ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Sequential change-point detection: Computation versus statistical performance

Haoyun Wang | Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence
Yao Xie, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Email: yao.xie@isye.gatech.edu

Edited by: David Scott, Review Editor and James E. Gentle, Commissioning Editor

Abstract
Change-point detection studies the problem of detecting the changes in the underlying distribution of the data stream as soon as possible after the change happens. Modern large-scale, high-dimensional, and complex streaming data call for computationally (memory) efficient sequential change-point detection algorithms that are also statistically powerful. This gives rise to a computation versus statistical power trade-off, an aspect less emphasized in the past in classic literature. This tutorial takes this new perspective and reviews several sequential change-point detection procedures, ranging from classic sequential change-point detection algorithms to more recent non-parametric procedures that consider computation, memory efficiency, and model robustness in the algorithm design. Our survey also contains classic performance analysis, which provides useful techniques for analyzing new procedures.

This article is categorized under:
  Statistical Models > Time Series Models
  Algorithms and Computational Methods > Algorithms
  Data: Types and Structure > Time Series, Stochastic Processes, and Functional Data

KEYWORDS
anomaly detection, sequential change-point detection, statistical signal processing

1 | INTRODUCTION

Sequential change-point detection has been a classic topic in statistics since the 1920s (Shewhart, 1925, 1931) with motivations in quality control—the objective is to monitor the manufacturing process by examining its statistical properties and signals a change in quality when the distribution of certain features of the products deviates from the desired one. Since then, change-point detection finds many applications in other fields, including monitoring power networks (Chen et al., 2015), internet traffic (Lakhina et al., 2004), sensor networks (Hallac et al., 2015; Raghavan & Veeravalli, 2010), social networks (Peel & Clauset, 2015; Raginsky et al., 2012), medical image processing (Malladi et al., 2013), cybersecurity (Tartakovsky, Polunchenko, & Sokolov, 2012), video surveillance (Lee & Kriegman, 2005), COVID-19 intervention (Dehning et al., 2020), and so on. Recently, there has been much interest in out-of-distribution detection (Magesh et al., 2022; Ren et al., 2019), which aims to detect the shift of the underlying distribution of the test data from training data, which is related to sequential change-point detection.

This tutorial considers sequential (also known as the online or quickest, in some context) change point detection problems, where the goal is to monitor a sequence of data or time series for any change and raise an alarm as quickly as possible the change has occurred to prevent any potential loss. The above problem is different from another major type of change-point detection problem, which is offline and aims to detect and localize (possibly multiple) change-points from sequential data in retrospect (see Truong et al. (2020) for a review). This tutorial focuses on sequential change-point detection.

Shewhart (1925) introduced a control chart that computes a statistic from every sample, where a low statistic value represents the process still within the desired control. Later it is generalized to compute a statistic for several consecutive samples. Page (1954) proposed the CUSUM procedure based on the log-likelihood ratio between the two exactly known distributions, one for the control and one for the anomaly. Moustakides (1986) and Lorden (1971) proved that CUSUM has strong optimality properties. Shiryaev–Roberts (SR) procedure (Roberts, 1966; Shiryaev, 1963) is similar to CUSUM inspired by the Bayesian setting, which is also optimal in several senses (Pollak, 1985; Pollak & Tartakovsky, 2009; Polunchenko & Tartakovsky, 2010; Tartakovsky, Pollak, & Polunchenko, 2012). More recent works on change point detection focus on relaxing the strong assumptions of parametric and known distributions. The generalized likelihood ratio (GLR) procedure (Lorden, 1971; Siegmund & Venkatraman, 1995) aims to detect the change to an unknown distribution assuming a parametric form by searching for the most probable post-change scenario. These classic procedures and their variants can also be found in Basseville and Nikiforov (1993) and Tartakovsky et al. (2014). Sparks (2000) proposed the first adaptive CUSUM, which detects an unknown mean shift by inserting an adaptive estimate of new mean to the CUSUM recursion, followed by Lorden and Pollak (2005), Abbasi and Haq (2019), Cao et al. (2018), and Xie et al. (2022). Such procedures allow an unknown post-change distribution while also having the computational benefit of CUSUM. More recently, Romano et al. (2023) and Ward et al. (2022) present a novel computationally efficient approach that implements the GLR test statistic for the Gaussian mean shift and the Poisson parameter shift detection, respectively. The above works belong to parametric change-point detection, that is, assuming the pre- and post-change distributions belong to the parametric family and detect a certain type of change (for instance, the mean shift and covariance change). There have also been many non-parametric and distribution-free change-point detection algorithms developed, such as kernel-based methods (Harchaoui et al., 2008; Li et al., 2019; Song & Chen, 2022) and graph-based method (Chen & Zhang, 2015; Chu & Chen, 2019).

Sequential change-point detection is different from the traditional fixed-sample hypothesis test. It is fundamentally a repeated likelihood ratio test because of the unknown change point. Moreover, the test statistics are correlated when scanning through time to detect a potential change point. Such correlation is significant, complicates the analysis, and must be explicitly characterized. For instance, in the GLR procedure, we are interested in the probability that the maximum likelihood ratio of all segments of consecutive observations exceeds a certain threshold. In this article, we review two different analytical techniques to tackle such a challenge: an earlier technique developed based on renewal theory (Siegmund, 1985), and a more recently developed method based on a change-of-measure technique for the extreme value of Gaussian random fields (Yakir, 2013).

Despite many methodological and theoretical development for sequential change-point detection, the aspect of (computational and memory) efficiency versus performance tradeoff has been less emphasized. Many change-point detection algorithms are designed, taking such considerations implicitly. This trade-off is becoming more prominent in modern applications due to the need to process large-scale streaming data in real-time and detect changes as quickly as possible. For example, in social network monitoring, one Twitter community will contain at least hundreds of users, let alone the entire network.

Thus, in this tutorial, we aim to contrast classic and new approaches through the lens of *computational efficiency versus statistical performance tradeoff*; we hope this can enlighten developing new techniques for analyzing new algorithms. We highlight the tradeoff by presenting the classic and new performance analysis techniques, ranging from the most standard parametric models to the more recent nonparametric and distribution-free models introduced to achieve computational benefits, such as robustness and flexibility in handling real data. We focus on proof techniques such as renewal theory and the change-of-measure technique, as they are essential for analyzing the classic performance metrics such as the average run length (ARL) and expected detection delay (EDD).

The rest of the article is organized as follows. Section 2 describes the formulation of a sequential change-point detection problem and several classic detection procedures. In Section 3, we define the performance metric and provide the classic analysis of CUSUM in renewal theory. Sections 4–6 each give a detailed review of recent literature on managing unknown distributions, from parametric to distribution-free. Section 7 concludes our review.

We would like to acknowledge that there are other recent techniques for analyzing alternative performance metrics, such as Maillard (2019), Yu et al. (2020), and Chen et al. (2022), which we do not discuss in this article due to space limitations.

2 | BASICS OF CHANGE-POINT DETECTION AND PROBLEM SETUP

2.1 | Problem setup

Consider a basic sequential change-point detection problem, where the aim is to detect an unknown change-point. We are given samples $x_1, x_2, ...$ such that before the change happens, the data distribution follows distribution $f_0 \in \mathscr{F}_0$, and after the change happens, the distribution shifts to $f_1 \in \mathscr{F}_1$. At each time step t after collecting the sample x_t, we can decide whether to raise an alarm or not. Our goal is to raise an alarm as soon as the change has happened under the false alarm constraint.

The simplest setting is when the sets are singletons, that is, the pre- and post-change distributions are f_0 and f_1, respectively. This happens when we understand the physical nature of the process or there is enough historical data to estimate the pre-change distribution f_0. The post-change distribution f_1 can be either a target anomaly or the smallest change we wish to detect. The anomaly may occur at some time $v \in \mathbb{N}$, resulting in $x_1, ..., x_{v-1} \overset{\text{i.i.d.}}{\sim} f_0$ and $x_v, x_{v+1}, ... \overset{\text{i.i.d.}}{\sim} f_1$. If the change never occurs, we write $v = \infty$. Here we assume the change-point v is deterministic but unknown, and the observations before and after the change-point are all independent and identically distributed.

A change-point detection procedure is a *stopping time* τ, which has the following property. For each $t = 1, 2, ...$, the event $\{\tau > t\}$ is measurable with respect to $\sigma(x_1, ..., x_t)$, the σ-field generated by data till time t. In other words, whether the procedure decides to stop right after observing the t-th sample depends solely on the history, not the future. Common detection procedures are based on the choice of a certain detection statistic, computing the detection statistic using the most recent data to form T_t, and stopping the first time that the detection statistics T_t signals a potential change-point (typically by comparing with a predetermined threshold b; the choice of the threshold b balances the false alarm and detection delay). Thus, the detection procedure can be defined as a stopping time

$$\tau = \min\{t \geq 1 : T_t > b\}.$$

Since, in practice, there are usually abundant pre-change samples for us to estimate f_0 with high accuracy, here we only consider the case with known pre-change distribution f_0. Uncertainties in the pre-change distribution can also be addressed though, and we give such an example in Sections 5 and 6. In such cases, it will be harder to control false alarms, and the detection procedure will be more complex computationally in general. We would often like to treat the post-change distribution f_1 as unknown since it is due to an unknown anomaly. Then instead of considering a single distribution f_1, we consider the post-change distribution belonging to a parametric distribution family $f_1(\cdot, \theta)$ with parameter $\theta \in \mathbf{\Theta}_1$.

2.2 | Computation and robustness considerations versus statistical performance

Two commonly used performance metrics for change-point detection are the ARL and the EDD, which we describe below and introduce more formally in Section 3.1. The ARL (related to the false alarm rate) is the expected stopping time under the pre-change distribution (i.e., where there is no change point), and the EDD is the expected stopping time after a change point has happened. Typically, one would choose a threshold to control the ARL to satisfy ARL $\geq \gamma$ for some chosen large lower bound γ, and measure the statistical performance by the EDD for a fixed γ. The well-known lower bound (see, e.g., Lorden, 1971) is EDD $= \log \gamma(1 + o(1))/D(f_1 \| f_0)$ as $\gamma \to \infty$, where $D(f_1 \| f_0)$ is the Kullback–Leibler (KL) divergence between f_0 and f_1. A larger KL divergence means it is easier to distinguish f_1 from f_0 and thus a smaller EDD is possible.

Another way to describe the objective is to minimize

$$\text{ARL}^{-1} + \lambda \, \text{EDD}$$

with some hyper-parameter λ. For some detection procedure τ which minimizes the above for certain λ, it also minimizes the EDD among all detection procedures with $\mathrm{ARL} \geq \gamma = \mathrm{ARL}(\tau)$. Most of the existing change-point detection procedures can be regarded as trying to solve the following minimax problem (with slightly varying definitions on the ARL and EDD):

$$\min_{\tau \in \mathcal{T}} \max_{f_1 \in \mathcal{F}_1} \mathrm{ARL}_{f_0}(\tau)^{-1} + \lambda \, \mathrm{EDD}_{f_1}(\tau), \tag{1}$$

where $\mathcal{T}$ is the set of stopping times, and $\mathcal{F}_1$ is the set of possible post-change distributions. We include f_0, f_1 in the subscript to show the dependence of the performance metrics on those distributions, but since there is often a large amount of reference data to estimate the pre-change distribution, here we only consider f_1 to be unknown. The statistical performance of a detection procedure can then be represented by the value $\max_{f_1 \in \mathcal{F}_1} \mathrm{ARL}_{f_0}(\tau)^{-1} + \lambda \, \mathrm{EDD}_{f_1}(\tau)$. One can expect that as the size of $\mathcal{F}_1$ grows, a detection procedure either becomes more complex or has worse statistical performance. This leads to the trade-off between computation, model robustness (the size of $\mathcal{F}_1$), and statistical performance.

Given observations $x_1, x_2, ..., x_t$, intuitively, to utilize observations and achieve the best statistical performance fully, we would use all the past samples in the detection statistic—as is in the case for the CUSUM and GLR procedures. However, this can become prohibitive in practice as t, the duration we have run the detection procedure grows larger and larger (unless the algorithm is fully recursive such as CUSUM).

A practical online change-point detection algorithm should have constant computation complexity and memory requirement $O(1)$ per iteration (unit time), but this is not likely to be statistically powerful for various cases. However, in some situations, the statistically powerful algorithm will require $O(t)$ computation per iteration, which grows with time and thus is not practical. Therefore, we constantly face a computation and statistical performance tradeoff. Due to this consideration, a commonly adopted simple (yet effective in many cases) strategy is to use *sliding window*: which stores historical data within a sliding window of length w and computes the statistic T_t using data in the sliding window $(x_{t-w+1}, x_{t-w+2}, ..., x_t)$ for every $t \geq w$; this way, both memory and computational complexities are constant in the duration t. We wish to have w as small as possible to minimize the algorithm's memory complexity and computation complexity. However, w also cannot be too small to sacrifice performance. Thus, the critical question is to decide the window length w—how much data needs to be remembered. We will see a trade-off between the memory complexity with respect to the window length w and statistical performance in Section 2.4.

The popularity of CUSUM in practice is probably due to that it achieves constant memory and computational complexity (only use the current sample). Furthermore, its statistical performance is asymptotically optimal (which we will specify more precisely later). However, this optimality of CUSUM requires precise knowledge of the pre-change and post-change distributions, which is not robust to model misspecification. In improving the model's robustness by relaxing such requirements (especially the assumption of known post-change distribution), the tradeoff arises between computation and memory complexity versus statistical performance. For example in window-limited GLR, the memory requirement is $O(w)$ where the window length is directly related to the gap between the pre-change distribution f_0 and $\mathcal{F}_1$ to have good statistical performance. And the computation complexity, in general, is at least $O(w^2)$ per time unit.

Another direction in developing robust change-point detection procedures is to utilize distribution-free methods and non-parametric statistics. Arguably, when the distributional models can be specified more or less reasonably, the non-parametric models are not needed, and they may not be as good as the parametric change-point detection algorithms. However, they gain robustness when the data distributions are not easy to specify using parametric models. While the construction of the non-parametric detection statistics can be straightforward in many cases (such as the kernel-based sequential change-point detection procedure Li et al. (2019)), the performance analysis is much harder than the parametric cases due to a lack of handle through the probability density function in parametric distributional models; the distribution of the non-parametric statistics can be unknown functional form. The asymptotic optimality is also harder to analyze in the sense that a meaningful lower bound is unclear. There are several recent interesting works in this area, and we discuss them in Sections 5 and 6.

In the following, we will describe several common procedures. The comparisons of the detecting statistics are summarized in Figure 1 and Table 1.
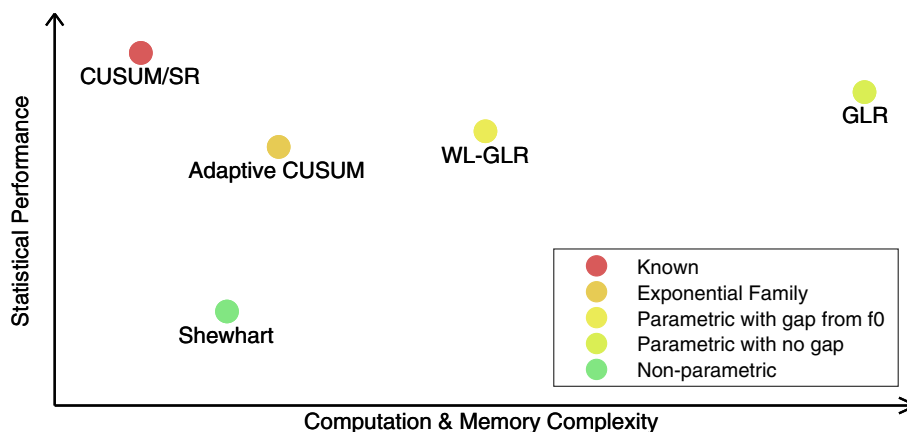
FIGURE 1 Illustration of the statistical performance, computational and memory efficiency, model robustness trade-off for the classic procedures in sequential change detection. The color shows the assumption on post-change distribution.

TABLE 1 Comparisons of classic parametric change-point detection procedures.

	Model robustness	Computation and memory efficiency		Asymptotic optimality
CUSUM/SR	f_1 is known	$O(1)$	$O(1)$	✓
Adaptive CUSUM/SR	Exponential family	$O(1)$	$O(1)$	✓
GLR	Parametric	$\geq O(t)$	$O(t)$	✓
Window-limited GLR	Parametric	$\geq O(w)$	$O(w)$	✓
Shewhart chart	Can be nonparametric	Varies	$O(w)$	✗

Note: Model robust means whether the detection procedure requires an accurate post-change distribution as an input. Efficiency includes both computational and memory requirements per time unit, and we say the procedure is efficient if both are a constant with respect to the duration t we monitor the process. Asymptotic optimality is reached when the performance satisfies the theoretical lower bound given in Section 3.1, Theorem 1. Note that the Shewhart chart contains a broad range of detecting procedures and can be based on either parametric or non-parametric statistics.

2.3 | Classic CUSUM procedure

The CUSUM procedure is derived based on likelihood ratios. For an assumed change-point location ν, the log-likelihood ratio for the hypothesis $H_0 : x_1, x_2, \ldots \sim f_0$ versus the alternative hypothesis $H_\nu : x_1, \ldots, x_{\nu-1} \sim f_0, x_\nu, x_{\nu+1} \ldots \sim f_1$ is given by

$$\sum_{i=\nu}^{t} \log\left(\frac{f_1(x_i)}{f_0(x_i)}\right)$$

Since change-point ν is unknown, the detection statistic needs to consider the maximization of the above with respect to all possible change-point locations. This gives rise to the following detection statistic for each time t

$$T_t^{\text{CUSUM}} = \max_{1 \leq k \leq t+1} \sum_{i=k}^{t} \log\left(\frac{f_1(x_i)}{f_0(x_i)}\right). \tag{2}$$

Here when $k = t+1$, we set the empty summation to 0. The CUSUM procedure computes the detection statistic for each time t, stops the first time that the detection statistic exceeds a certain threshold $b > 0$, and claims there has been a change in the past:

$$\tau^{\text{CUSUM}}(b) = \inf\{t : T_t^{\text{CUSUM}} > b\}. \tag{3}$$

The popularity of CUSUM is possibly due to the following recursive computation of the detection statistic. Let $z_i = \log(f_1(x_i)/f_0(x_i))$ be the increment in the log-likelihood ratio, $i = 1, 2, \ldots$, and $S_t = \sum_{i=1}^{t} z_i$, $\forall t \geq 0$. Then the CUSUM statistic has the following recursive expression because the partial sums starting from each potential change-point k shares the same increment when updating from time step t to $t+1$:

$$T_{t+1}^{\text{CUSUM}} = \max_{0 \leq k \leq t+1}(S_{t+1} - S_k) = \max\{z_{t+1} + \max_{0 \leq k \leq t}(S_t - S_k), 0\} = \max\{z_{t+1} + T_t^{\text{CUSUM}}, 0\}, \tag{4}$$

with $T_0^{\text{CUSUM}} = 0$. This convenient recursive CUSUM evaluation (4) means that we do not need to remember any data and merely update the detection statistic every time we observe a new sample. More precisely, for each time slot, one only needs to compute the new log-likelihood ratio increment z_i and update the CUSUM statistic T_t^{CUSUM} according to (4). In other words, both the computation and memory complexity per update are constant for CUSUM.

Moreover, CUSUM's statistical performance enjoys asymptotic optimality (Lorden, 1971), which we will explain in Section 3.1. It is exactly optimal shown later by (Moustakides, 1986). Largely speaking, for CUSUM to enjoy a good performance, we need the property that the expected value of the increment term before the change happens is negative ($\mathbb{E}_0[z_t] < 0$), and positive after the change ($\mathbb{E}_1[z_t] > 0$). This can be verified using Jensen's inequality (when the data distribution specifications are precise).

However, a known drawback of CUSUM is that it requires both the pre-change and post-change distributions to be parametric and specified exactly, which can be too restrictive for real-world applications. When the true distribution deviates from the assumed distributions, CUSUM is no longer optimal and suffers from performance loss depending on the level of model misspecification. For modern data, especially high-dimensional data, having an exact specification of data distribution is difficult, and performance degradation of CUSUM due to model mismatch becomes inevitable.

Multiple variants have been developed to make CUSUM more robust to unknown post-change distributions. One possibility is to run multiple CUSUM procedures in parallel, each detecting against a different possible anomaly outcome (Lorden & Eisenberger, 1973; Lucas, 1982; Tartakovsky, 2005; Xie & Siegmund, 2013). Another approach is to consider the so-called *adaptive CUSUM*, to adaptively estimate the post-change distribution while running the CUSUM recursion (Abbasi & Haq, 2019; Lorden & Pollak, 2005; Sparks, 2000; Xie et al., 2020). Uncertainties in the pre-change distribution can be treated similarly (Krieger et al., 1999; Mei, 2006; Pollak & Siegmund, 1991); however, to the best of our knowledge, there is yet a study regarding how to obtain recursive expression when the pre-change distribution is unknown. A more detailed review of adaptive CUSUM and its recent progress can be found in Section 4.

2.4 | Other classic procedures

2.4.1 | Shiryaev–Roberts procedure

The SR procedure, similar to the CUSUM, is based on the log-likelihood ratio between the completely specified pre-change and post-change distribution. It is first inspired by putting a prior on the change-point ν and assuming it follows a geometric distribution (Shiryaev, 1963). The SR procedure can now be seen as the resulting procedure when the geometric mean goes to infinity. This is reflected in the detection statistic, where the SR procedure takes the summation instead of the maximum over all the potential change-point k (as done in CUSUM),

$$T_t^{\text{SR}} = \sum_{k=1}^{t} \prod_{i=k}^{t} \frac{f_1(x_i)}{f_0(x_i)}.$$

And the recursion (similar to CUSUM) is in the form of

$$T_{t+1}^{\text{SR}} = \left(1 + T_t^{\text{SR}}\right) \frac{f_1(x_{t+1})}{f_0(x_t)}.$$

The SR procedure is exactly optimal under the integral expected detection delay (Pollak & Tartakovsky, 2009) and later Tartakovsky, Pollak, and Polunchenko (2012) proved that the SR procedure is third-order asymptotically optimal following Pollak's (1985) definition of the EDD. They also compared different initial values of T_0^{SR} and discussed the detection delay conditioned on the change-point k. See Polunchenko and Tartakovsky (2012) for a review on sequential change-point detection with known pre- and post-change distributions.

2.4.2 | Generalized likelihood ratio procedure

The GLR procedure is adopted when the post-change distribution is unknown but still parametric. The test statistic scan through the log-likelihood ratio over all the potential change-point k and for each assumed k and maximum likelihood estimator of the post-change parameter is used in forming the likelihood ratio:

$$T_t^{\mathrm{GLR}} = \max_{1 \le k \le t} \sup_{\theta \in \boldsymbol{\Theta}_1} \sum_{i=k}^{t} \log\left(\frac{f_1(x_i, \theta)}{f_0(x_i)}\right).$$

The GLR procedure is asymptotically optimal in its statistical performance and handles the problem of unknown post-change distribution automatically by searching through all possible parameters in the feasible region. Lorden (1971) proves the first-order asymptotic optimality for the GLR procedure with univariate exponential family, and later Siegmund and Venkatraman (1995) provides a more precise characterization of the ARL on the problem of a mean shift with known variance, which turns out to be very useful in analyzing modern change-point detection procedures as well. We will review this method in Section 6. Unfortunately, in general, T_{t+1}^{GLR} cannot be updated recursively from T_t^{GLR} and the computation and memory needed per update is often at least linear in t. An exception with univariate exponential family can be found in Romano et al. (2023) where the complexity is reduced to $O(\log t)$.

2.4.3 | Window-limited GLR

A more computationally efficient procedure than the above vanilla GLR is the window-limited GLR (Willsky & Jones, 1976) developed by taking the maximum over all potential change-points within a sliding window of fixed length w:

$$T_t^{\mathrm{WL\text{-}GLR}} = \max_{t-w \le k \le t} \sup_{\theta \in \boldsymbol{\Theta}_1} \sum_{i=k}^{t} \log\left(\frac{f_1(x_i, \theta)}{f_0(x_i)}\right).$$

At each time step, the computational and memory requirement of window-limited GLR is constant with respect to t. It appears that by forgetting previous samples before time $t - w$, we may lose information. But it can be shown that with a proper window length w, the window-limited GLR can still be asymptotically optimal (Lai & Shan, 1999); this requirement is $w \ge \log \gamma / D_{\min}$, where γ is the ARL and $D_{\min}$ is the smallest KL-divergence between f_0 and potential f_1 we want to detect. Because the window-limited GLR still scans through all potential change points within the sliding window, the computational cost scales with w; in general, it is reasonable to assume at least $O(w)$ operations are needed to find the supremum over $\theta \in \boldsymbol{\Theta}_1$. There are w potential change points per update so the computation complexity would be at least $O(w^2)$. When working with exponential family, however, the computation complexity is reduced to $O(w)$ because the partial sum $\sum_{=k}^{t} x_i$ is a sufficient statistic for computing the maximum likelihood ratio for each potential change point k.

Shewhart chart is one of the earliest sequential change-point procedures (Shewhart, 1925, 1931). Still, many recent detecting procedures fall into this type due to its simplicity: an offline test can be converted into a sequential change-point detection procedure easily by applying the test to a *sliding window of samples*. More precisely, the detection statistic T_t using a sliding window of fixed length w. For instance, the simplest example is by letting T_t be the average of the past w samples to detect a mean shift. If the problem is parametric, it can be the generalized likelihood ratio:

$$T_t^{\text{SH-GLR}} = \sup_{\theta \in \Theta_1} \sum_{i=t-w}^{t} \log\left(\frac{f_1(x_i, \theta)}{f_0(x_i)}\right), \tag{5}$$

And again when dealing with exponential families, the above turns into an explicit function of the partial sum $\sum_{i=t-w}^{t} x_i$. Another example of Shewhart chart is based on the score statistic (e.g., Chen et al. (2020)) utilizing the locally most powerful score statistic,

$$T_t^{\text{Score}} = \frac{1}{2w} \nabla_\theta^T \left(\sum_{i=t-w}^{t} \log f_1(x_i, \theta)\right) \mathcal{I}^{-1}(0) \nabla_\theta \left(\sum_{i=t-w}^{t} \log f_1(x_i, \theta)\right),$$

where $\mathcal{I}(0)$ is the Fisher information Casella and Berger (2021) at $\theta = 0$ and it can be pre-computed.

The score statistic can sometimes lead to simple detection statistics. Since we only consider the gradient of the likelihood ratio at the single parameter value under the null hypothesis, sometimes this avoids solving optimization problems or matrix inversion. Similar to the window-limited GLR, the choice of the window length w depends on the smallest change we are interested in detecting. To further simplify computation, one can replace computing the statistic T_t at every time step with updating it every δ time unit for δ up to w.

We would like to emphasize that a distinction is that the Shewhart detection statistic, unlike CUSUM, does not scan over all potential change-point locations. Because the Shewhart chart no longer scans over potential change points, its statistical performance is not asymptotically optimal with respect to the ARL EDD metric (with the exception that if we consider a different metric: maximizing the probability of detection [Moustakides, 2014; Pollak & Krieger, 2013]). It is commonly believed that CUSUM is more statistically powerful for detecting small changes.

3 | STATISTICAL PERFORMANCE

We start by reviewing statistical performance analysis for change-point detection algorithms. The techniques used for statistical performance analysis include renewal theory, Wald's identity, and change-of-measure techniques (see, e.g., Lai, 1998; Lorden, 1971; Lorden & Eisenberger, 1973; Page, 1954; Pollak, 1985; Siegmund, 1985; a recent survey in quickest change-point detection can be found in Xie et al. (2021)). We review such techniques because they are still useful for analyzing new change-point detection procedures (Tartakovsky, 2019).

3.1 | Standard performance metrics: ARL and EDD

The performance of a change-point detection procedure with stopping time τ can be evaluated by the average stopping time after a change with a false alarm constraint. Let $\mathbb{P}_v, \mathbb{E}_v$ be the probability measure and expectation given time of change $v = 1, \dots$ (with $\nu = \infty$ denoting no change).

Consider the following problem. We would like to find a detection procedure that can minimize the worst-case expected detection delay (Pollak, 1985):

$$\text{EDD}(\tau) := \sup_{v \geq 1} \mathbb{E}_v[\tau - v + 1 | \tau \geq v], \tag{6}$$

subject to average run length

$$\text{ARL}(\tau) := \mathbb{E}_0[\tau] \geq \gamma \tag{7}$$

for some large constant $\gamma > 0$. (Note that there are other definitions of the worst-case detection delay, such as that in Lorden (1971), which we do not consider here.) The theoretical lower-bound of the EDD has related to the Kullback–Leibler (KL) divergence between the post-change and pre-change distributions:

$$D(f_1 \| f_0) = \mathbb{E}_1[\log(f_1(x)/f_0(x))].$$

Theorem 1. *Lower bound for EDD (Lai, 1998; Pollak, 1985).*

$$\inf_{\tau:\mathrm{ARL}(\tau) \geq \gamma} \mathrm{EDD}(\tau) \geq \frac{\log \gamma}{D(f_1 \| f_0)}(1 + o(1)),$$

as $\gamma \to \infty$.

It can be shown that the CUSUM procedure satisfies

$$\mathrm{ARL}\big(\tau^{\mathrm{CUSUM}}(b)\big) = \Theta\big(e^b\big), \quad \mathrm{EDD}\big(\tau^{\mathrm{CUSUM}}(b)\big) = \frac{b}{D(f_1 \| f_0)}(1 + o(1)), \tag{8}$$

as the threshold $b \to \infty$, and hence it is asymptotically optimal; here, Θ is the big-theta notation. In the next subsection, we provide the analysis leading to (8) for the basic i.i.d. setup. For non-i.i.d. data, (8) still holds when the average log-likelihood ratio between $\mathbb{P}_0$ and $\mathbb{P}_v$ converges to some constant in probability as $t \to \infty$ (Lai, 1998).

3.2 | Analysis via renewal theory and Wald's identity

Now we demonstrate the analysis of the ARL and EDD approximation (8) for the CUSUM procedure when the data are i.i.d. By the recursive rule (4), $\big(T_t^{\mathrm{CUSUM}}\big)_{t \geq 1}$ in CUSUM can be viewed as a random walk. Each time the random walk is increased by $\log(f_1(x_t)/f_0(x_t))$, with the exception that if the path steps below 0, there is a renewal and the path is reset to 0. The random walk ends the first time when the path exceeds the threshold b.

To analyze the EDD of CUSUM procedure $\tau^{\mathrm{CUSUM}}(b)$ for i.i.d. data, first, we have the following fact that greatly simplifies EDD's computation. For any $b \geq 0$, for CUSUM procedure $\tau^{\mathrm{CUSUM}}(b)$ (which is abbreviated as τ in the informal proof to simplify notation)

$$\sup_{v \geq 1} \mathbb{E}_v[\tau - v + 1 | \tau \geq v] = \mathbb{E}_1[\tau]. \tag{9}$$

Proof. (Informal) We first show that for any $v \geq 1$,

$$\mathbb{E}_v[\tau - v + 1 | \tau \geq v] \leq \mathbb{E}_1[\tau].$$

Thus, the conditional expectation of detection delay in EDD's definition (6) reaches its supremum at $v = 1$. Conditioned on $\tau \geq v$ and T_{v-1}^{CUSUM}, τ is decided by the random walk starting from $T_{v-1}^{\mathrm{CUSUM}} \geq 0$ with increments $z_v, z_{v+1}, z_{v+2}, \ldots$, which are function values of i.i.d. random variables $x_v, x_{v+1}, \ldots, \sim f_1$. By the independence between samples, the increments $z_v, z_{v+1}, \ldots$, are independent of T_{v-1}^{CUSUM} and the event $\tau \geq v$. For every $v \geq 2$ and $s \geq 0$, there must be

$$\mathbb{E}_v\big[\tau - v + 1 | \tau \geq v, T_{v-1}^{\mathrm{CUSUM}} = s\big] \leq \mathbb{E}_v\big[\tau - v + 1 | \tau \geq v, T_{v-1}^{\mathrm{CUSUM}} = 0\big]$$

because the random walk with renewal must be at least as large as if it starts from $T_{v-1}^{\mathrm{CUSUM}} = 0$, and hence stops earlier. Also note that the random walk $\big(T_{v-1+t}^{\mathrm{CUSUM}}\big)_{t \geq 0}$ conditioned on $T_{v-1}^{\mathrm{CUSUM}} = 0$ under $\mathbb{P}_v$ is identically distributed with $\big(T_t^{\mathrm{CUSUM}}\big)_{t \geq 0}$ starting from $T_0^{\mathrm{CUSUM}} = 0$ under $\mathbb{P}_1$, so their expected stopping time should be the same, which gives $\mathbb{E}_v\big[\tau - v + 1 | \tau \geq v, T_{v-1}^{\mathrm{CUSUM}} = 0\big] = \mathbb{E}_1[\tau]$. Then

$$\mathbb{E}_\nu[\tau - \nu + 1 | \tau \geq \nu] = \mathbb{E}_\nu \left[\mathbb{E}_\nu \left[\tau - \nu + 1 | \tau \geq \nu, T_{\nu-1}^{\mathrm{CUSUM}} \right] | \tau \geq \nu \right] \leq \mathbb{E}_\nu \left[\tau - \nu + 1 | \tau \geq \nu, T_{\nu-1}^{\mathrm{CUSUM}} = 0 \right] = \mathbb{E}_1[\tau].$$

Now with (9), we can compute the EDD using renewal properties under $\mathbb{P}_1$. Let N_1 be the first time the random walk $\left(T_t^{\mathrm{CUSUM}} \right)_{t \geq 1}$ falls out of the interval $(0, b)$. If $T_{N_1}^{\mathrm{CUSUM}} \geq b$, the random walk stops, otherwise there is a renewal, and $T_{N_1}^{\mathrm{CUSUM}}$ is reset to be 0. We can continue and define $N_2, N_3, \ldots$ to be the time the renewed random walk $\left(T_{N_1+t}^{\mathrm{CUSUM}} \right)_{t \geq 1}$ first falls out of $(0, b)$, until finally for some M, $T_{N_1+\ldots+N_M}^{\mathrm{CUSUM}} \geq b$. Since the increments $z_t, t = 1, 2, \ldots$ are i.i.d., so are the random walks after renewal. This means $N_1, N_2, \ldots$ are i.i.d, and the conditional probability $\mathbb{P}_1(M = m | M \geq m)$ is a constant with respect to m. So M follows a geometric distribution with expectation $1/\mathbb{P}_1(M = 1)$. Then we need Wald's identity on the expectation of summation of random variables with a stopping time.

Theorem 2. *(Wald's identity)* For i.i.d. random variables $X_1, X_2, \cdots$ and a stopping time τ, if $\mathbb{E}[X_1], \mathbb{E}[\tau] < \infty$, then

$$\mathbb{E}\left[\sum_{i=1}^{\tau} X_i \right] = \mathbb{E}[X_1]\mathbb{E}[\tau].$$

By the Wald's identity,

$$\mathbb{E}_1\left[\tau^{\mathrm{CUSUM}}(b) \right] = \mathbb{E}_1\left[\sum_{i=1}^{M} N_i \right] = \mathbb{E}_1[N_1]\mathbb{E}_1[M] = \mathbb{E}_1[N_1]/\mathbb{P}_1(M = 1). \tag{10}$$

For the ARL, the same argument applies with the probability measure $\mathbb{P}_0$ and

$$\mathbb{E}_0\left[\tau^{\mathrm{CUSUM}}(b) \right] = \mathbb{E}_0[N_1]/\mathbb{P}_0(M = 1). \tag{11}$$

Now we have established that the ARL equals to $\mathbb{E}_0[N_1]/\mathbb{P}_0(M = 1)$ and the EDD equals to $\mathbb{E}_1[N_1]/\mathbb{P}_1(M = 1)$. Next, we analyze these values using martingale properties. Let $(S_t)_{t \geq 0}$ be the random walk with increments $z_1, z_2, \ldots$ without renewal, that is, $S_t = \sum_{i=1}^{t} z_i, \forall t \geq 0$. Let the stopping time
$N = N_1 = \min\{t \geq 1 : S_t \notin (0, b)\}$.

3.2.1 | Ladder variables

For now, the increments z_i can be either positive or negative. To connect with the renewal theorem, we introduce the ladder variable z^+, which is the value when $(S_t)_{t \geq 0}$ first hits the positive axis, and let

$$\tau_+ = \inf\{t : S_t > 0\}.$$

By the law of large numbers, since $\mathbb{E}_1[z_1] > 0$, τ_+ must be finite under $\mathbb{P}_1$. So $z^+ = S_{\tau_+}$ is well defined under $\mathbb{P}_1$. We also define two other stopping times for later use.

$$\tau_b = \inf\{t : S_t \geq b\}, \tau_- = \inf\{t : S_t < 0\}.$$

As we will see later, the value of interest is the overshoot $S_{\tau_b} - b$ under $\mathbb{P}_1$, and it suffices to consider only the distribution of z^+. This is because $(S_t)_{0 \leq t \leq \tau_b}$ can break down into several pieces, each ending when the random walk reaches a new highest. Therefore, S_{τ_b} can be expressed as the sum of i.i.d. copies of z^+.

3.2.2 | Approximating ARL

For $i = 0$, $(\exp(S_t))_{t=0}^{N}$ is a martingale under $\mathbb{P}_0$ because for every $0 \leq t < N$,

$$\mathbb{E}_0[\exp(S_{t+1})|\exp(S_t)] = \exp(S_t)\mathbb{E}_0[f_1(x_{t+1})/f_0(x_{t+1})] = \exp(S_t).$$

By Doob's Martingale inequality,

$$\mathbb{P}_0(S_N \geq b) = \mathbb{P}_0\big(\exp(S_N) \geq e^b\big) \leq \frac{\mathbb{E}_0[\exp(S_0)]}{e^b} = e^{-b}. \tag{12}$$

And obviously $N \geq 1$, so

$$\mathrm{ARL}\big(\tau^{\mathrm{CUSUM}}(b)\big) = \mathbb{E}_0[N]/\mathbb{P}_0(S_N \geq b) \geq e^b. \tag{13}$$

To better approximate the ARL and improve the multiplicative constant, first, we look at $\mathbb{E}_0[N]$. As $b \to \infty$, $N \to \tau_-$ and $\mathbb{E}_0[N] \to \mathbb{E}_0[\tau_-]$ because of the monotone convergence theorem. The divider in (11) can be evaluated using a change-of-measure trick,

$$\mathbb{P}_0(S_N \geq b) = \mathbb{E}_0[\mathbf{1}\{S_N \geq b\}] = \mathbb{E}_1[\exp(-S_N)\mathbf{1}\{S_N \geq b\}] = e^{-b}\mathbb{E}_1[\exp(-(S_N - b))\mathbf{1}\{S_N \geq b\}], \tag{14}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function of an event. The second equality is due to the fact that $\exp(-S_N)$ is the likelihood ratio between $\mathbb{P}_0$ and $\mathbb{P}_1$. We use the renewal theorem to estimate the right-hand side of (14).

Theorem 3. *(Renewal theorem) (*Siegmund, *1985, chapter 8) For any $y \geq 0$,*

$$\lim_{b \to \infty} \mathbb{P}_1(S_{\tau_b} - b > y) = \mathbb{E}_1[z^+]^{-1} \int_y^\infty \mathbb{P}_1(z^+ > x)dx.$$

This result specifies the asymptotic tail distribution of the "over-shoot" term as $b \to \infty$, useful in deriving many properties of the stopping time. Then the right hand side of (14) can be approximated using

$$\mathbb{E}_1[\exp(-(S_N - b))\mathbf{1}\{S_N \geq b\}] = \mathbb{P}_1(S_N \geq b)\mathbb{E}_1[\exp(-(S_N - b))|S_N \geq b] \approx \mathbb{P}_1(S_N \geq b)\mathbb{E}_1[\exp(-(S_{\tau_b} - b))]$$

as done in Lorden and Eisenberger (1973). When $b \to \infty$, $\mathbb{P}_1(S_N \geq b) \to \mathbb{P}_1(\tau_- = \infty)$, and

$$\mathbb{E}_1[\exp(-(S_{\tau_b} - b))] \to \mathbb{E}_1[z^+]^{-1}(1 - \mathbb{E}_1[\exp(-z^+)])$$

as a result of the renewal theorem above. So putting the results above together, we have

$$\mathrm{ARL}\big(\tau^{\mathrm{CUSUM}}(b)\big) \approx \frac{\mathbb{E}_0[\tau_-]\mathbb{E}_1[z^+]}{\mathbb{P}_1(\tau_- = \infty)(1 - \mathbb{E}_1[\exp(-z^+)])} e^b.$$

3.2.3 | Approximating EDD

Since $(S_t - D(f_1\|f_0)t)_{t \geq 0}$ is a martingale under $\mathbb{P}_1$, by the optional stopping theorem,

$$\mathbb{E}_1[S_{\tau_b} - D(f_1\|f_0)\tau_b] = 0,$$

and $\mathbb{E}_1[\tau_b] = (b + \mathbb{E}_1[S_{\tau^{\mathrm{CUSUM}}} - b])/D(f_1\|f_0)$. Note that $S_t \geq T$ for any t, so CUSUM must stop no later than τ_b,

$$\mathrm{EDD}\big(\tau^{\mathrm{CUSUM}}(b)\big) \leq \mathbb{E}_1[\tau_b] = \frac{b + \mathbb{E}_1[S_{\tau_b} - b]}{D(f_1 \| f_0)},$$

where as a result of Theorem 3,

$$\lim_{b \to \infty} \mathbb{E}_1[S_{\tau_b} - b] = \lim_{b \to \infty} \int_0^\infty \mathbb{P}_1(S_{\tau_b} - b > y) dy = \int_0^\infty \lim_{b \to \infty} \mathbb{P}_1(S_{\tau_b} - b > y) dy$$

$$= \mathbb{E}_1[z^+]^{-1} \int_0^\infty \int_y^\infty \mathbb{P}_1(z^+ > x) dx dy = \mathbb{E}_1[z^+]^{-1} \int_0^\infty x \mathbb{P}_1(z^+ > x) dx$$

$$= \mathbb{E}_1\left[(z^+)^2\right] / 2\mathbb{E}_1[z^+].$$

Here we assume we can change the order of taking limit and integral but this is not an issue using a more complete version of the renewal theorem. For details, we refer to Siegmund (1985).

For the stopping time N_1, similarly, there is $\mathbb{E}_1[S_N - D(f_1 \| f_0)N] = 0$, and

$$\mathbb{E}_1[N] = \frac{\mathbb{E}_1[S_N]}{D(f_1 \| f_0)} \geq \frac{\mathbb{E}_1[S_N \mathbf{1}\{S_N \geq b\}]}{D(f_1 \| f_0)} \geq \frac{\mathbb{P}_1(S_N \geq b)b}{D(f_1 \| f_0)},$$

with (10), together there is

$$\frac{b}{D(f_1 \| f_0)} \leq \mathrm{EDD}\big(\tau^{\mathrm{CUSUM}}(b)\big) \leq \frac{b + \mathbb{E}_1\left[(z^+)^2\right] / 2\mathbb{E}_1[z^+] + o(1)}{D(f_1 \| f_0)}.$$

4 | ADAPTIVE CUSUM

When the post-change distribution belongs to some family parametrized by $\theta \in \Theta_1$, adaptive CUSUM provides another framework other than GLR. At each time slot, the statistic T_t is computed using the CUSUM recursive rule (4), where the increment z_t is the log-likelihood ratio $\log\big(f_1\big(x_t, \widehat{\theta}\big)/f_0(x_t)\big)$ for some estimator $\widehat{\theta}_t$. Such method first appears in sequential hypothesis testing where the alternative hypothesis is composite, and a non-anticipating estimator replaces the simple alternative distribution in the sequential likelihood ratio test (Pavlov, 1991; Robbins & Siegmund, 1972, 1974). Tartakovsky (2014) compared the adaptive likelihood ratio with the generalized likelihood ratio type-of method, and found the former more convenient in controlling the probability of false alarm. In change-point detection, the key difference between adaptive CUSUM and (even Shewhart-chart type of) GLR is that while both search for the most probable post-change parameter $\widehat{\theta}_t$, GLR estimates the post-change parameter and finds evidence for the change in the same window, and adaptive CUSUM do these separately. It turns out that we need fewer observations to estimate the post-change parameters than to find evidence, and this can be one reason why adaptive CUSUM outperforms GLR in computational aspects.

An earlier work we found that uses adaptive CUSUM is (Sparks, 2000), which considers the case where the change-point causes a mean shift with an unknown scale. They proposed to estimate the shifted mean $\widehat{\theta}_t$ using the exponentially weighted moving average (EWMA) and update the CUSUM statistic using the new sample x_t and $\widehat{\theta}_t$ as the targeted change. Then the estimated change is updated using $\widehat{\theta}_{t+1} = \alpha x_t + (1-\alpha)\widehat{\theta}_t$ for some $\alpha \in (0,1)$. Following the same spirit, Cao et al. (2018) propose to update the estimate $\widehat{\theta}_t$ using online convex optimization algorithms to manage exponential families while the estimated change $\widehat{\theta}_t$ remains to be computed recursively. More recently, Xie et al. (2022) studied a window-limited CUSUM which can be applied to general distribution families of parametric forms. This procedure estimates the $\widehat{\theta}_t$ in a sliding window of length w, and the statistic is given by the recursion

$$T_{t+1}^{\text{WL-CUSUM}} = \max\left\{ T_t^{\text{WL-CUSUM}} + \log\frac{f_1\left(x_{t+1}, \widehat{\theta}_{t+1}^w\right)}{f_0(x_{t+1})}, 0 \right\},$$ (15)

where $\widehat{\theta}_t^w$ is a consistent estimator (e.g., the maximum likelihood estimator) of θ using a sliding window of past w samples: $x_{t-w}, \ldots, x_{t-1}$. Unlike the window-limited GLR where $w \sim \Theta(\log\gamma)$ needs to be in the same order with EDD (or the threshold b) to collect enough information to detect a change, the window-limited CUSUM has smaller memory requirement $w \sim \Theta(\sqrt{\log\gamma})$, possibly explained by the fact that in window-limited CUSUM we only need enough samples to obtain a reasonably good estimate of $\widehat{\theta}_t$. The optimal window length is explained by the following upper bound on the EDD, and with this window length window-limited CUSUM is asymptotically optimal in its statistical performance:

Theorem 4. *(EDD of window-limited CUSUM) (Xie et al., 2022)*

$$\text{EDD} \leq \frac{b + \widehat{J}_0/\widehat{D}(f_{1,\theta}\|f_0) + \left(b\widehat{J}_0/\widehat{D}(f_{1,\theta}\|f_0)\right)^{1/2} + wD(f_{1,\theta}\|f_0) + \left(D(f_1\|f_0)w\widehat{J}_0/\widehat{D}(f_{1,\theta}\|f_0)\right)^{1/2}}{\widehat{D}(f_{1,\theta}\|f_0)},$$

where $D(f_{1,\theta}\|f_0)$ is the KL-divergence and J_0 the second-order moment of the log-likelihood between the (true) post-change and pre-change distribution,

$$D(f_{1,\theta}\|f_0) = \mathbb{E}_{1,\theta}\left[\log\frac{f_1(x_1,\theta)}{f_0(x_1)}\right], J_0 = \mathbb{E}_{1,\theta}\left[\log^2\frac{f_1(x_1,\theta)}{f_0(x_1)}\right],$$

$\widehat{D}(f_{1,\theta}\|f_0), \widehat{J}_0$ is defined similarly with true parameter θ replaced by the estimator,

$$\widehat{D}(f_{1,\theta}\|f_0) = \mathbb{E}_{1,\theta}\left[\log\frac{f_1\left(x_t,\widehat{\theta}_t\right)}{f_0(x_t)}\right] = D(f_{1,\theta}\|f_0) + O(1/w), \widehat{J}_0 = \mathbb{E}_{1,\theta}\left[\log^2\frac{f_1\left(x_t,\widehat{\theta}_t\right)}{f_0(x_t)}\right] = J_0 + O(1/w), \text{for } t \geq w+1.$$

The computation complexity of this window-limited CUSUM depends on the parametric family and the optimization method to obtain the estimator $\widehat{\theta}_t$. For the exponential family, $\widehat{\theta}_t$ can be updated recursively in $O(1)$ time from $\widehat{\theta}_{t-1}$ and for the general parametric family, this depends on the precision we want in solving optimization and the computation complexity may scale with window length w. Apart from this, the statistic $T_{t+1}^{\text{WL-CUSUM}}$ can be updated in $O(1)$ time using the usual CUSUM recursion. Either way, it has better computation complexity than window-limited GLR because the latter one still needs to search for all possible change points within the much bigger sliding window.

5 | E-DETECTORS: NONPARAMETRIC CHANGE-POINT DETECTION

Recently, Shin et al. (2022) discuss a general framework based on the CUSUM procedure, which can be applied to nonparametric and composite pre- and post-change distribution families $\mathscr{F}_0$ and $\mathscr{F}_1$, named *E-detectors*. A $\mathscr{F}_0$-E-detector is defined as a non-negative process $(M_t)_{t \geq 0}$ such that

$$\mathbb{E}_{f,0}[M_\tau] \leq \mathbb{E}_{f,0}[\tau], \forall f \in \mathscr{F}_0, \forall \tau \in \mathcal{T},$$

where $\mathbb{E}_{f,0}$ is the expectation when the pre-change distribution is f with no change-point, and $\mathcal{T}$ is the set of finite stopping times. It is easily seen by the optional stopping theorem that if $(M_t - t)_{t \geq 0}$ is a super-martingale under all probability measure in $\mathscr{F}_0$, then $(M_t)_{t \geq 0}$ must be an $\mathscr{F}_0$-E-detector. Let the stopping time

$$\tau^{\text{E}}(b) = \min\{t : M_t \geq \exp(b)\}.$$

Then the ARL is lower bounded by

$$\mathbb{E}_{f_0,0}\big[\tau^{\mathrm{E}}(b)\big] \geq \mathbb{E}_{f_0,0}\big[M_{\tau^{\mathrm{E}}(b)}\big] \geq \exp(b). \tag{16}$$

The classic CUSUM and SR procedures all satisfy the definition of E-detectors, and (16) provides a fast and easy way to lower bound the ARL, which is good enough for showing the asymptotic optimality of the procedure.

For the CUSUM procedure, the final E-detector used for detecting an unknown anomaly is a mixture of baseline E-detectors. Each baseline E-detector is constructed via *E-processes* $\left(M_t^{(k)}\right)_{t \geq 0}$, $k = 1, 2, 3, \ldots$, which are generalizations of the likelihood ratio. Each E-process is responsible for detecting a change point at k. An example of the most basic setup described in Section 2 is

$$M_t^{(k)} = \begin{cases} \prod_{i=k}^t \dfrac{f_1(x_i)}{f_0(x_i)} & t \geq k, \\ 1 & 1 \leq t < k. \end{cases} \tag{17}$$

In the non-parametric setup, the baseline E-process is in the form of

$$M_t^{(k)} = \begin{cases} \prod_{i=k}^t L_i & t \geq k, \\ 1 & 1 \leq t < k, \end{cases} \tag{18}$$

where the multiplicative increments L_i are non-negative and shared between all k to keep the recursive rule (4). L_i also satisfies for any $f_0 \in \mathcal{F}_0$, $\mathbb{E}_{f_0}[L_i|x_1, \ldots, x_{i-1}] \leq 1$. So the E-process $\left(M_t^{(k)}\right)_{t \geq 0}$ is a non-negative super-martingale under the null hypothesis. An example of the multiplicative increments L_i is by leveraging the concentration inequalities used to quantify the uncertainties in the (non-parametric) pre-change distribution f_0 (e.g., Howard et al. (2020)), which is in the form of

$$\mathbb{E}_{f_0}[\exp(\lambda s(x_i) - \varphi(\lambda)\nu(x_i))|x_1, \ldots, x_{i-1}] \leq 1, \ \ \forall \lambda \in \Lambda$$

for some known real function s, continuously differentiable convex function φ and non-negative function ν. Then the multiplicative increment is

$$L_i(\lambda) = \exp(\lambda s(x_i) - \varphi(\lambda)\nu(x_i)), \tag{19}$$

and the baseline E-detector is

$$M_t(\lambda) = \max_{1 \leq k \leq t} M_t^{(k)}(\lambda), t = 0, 1, 2, \ldots, \tag{20}$$

where λ remains to be chosen to optimize the EDD. To account for the unknown post-change distribution, a mixture of baseline E-detectors

$$M_t = \sum_{i=1}^K \omega_i M_t(\lambda_i), \sum_{i=1}^K \omega_i = 1$$

can be taken so that the EDD for the worst possible anomaly will be small. Since the set of E-detectors is closed under convex operations, the mixture will still be an E-detector, and the ARL is controlled by (16). Let $\mathbb{E}_{1,f_1}, \mathbb{V}_{1,f_1}$ represent the expectation and variance when the change starts with the first sample and the post-change distribution is f_1. Let $\left(L_1^{(\lambda)}\right)_{\lambda \in \Lambda}$ be the set of all possible baseline increments in an E-process, and λ^* be the parameter maximizing $\mathbb{E}_{1,f_1}\left[\log L_1^{(\lambda)}\right]$. Let $D(f_1 \| \mathcal{F}_0) = \mathbb{E}_{1,f_1}\left[\log L_1^{(\lambda^*)}\right]$ be the divergence between f_1 and $\mathcal{F}_0$. Let Δ_U, Δ_L be the upper and lower

bound of the gap $\mathbb{E}_{1f_1}[s(x_1)]/\mathbb{E}_{1f_1}[\nu(x_1)]$ between $\mathscr{F}_0$ and all possible f_1, and let φ^* be the convex conjugate of φ. For i.i.d. data and a properly chosen mixture, the EDD is upper bounded by the following.

Theorem 5. *(EDD of a proper mixture) (Shin et al., 2022, theorem 4.3)* For every $f_1 \in \mathscr{F}_1$,

$$\text{EDD}\big(\tau^{\text{E}}(b)\big) \leq \frac{g_b}{D(f_1\|\mathscr{F}_0)} + \frac{\mathbb{V}_{1f_1}\Big(\log L_1^{(\lambda^*)}\Big)}{D(f_1\|\mathscr{F}_0)^2} + 1,$$

where

$$g_b = \inf_{\eta > 1} \eta\left(b + \log\left(1 + \left\lceil \log_\eta \frac{\varphi^*(\Delta_U)}{\varphi^*(\Delta_L)} \right\rceil \right) \right),$$

given that there are sufficiently many baseline E-detectors in the mixture.

The first inequality is based on results in Lorden (1970). Note that $D(f_1\|\mathscr{F}_0)$ is no longer the KL divergence, as there are no probability density functions. It is instead related to the construction of the set of baseline increments, in this case, the functions s, ν, φ in (19) and Λ the set of possible λ. So are the gap bounds Δ_U, Δ_L. Here $D(f_1\|\mathscr{F}_0)$ will be smaller than the KL divergence between the true f_0 and the true f_1, thus this procedure is not asymptotically optimal in the sense of Theorem 1. We believe no non-parametric methods can match this theoretical lower bound on EDD, as a cost of admitting better model robustness.

The computation and memory complexity is constant for each baseline E-detector, the same as the classic CUSUM procedure. The number of baseline E-detectors depends on the threshold b for ARL control and the size of the gap between $\mathscr{F}_0$ and $\mathscr{F}_1$. For details, we refer to Shin et al. (2022).

6 | DISTRIBUTION-FREE CHANGE-POINT DETECTION PROCEDURES

Like has mentioned above, to our knowledge, existing distribution-free change-point detection procedures largely follow a similar strategy based on the Shewhart chart: scanning through different blocks of the data and finding the statistics $\{T_t\}_{t \geq 0}$ where a higher value suggests the distributions of the two blocks are different. Then the procedure declares a change-point when $T_t > b$ for some t. Such methods replace the log-likelihood ratio used in a parametric setup with alternative divergence, for example, the maximum mean discrepancy (MMD) (Li et al., 2019; Song & Chen, 2022), kernel-based Fisher discriminant ratio (Harchaoui et al., 2008), similarity graphs (Chen & Zhang, 2015; Chu & Chen, 2019), marginal ranking (Lung-Yut-Fong et al., 2015), interpoint distance (Matteson & James, 2014) to remove the strong assumption of known distribution family. Typically theoretical analysis for such procedures can be more challenging because of the lack of a handle through the likelihood ratio. Here we review a technique based on change-of-measure that can approximate the probability of extremely rare events leading to the ARL. We start from the analysis under the parametric setup (Section 6.1) and discuss its generalization to distribution-free change-point detection in Section 6.2. The complete and rigorous version is given in Yakir (2013). The EDD analysis remains an open question.

6.1 | ARL analysis via change-of-measure for extreme of random fields

In this section, we review the ARL analysis of the GLR in Siegmund and Venkatraman (1995). In this setup, the pre-change distribution $f_0 \sim \mathcal{N}(0,1)$ and the post-change distribution $f_1 \sim \mathcal{N}(\theta,1)$ where θ is unknown. The (window-limited) GLR statistic is in the form of

$$T_t^{\text{GLR}} = \max_{\min\{t-w,1\} \leq k \leq t} \frac{\left(\sum_{i=k}^t x_i\right)^2}{2(t-k+1)},$$

where for GLR, we use $w = \infty$; note that the maximum over θ part is absorbed into the equation using a plug-in maximum likelihood estimator.

For each (t,k), we denote by $Z_{t,k} = \left(\sum_{i=k}^{t} x_i\right)/\sqrt{t-k+1}$ and each follows $N(0,1)$ under the null distribution. This kind of analysis can be generalized to setups where the log-likelihood ratios do not follow Gaussian distribution and have a possibly continuous-valued index θ. Recall that a general change-point detection procedure is defined as the first time the detecting statistic exceeds threshold b, and in this case, we consider the one-sided stopping rule

$$\tau^{\text{GLR}} = \inf\left\{ t : \max_{\min\{t-w,1\} \le k \le t} Z_{t,k} \ge \sqrt{2b} \right\}.$$

6.1.1 | Poisson approximation

By examining the correlation between the $Z_{t,k}$s, it can be proved that τ^{GLR} asymptotically follows exponential distribution (Siegmund & Venkatraman, 1995; Yakir, 2013). Next we want to find out the mean by calculating $\mathbb{P}_0(\tau^{\text{GLR}} \le m)$ for $\log b < < m < < b^{-1/2} e^b$,

$$\mathbb{P}_0\left(\tau^{\text{GLR}} \le m\right) = \mathbb{P}_0\left(\max_{\min\{t-w,1\} \le k \le t \le m} Z_{t,k} \ge \sqrt{2b}\right). \tag{21}$$

And there would be $\text{ARL}(\tau^{\text{GLR}}) \sim m/\mathbb{P}_0(\tau^{\text{GLR}} \le m)$, where the symbol $\sim$ means the ratio between the two sides converges to 1.

Decomposition on the local field of each (t,k). Note that the random field $\{Z_{t,k}\}$ is highly correlated because it is computed using highly overlapping data segments. For convenience, we denote the set of all possible (t,k) by $S_{m,w} = \{(t,k) \in \mathbb{N}^2 : \min\{t-w,1\} \le k \le t \le m\}$ and the event $\{\max_{(t,k) \in S_{m,w}} Z_{t,k} \ge \sqrt{2b}\}$ by $A_{m,w}$. To analyze the tail probability on the right-hand-side of (21), we decompose it as the sum of the probability $Z_{t,k} \ge \sqrt{2b}$ for each $(t,k) \in S_{m,k}$ while also take into account their correlation. Let $\mathbf{1}$ be the indicator function. For any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{P}_0\left(\max_{(t,k) \in S_{m,w}} Z_{t,k} \ge \sqrt{2b}\right)$$
$$= \mathbb{E}_0\left[\frac{\sum_{(t,k) \in S_{m,w}} \exp(g(Z_{t,k}))}{\sum_{(t',k') \in S_{m,w}} \exp(g(Z_{t',k'}))} \mathbf{1}\{A_{m,w}\}\right]$$
$$= \sum_{(t,k) \in S_{m,w}} \mathbb{E}_0\left[\frac{\exp(g(Z_{t,k}))}{\sum_{(t',k') \in S_{m,w}} \exp(g(Z_{t',k'}))} \mathbf{1}\{A_{m,w}\}\right]$$

6.1.2 | Exponential tilting and change-of-measure

We use a change-of-measure technique similar to that has been used in Section 3, Equation (14) so that under the new probability measure, $\{Z_{t,k} \ge \sqrt{2b}\}$ happens with a much higher rate. In this setup for each (t,k) we change to the measure $\mathbb{P}_{t,k}$ and its expectation $\mathbb{E}_{t,k}$, under which $x_k, x_{k+1}, \ldots, x_t \sim \mathcal{N}(\sqrt{2b}/\sqrt{t-k+1}, 1)$ while the rest of the samples still follow $\mathcal{N}(0,1)$. So now $Z_{t,k} \sim \mathcal{N}(\sqrt{2b}, 1)$ under $\mathbb{P}_{t,k}$, and the event $A_{m,w}$ happens with high probability. Let $g(Z_{t,k}) = \sqrt{2b} Z_{t,k} - b$ be the log-likelihood ratio between $\mathbb{P}_{t,k}$ and $\mathbb{P}_0$. When $Z_{t,k}$ is not normally distributed, exponential tilting is done by letting $g(z) = \gamma z - \varphi(\gamma)$, where φ is the log moment generating function of $Z_{t,k}$ under the null distribution. Then $\mathbb{E}_0[\exp(g(Z_{t,k}))] = 1$ and it can be regarded as the likelihood ratio between $\mathbb{P}_0$ and some distribution $\mathbb{P}_{t,k}$. By choosing a proper γ such that $\mathbb{E}_{t,k}[Z_{t,k}]$ equals to the threshold b, $A_{m,w}$ happens with high probability.

$$\mathbb{P}_0\left(\max_{(t,k) \in S_{m,w}} Z_{t,k} \ge \sqrt{2b}\right)$$
$$= \sum_{(t,k) \in S_{m,w}} \mathbb{E}_{t,k}\left[\frac{1}{\sum_{(t',k') \in S_{m,w}} \exp(\sqrt{2b} Z_{t',k'} - b)} \mathbf{1}\{A_{m,w}\}\right].$$

Next step, we decompose the expectation into a local and a global term which are "asymptotically independent," the meaning of which will be explained later.

$$\sum_{(t,k) \in S_{m,w}} \mathbb{E}_{t,k} \left[\frac{1}{\sum_{(t',k') \in S_{m,w}} \exp\left(\sqrt{2b}Z_{t',k'} - b\right)} \mathbf{1}\{A_{m,w}\} \right]$$

$$= \sum_{(t,k) \in S_{m,w}} \mathbb{E}_{t,k} \left[\frac{\max_{(t',k') \in S_{m,w}} \exp\left(\sqrt{2b}(Z_{t',k'} - Z_{t,k})\right)}{\sum_{(t',k') \in S_{m,w}} \exp\left(\sqrt{2b}(Z_{t',k'} - Z_{t,k})\right)} \exp\left(-\max_{(t',k') \in S_{m,w}} \left(\sqrt{2b}Z_{t',k'} - b\right)\right) \mathbf{1}\{A_{m,w}\} \right].$$

For convenience, let

$$S_{t,k} = \sum_{(t',k') \in S_{m,w}} \exp\left(\sqrt{2b}(Z_{t',k'} - Z_{t,k})\right), \quad M_{t,k} = \max_{(t',k') \in S_{m,w}} \exp\left(\sqrt{2b}(Z_{t',k'} - Z_{t,k})\right).$$

Now after replacing certain terms with $S_{t,k}$ and $M_{t,k}$,

$$\mathbb{P}(A_{m,w}) = \sum_{(t,k) \in S_{m,w}} \mathbb{E}_{t,k} \left[\frac{M_{t,k}}{S_{t,k}} \exp\left(-\left(\log M_{t,k} + \sqrt{2b}Z_{t,k} - b\right)\right) \mathbf{1}\left\{\log M_{t,k} + \sqrt{2b}Z_{t,k} - b \geq b\right\} \right]. \tag{22}$$

Next we discuss the expectation as

$$b \to \infty, t - k \to \infty, \text{and } b/(t - k + 1) \to c \text{ for some constant } c. \tag{23}$$

6.1.3 | The local term

The first term $M_{t,k}/S_{t,k}$ in (22) for every $(t,k) \in S_{m,w}$, or the tuple $(M_{t,k}, S_{t,k})$, is a local random variable meaning it can be well approximated with high probability using values in the random field $\left(Z_{t',k'}\right)_{(t',k') \in S_{m,w}}$ which index is constantly close to (t,k). And $(M_{t,k}, S_{t,k})$ converge in distribution to some local random variables which we will define later. Since it is a Gaussian random field with mean 0 under $\mathbb{P}_0$, the local properties are decided by the covariance under the asymptotic regime (23). For fixed $i, j, i', j' \in \mathbb{Z}$, it can be verified that

$$\text{Cov}_0\left[Z_{t+i,k-j}, Z_{t+i',k-j'}\right] = 1 - \frac{c}{2b}|i - i'| - \frac{c}{2b}|j - j'| + o\left(b^{-1}\right). \tag{24}$$

And consequently,

$$\mathbb{E}_{t,k}\left[\sqrt{2b}(Z_{t+i,k-j} - Z_{t,k})\right] = -c|i| - c|j| + o(1),$$

$$\text{Cov}_{t,k}\left[\sqrt{2b}(Z_{t+i,k-j} - Z_{t,k}), \sqrt{2b}(Z_{t+i',k-j'} - Z_{t,k})\right] = 2c\mathbf{1}\{ii' > 0\} \min\{|i|, |i'|\} + 2c\mathbf{1}\{jj' > 0\} \min\{|j|, |j'|\}.$$

The local random field $\left\{\sqrt{2b}(Z_{t+i,k-j} - Z_{t,k})\right\}_{i,j}$ under $\mathbb{P}_{t,k}$ converge in distribution to

$$\left\{Y_i + Y'_j\right\}_{i,j}, \tag{25}$$

where $\{Y_i\}_{i \in \mathbb{Z}}$ is a two-sided Gaussian random walk with a negative drift c and variance $2c$, that is,

$$\mathbb{E}[Y_i] = -c\,|\,i\,|, \quad \mathrm{Cov}[Y_i, Y_{i'}] = 2c\mathbf{1}\{ii' > 0\}\min\{|i|, |i'|\}.$$

$\left\{Y'_j\right\}_{j \in \mathbb{Z}}$ is an i.i.d. copy of $\{Y_i\}_{i \in \mathbb{Z}}$. Let

$$\mathscr{M}_c = \max_{i \in \mathbb{Z}} \exp(Y_i), \quad \mathscr{S}_c = \sum_{i \in \mathbb{Z}} \exp(Y_i),$$

The expectation $\mathbb{E}[\mathscr{M}_c/\mathscr{S}_c]$ is called the Mill's ratio (see chapter 2.2 in Yakir (2013)) and it has the following form.

$$\mathbb{E}[\mathscr{M}_c/\mathscr{S}_c] = c\nu\left(\sqrt{2c}\right), \tag{26}$$

where the special function

$$\nu(x) = 2x^{-2}\exp\left(-2\sum_{i=1}^{\infty} i^{-1}\Phi\left(-xi^{1/2}/2\right)\right),$$

is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A numerical approximation to the Mill's ratio is

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2) - 0.5)}{(x/2)\Phi(x/2) + \phi(x/2)},$$

here ϕ, Φ is the p.d.f. and c.d.f. of the standard normal distribution.

Back to the GLR analysis, it can be shown that $(M_{t,k}, S_{t,k})$ converge in distribution to $\left(\mathscr{M}_c\mathscr{M}'_c, \mathscr{S}_c\mathscr{S}'_c\right)$ where $\left(\mathscr{M}'_c, \mathscr{S}'_c\right)$ is an i.i.d. copy of $(\mathscr{M}_c, \mathscr{S}_c)$. This is because our local random field is two-dimensional and asymptotically we can break it down using (25) into the sum of two independent Gaussian random walks. Using the Mill's ratio, the expectation $\mathbb{E}_{t,k}[M_{t,k}/S_{t,k}]$ over the two-dimensional random field is

$$\mathbb{E}_{t,k}[M_{t,k}/S_{t,k}] \to \mathbb{E}[\mathscr{M}_c/\mathscr{S}_c]^2 = c^2\nu^2\left(\sqrt{2c}\right).$$

6.1.4 | The global term

The second term $\exp\left(-\left(\log M_{t,k} + \sqrt{2b}Z_{t,k} - b\right)\right)\mathbf{1}\left\{\log M_{t,k} + \sqrt{2b}Z_{t,k} - b \geq b\right\}$ in (22) for every $(t,k) \in S_{m,w}$ is considered a global one because asymptotically it is not relevant with the constantly many observations around time t and k, but only those in between. Under the asymptotic regime (23), $\sqrt{2b}Z_{t,k} - b \sim \mathcal{N}(b, 2b)$ but the distribution of $M_{t,k}$ converge to $\mathscr{M}_c\mathscr{M}'_c$ which does not scale with b. It can be shown that the distribution of $\log M_{t,k} + \sqrt{2b}Z_{t,k} - b$ together is dominated by $\sqrt{2b}Z_{t,k} - b$ and it is so close to $\mathcal{N}(b, 2b)$ that for each (t,k) in (22) the global term can be treated as such and is "independent" of the local term. And the expectation of the global term converges to

$$\mathbb{E}_{t,k}\left[\exp\left(-\left(\sqrt{2b}Z_{t,k} - b\right)\right)\mathbf{1}\left\{\sqrt{2b}Z_{t,k} - b \geq b\right\}\right] \to \int_b^{\infty} \frac{1}{\sqrt{4\pi b}}\exp(-x)dx = e^{-b}/\sqrt{4\pi b}$$

because the probability density function of $\sqrt{2b}Z_{t,k} - b$ is almost always $1/\sqrt{4\pi b}$ around b.

Together, we know that asymptotically the expectation in (22) is

$$\mathbb{E}_{t,k}\left[\frac{M_{t,k}}{S_{t,k}}\exp\left(-\left(\log M_{t,k} + \sqrt{2b}Z_{t,k} - b\right)\right)\mathbf{1}\left\{\log M_{t,k} + \sqrt{2b}Z_{t,k} - b \geq b\right\}\right] \sim \frac{e^{-b}}{\sqrt{4\pi b}}c^2\nu^2\left(\sqrt{2c}\right). \tag{27}$$

6.1.5 | The ARL

After checking several good conditions on $\mathbb{E}_{t,k}[\cdot]$ (e.g., uniform integrability), we can now calculate the probability $\mathbb{P}(A_{m,w})$ by substituting the summation with integral. Assume as $b \to \infty$, the window length w satisfies $b/w \to C$, and $m >> w$,

$$\mathbb{P}(A_{m,w}) \sim \frac{e^{-b}}{\sqrt{4\pi b}} \sum_{(t,k) \in S_{m,w}} \mathbb{E}\left[\mathscr{M}_{b/(t-k+1)}/\mathcal{S}_{b/t-k+1}\right]^2$$

$$\sim \frac{e^{-b}}{\sqrt{4\pi b}} b \int_0^{C^{-1}} m\mathbb{E}[\mathscr{M}_{x^{-1}}/\mathcal{S}_{x^{-1}}]^2 dx,$$

where we take $x = (t-k+1)/b$ when changing to integral. By Poisson approximation,

$$\mathrm{ARL}\left(\tau^{\mathrm{GLR}}\right) \sim \sqrt{4\pi} b^{-0.5} e^b \left(\int_0^{C^{-1}} x^{-2} \nu^2\left(\sqrt{2/x}\right) dx\right)^{-1}.$$

6.2 | Scan statistic via kernel-based MMD

In (Li et al., 2019), the authors propose to use the unbiased kernel-based maximum mean divergence (MMD) statistic to measure the difference between two blocks of samples. For $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$, the unbiased MMD is

$$\mathrm{MMD}(X, Y) = \frac{1}{n(n-1)} \sum_{i \neq j} \left(k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(y_i, x_j)\right),$$

where k is a kernel in the reproducing kernel Hilbert space. The expectation of such measure is 0 only when samples in X and Y share the same distribution; otherwise, the expectation is strictly positive. Assuming the abundance of reference data, at every time t the average unbiased MMD between the most recent B samples and N reference blocks is computed, and then normalized to get the test statistic T_t^{MMD}:

$$Z_t = \frac{1}{N} \sum_{i=1}^N \mathrm{MMD}_u(X_t, Y_{t,i}), \qquad T_t^{\mathrm{MMD}} = \frac{Z_t}{\mathbb{V}_0(Z_t)^{1/2}}.$$

Here X_t is the collection of the most recent B samples, and the reference blocks $Y_{t,i}, i = 1, ..., N$ is formed by taking NB samples without replacement from reference data. The variance $\mathbb{V}_0(Z_t)^{1/2}$ can be pre-computed from reference data. And procedure raises an alarm at $\tau^{\mathrm{MMD}} = \min\{t : T_t^{\mathrm{MMD}} > b\}$.

During each update, the statistic T_{t+1}^{MMD} can be computed recursively from T_t^{MMD}. This is because the reference block $Y_{t+1,i}$ is formed by adding a new sample from reference data to $Y_{t,i}$ and removing the oldest one, for $i = 1, ..., N$. X_{t+1} is formed using the same way. For convenience, let

$$h(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') + k(y, x'),$$

and let $Y_{t,i} = (y_{t-B+1,i}, ..., y_{t,i})$. for every $i = 1, ..., N$. Then T_{t+1}^{MMD} can be updated from T_t^{MMD} by adding and subtracting the values $h(x_s, x_{t+1}, y_{s,i}, y_{t+1,i}), h(x_s, x_{t-B+2}, y_{s,i}, y_{t-B+2,i}) s = t - B + 2, ..., t, i = 1, ..., N$. This leads to a computation and memory complexity of $O(NB)$.

The ARL analysis of such statistics follows the example given in Section 6.1. The covariance between the statistics $\{T_t\}_{t \geq 0}$ is provided in a close form and can be computed easily from reference data. With a proper asymptotic regime,

the local covariance can be found in the form of (24). The rest of the analysis carries through as if $\{T_t\}_{t \geq 0}$ follows Gaussian distribution.

Of course, the above analysis can be less accurate when $\{Z_t\}_{t \geq 0}$ does not converge to a Gaussian distribution. Skewness correction can be performed to account for the non-zero third-order moment. This method first appears in Tu and Siegmund (1999) and then is modified in Tang and Siegmund (2001) and becomes useful in analyzing the ARL or false alarm rate in many change-point detection procedures based on scanning statistics.

7 | CONCLUSION

In this article, we review the classic procedures in sequential change-point detection featuring the computation versus performance trade-off and the classic performance analysis methods, which are still quite useful in analyzing new change-point detection procedures and other problems (for details, we refer to Yakir, 2013). We also discuss several recent articles covering computation and model robustness considerations. For completeness, a longer survey on the extensions and modern applications of sequential change-point detection can be found in Xie et al. (2021). Finally, another important direction we have not mentioned is machine learning algorithm-based sequential change-point detection to handle situations such as missing data (Londschien et al., 2021; Xie et al., 2012), the existence of outliers (Fearnhead & Rigaill, 2019), and cost-sensitive active sampling (Gundersen et al., 2021).

AUTHOR CONTRIBUTIONS
Haoyun Wang: Writing – original draft (equal); writing – review and editing (equal). **Yao Xie:** Conceptualization (lead); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

CONFLICT OF INTEREST STATEMENT
The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT
Not available.

ORCID
Yao Xie https://orcid.org/0000-0001-6777-2951

RELATED WIREs ARTICLES
Statistical methods for complex event processing and real time decision making
Energy distance

REFERENCES
Abbasi, S., & Haq, A. (2019). Optimal CUSUM and adaptive CUSUM charts with auxiliary information for process mean. *Journal of Statistical Computation and Simulation*, *89*, 337–361.

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: Theory and application* (Vol. 104). Prentice-Hall Englewood Cliffs.

Cao, Y., Xie, L., Xie, Y., & Xu, H. (2018). Sequential change-point detection via online convex optimization. *Entropy*, *20*, 108.

Casella, G., & Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Chen, H., & Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, *43*, 139–176.

Chen, J., Kim, S.-H., & Xie, Y. (2020). S3t: An efficient score-statistic for spatio-temporal surveillance. *Sequential Analysis*, *39*, 563–592.

Chen, Y., Wang, T., & Samworth, R. J. (2022). High-dimensional, multiscale online changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*, 234–266.

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

Chu, L., & Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics*, *47*, 382–414.

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., & Priesemann, V. (2020). Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, *369*, eabb9789.

Fearnhead, P., & Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, *114*, 169–183.

Gundersen, G. W., Cai, D., Zhou, C., Engelhardt, B. E., & Adams, R. P. (2021). Active multi-fidelity Bayesian online changepoint detection. *arXiv Preprints*, 1916–1926.

Hallac, D., Leskovec, J., & Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 387–396.

Harchaoui, Z., Moulines, E., & Bach, F. (2008). Kernel change-point analysis. *Advances in Neural Information Processing Systems*, *21*, 1–8.

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, *17*, 257–317.

Krieger, A. M., Pollak, M., & Yakir, B. (1999). Detecting a change in regression: First-order optimality. *The Annals of Statistics*, *27*, 1896–1913.

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

Lai, T. L., & Shan, J. Z. (1999). Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, *44*, 952–966.

Lakhina, A., Crovella, M., & Diot, C. (2004). Diagnosing network-wide traffic anomalies. *ACM SIGCOMM Computer Communication Review*, *34*, 219–230.

Lee, K.-C., & Kriegman, D. (2005). Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 852–859.

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

Londschien, M., Kovács, S., & Bühlmann, P. (2021). Change-point detection for graphical models in the presence of missing values. *Journal of Computational and Graphical Statistics*, *30*, 768–779.

Lorden, G. (1970). On excess over the boundary. *The Annals of Mathematical Statistics*, *41*, 520–527.

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

Lorden, G., & Eisenberger, I. (1973). Detection of failure rate increases. *Technometrics*, *15*, 167–175.

Lorden, G., & Pollak, M. (2005). Non-anticipating estimation applied to sequential analysis and changepoint detection. *The Annals of Statistics*, *33*, 1422–1454.

Lucas, J. M. (1982). Combined Shewhart-CUSUM quality control schemes. *Journal of Quality Technology*, *14*, 51–59.

Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, *156*, 133–162.

Magesh, A., Veeravalli, V. V., Roy, A., & Jha, S. (2022). Multiple testing framework for out-of-distribution detection. *arXiv Preprint*.

Maillard, O.-A. (2019). Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In A. Garivier & S. Kale (Eds.), *Proceedings of the 30th international conference on algorithmic learning theory* (Vol. 98, pp. 610–632). Proceedings of Machine Learning Research.

Malladi, R., Kalamangalam, G. P., & Aazhang, B. (2013). Online Bayesian change point detection algorithms for segmentation of epileptic activity. In 2013 Asilomar conference on signals, systems and computers, pp. 1833–1837.

Matteson, D. S., & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, *109*, 334–345.

Mei, Y. (2006). Sequential change-point detection when unknown parameters are present in the pre-change distribution. *The Annals of Statistics*, *34*, 92–122.

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

Moustakides, G. V. (2014). Multiple optimality properties of the Shewhart test. *Sequential Analysis*, *33*, 318–344.

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

Pavlov, I. V. (1991). Sequential procedure of testing composite hypotheses with applications to the Kiefer-Weiss problem. *Theory of Probability & its Applications*, *35*, 280–292. https://doi.org/10.1137/1135036

Peel, L., & Clauset, A. (2015). Detecting change points in the large-scale structure of evolving networks. In Twenty-ninth AAAI conference on artificial intelligence, 29.

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

Pollak, M., & Krieger, A. M. (2013). Shewhart revisited. *Sequential Analysis*, *32*, 230–242. https://doi.org/10.1080/07474946.2013.774621

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

Pollak, M., & Tartakovsky, A. G. (2009). Optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica*, *19*(4), 1729–1739.

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

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

Raghavan, V., & Veeravalli, V. V. (2010). Quickest change detection of a Markov process across a sensor array. *IEEE Transactions on Information Theory*, *56*, 1961–1981.

Raginsky, M., Willett, R. M., Horn, C., Silva, J., & Marcia, R. F. (2012). Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, *58*, 5544–5562.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, *32*, 1–12.

Robbins, H., & Siegmund, D. (1972). A class of stopping rules for testing parametric hypotheses. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol 4, pp. 37–41.

Robbins, H., & Siegmund, D. (1974). The expected sample size of some tests of power one. *The Annals of Statistics*, *2*, 415–436.

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

Romano, G., Eckley, I. A., Fearnhead, P., & Rigaill, G. (2023). Fast online changepoint detection via functional pruning CUSUM statistics. *Journal of Machine Learning Research*, *24*, 1–36.

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

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Macmillan and Co Ltd.

Shin, J., Ramdas, A., & Rinaldo, A. (2022). E-detectors: A nonparametric framework for online changepoint detection. *arXiv Preprint*.

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

Siegmund, D. (1985). *Sequential analysis: Tests and confidence intervals*. Springer Science & Business Media.

Siegmund, D., & Venkatraman, E. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, *23*, 255–271.

Song, H., & Chen, H. (2022). New kernel-based change-point detection. *arXiv Preprint*.

Sparks, R. S. (2000). CUSUM charts for signalling varying location shifts. *Journal of Quality Technology*, *32*(2), 157–171.

Tang, H.-K., & Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, *2*, 147–162.

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

Tartakovsky, A. G. (2005). Asymptotic performance of a multichart CUSUM test under false alarm probability constraint. In Proceedings of the 44th IEEE conference on decision and control, pp. 320–325.

Tartakovsky, A. G. (2014). Nearly optimal sequential tests of composite hypotheses revisited. *Proceedings of the Steklov Institute of Mathematics*, *287*, 268–288.

Tartakovsky, A. G. (2019). *Sequential change detection and hypothesis testing: General non-iid stochastic models and asymptotically optimal rules*. Chapman and Hall/CRC.

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

Tartakovsky, A. G., Polunchenko, A. S., & Sokolov, G. (2012). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, *7*, 4–11.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299.

Tu, I.-P., & Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis. *Advances in Applied Probability*, *31*, 510–531.

Ward, K., Dilillo, G., Eckley, I., & Fearnhead, P. (2022). Poisson-FOCuS: An efficient online method for detecting count bursts with application to gamma ray burst detection. *arXiv Preprint*.

Willsky, A., & Jones, H. (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, *21*, 108–112.

Xie, L., Moustakides, G. V., & Xie, Y. (2022). Window-limited CUSUM for sequential change detection. *arXiv Preprint*, *2206*, 06777.

Xie, L., Xie, Y., & Moustakides, G. V. (2020). Sequential subspace change point detection. *Sequential Analysis*, *39*, 307–335.

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

Xie, Y., Huang, J., & Willett, R. (2012). Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, *7*, 12–27.

Xie, Y., & Siegmund, D. (2013). Sequential multi-sensor change-point detection. In 2013 Information theory and applications workshop (ITA), pp. 1–20.

Yakir, B. (2013). *Extremes in random fields: A theory and its applications*. John Wiley & Sons.

Yu, Y., Padilla, O. H. M., Wang, D., & Rinaldo, A. (2020). A note on online change point detection. *arXiv Preprint*.