consequence of Proposition 8:

- If there is no changepoint, then the probability that the FCS-Detector ever stops is equal to the probability that the running intersection of the CS $\{C_t : t \geq 1\}$ ever becomes empty. This is upper bounded by α by the definition of CSs.
- Suppose the changepoint occurs at some time $T \geq 1$. For detection to be possible by FCS-Detector, we require the changepoint to satisfy $T \geq T_1$, where $T_1 := \min\{t \geq 1 : w_t < d(\theta_0, \theta_1)\}$. Without this requirement, we do not have enough pre-change observations (relative to the change magnitude $d(\theta_0, \theta_1)$) to estimate the pre-change mean parameter (θ_0) sufficiently well. Assuming this requirement holds, Proposition 8 implies that the following upper bound on the detection delay holds with probability at least $1 - \alpha$:

$$\tau - T \leq \min \left\{ t - \tau : w(t) + w(T) \leq \left(1 - \frac{T}{t}\right) |\theta_0 - \theta_1| \right\}.$$

  As we discussed in Appendix B, this detection delay is $\mathcal{O}(T)$ when $T \approx \frac{\log\log(1/\Delta)}{\Delta^2}$, and $\mathcal{O}(\sqrt{T})$ when $T \to \infty$. In both cases, the detection delay can be made arbitrarily large by making the changepoint T large.

Hence, the FCS-Detector provides a very strong control of false positives at the cost of weak detection guarantees. We now show how a better trade-off is achieved by the BCS-Detector.

Performance Guarantees for `BCS-Detector`. By specializing the general result of Theorem 13 to our problem, we get the following performance guarantees:

- Instead of a high probability bound on false alarm rate, the BCS-Detector provides a guarantee on the average run length (ARL): $\mathbb{E}_0[\tau] \geq \frac{1}{2\alpha} - 3/2$. In other words, it guarantees that when there is no change, the BCS-Detector will raise an alarm roughly every $\frac{1}{2\alpha} - 3/2$ steps.
- When there is a change in distribution at some finite time T, then the scheme guarantees the following upper bound on the detection delay, as we showed in Corollary 17:

$$\mathbb{E}[(\tau - T)^+ | \mathcal{E}] = \mathcal{O}\left(\frac{\log\log(1/\Delta) + \log(1/\alpha)}{\Delta^2} \right), \quad \text{where} \Delta = |\theta_0 - \theta_1|,$$

  and $\mathcal{E}$ is the 'good event' of probability at least $1 - \alpha$, associated with the forward CS.

- Finally, note that if the pre-change mean parameter θ_0 is known, then the event $\mathcal{E}$ has probability 1. Hence, we can get an unconditional version of the above statement in this case. Furthermore, since for Gaussian distributions with unit variance, $d_{\text{KL}}(P_{\theta_1}, P_{\theta_0}) = \frac{1}{2}\Delta^2$, the expected detection delay is $\mathcal{O}\left(\frac{\log\log(1/d_{\text{KL}}(P_{\theta_1},P_{\theta_0})+\log(1/\alpha)}{d_{\text{KL}}(P_{\theta_1},P_{\theta_0})}\right)$, which as we discussed in Remark 18, is asymptotically near-optimal.

H. Additional Experiments

In this section, we compare the performance of our SCD scheme with the kernel-CUSUM strategy of Flynn and Yoo (2019), on the problem of testing for homogeneity in a stream of paired observations (Section 5.4). We consider the following datasets for binary classification from the UCI Machine learning repository: *Higgs*, *Banknote*, and *Occupancy*.

The kernel-CUSUM test proposed by Flynn and Yoo (2019) declares a detection the first time a running estimate of the kernel-MMD distance between the two streams exceeds a threshold a. That is,

$$\begin{aligned}
&\tau = \min\{n \geq 1 : L_t \geq a\}, \quad \text{where} \\
&L_0 = 0, \quad L_{2n} = L_{2n-2} + k(U_{2n-1}, U_{2n}) + k(V_{2n-1}, V_{2n}) - k(U_{2n-1}, V_{2n}) - k(U_{2n}, V_{2n-1}) - \delta.
\end{aligned}$$

Here k denotes a positive definite kernel, and δ is a lower bound on the change magnitude that is assumed to be known apriori. In many problems, a large volume of pre-change data is available, Flynn and Yoo (2019) suggest selecting the rejection threshold a as the smallest value that results in the ARL (computed on historical pre-change data) exceeding a required value. We followed this approach for calibrating the kernel-CuSum test and our proposed SCD scheme of Section 5.4, with a target ARL of 500 on all the three datasets. The results are tabulated in Table 1.

Dataset	ARL (BCS)	ARL (K-CuSum)	Delay (BCS)	Delay (K-CuSum)
Higgs	547.85	569.20	251.45	440.63
Banknote	601.68	594.96	38.47	61.92
Occupancy	505.25	521.80	22.75	58.60

Table 1. Comparison of the performance of our BCS detector for detecting changes in homogeneity with the kernel-CUSUM method of Flynn and Yoo (2019). Both the methods were calibrated to ensure an ARL of at least 500. The results indicate that our BCS detector can achieve a smaller detection delay than the kernel-CuSum while maintaining the required ARL.