Quickest Anomaly Detection in Sensor Networks With Unlabeled Samples

Zhongchang Sun and Shaofeng Zou

***Abstract*—The problem of quickest anomaly detection in networks with unlabeled samples is studied. At some unknown time, an anomaly emerges in the network and changes the data-generating distribution of some unknown sensor. The data vector received by the fusion center at each time step undergoes some unknown and arbitrary permutation of its entries (unlabeled samples). The goal of the fusion center is to detect the anomaly with minimal detection delay subject to false alarm constraints. With unlabeled samples, existing approaches that combines local cumulative sum (CuSum) statistics cannot be used anymore. Several major questions include whether detection is still possible without the label information, if so, what is the fundamental limit and how to achieve that. Two cases with static and dynamic anomaly are investigated, where the sensor affected by the anomaly may or may not change with time. For the two cases, practical algorithms based on the ideas of mixture likelihood ratio and/or maximum likelihood estimate are constructed. Their average detection delays and false alarm rates are theoretically characterized. Universal lower bounds on the average detection delay for a given false alarm rate are also derived, which further demonstrate the asymptotic optimality of the two algorithms.**

***Index Terms*—Quickest change detection, unlabeled samples, permuted samples, asymptotically optimal, fundamental limits.**

I. Introduction

IN SENSOR networks, samples may lack label information such as identity due to, e.g., malicious attacks and limited communication resources. For example, wireless ad-hoc sensor networks are usually vulnerable to spoofing attacks [2], and samples received by the fusion center may then lose their label information. Furthermore, in Internet-of-things (IoT) networks, where devices are commonly small and low-cost sensing devices powered by battery with limited communication bandwidth, and are usually deployed in a massive scale, the communication overhead of identifying individual sensors increases drastically with the number of sensors [3]. However, these battery-powered IoT devices are usually expected to survive for years without battery change. In this case, message that is delivered to the fusion center may be constrained not to contain the identity information. Furthermore, in social sensing applications, participants may choose to be anonymous in order to protect privacy, i.e., sharing the data without including identity information. Motivated by these applications, there is a recent surge of interest in the problem of signal processing with unlabeled data (see e.g., [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]), which refers to various signal processing problems where the data vector undergoes an unknown permutation of its entries, and the original position of each datum in the vector is unknown.

In this paper, we investigate the problem of quickest anomaly detection in sensor networks with unlabeled samples. Specifically, at some unknown time, an anomaly emerges in the network and leads to a change in the data-generating distribution of some unknown sensor. The fusion center sequentially receives unlabeled (arbitrarily permuted) samples from all the sensors at each time step. The goal of the fusion center is to detect the anomaly as quickly as possible, subject to false alarm constraints. This problem is of particular relevance to applications where an anomaly affects some sensor in the network, and the affected sensor may change over time [18], e.g., surveillance system, intrusion detection, environmental change detection, rumor detection, and seismic wave detection.

A. Contributions and Major Challenges

We first focus on the static anomaly, where the sensor affected by the anomaly does not change with time, but which sensor is affected is still unknown. We consider the detection delay under the worst-case affected sensor. The goal here is to minimize the detection delay subject to false alarm constraints. The major challenges here are two-fold. First of all, the labels of the samples are unknown and time-varying. Second, even if the labels are known, i.e., each sample is associated with its sensor, the sensor the anomaly affects is still unknown. For this problem, we construct a generalized mixture CuSum (GM-CuSum) algorithm. The basic idea is to estimate the unknown identity of the affected sensor using the maximum likelihood estimate (MLE), and further employ a mixture likelihood w.r.t. all possible labels. We prove that the GM-CuSum is second-order asymptotically optimal.[1]

Manuscript received 3 September 2022; revised 4 January 2023; accepted 6 March 2023. Date of publication 17 March 2023; date of current version 31 March 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ali Tajer. The work of Zhongchang Sun and Shaofeng Zou was supported by the National Science Foundation under Grants 1948165 and 2112693. This paper was presented in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9517771]. *(Corresponding author: Shaofeng Zou.)*

The authors are with the Department of Electrical Engineering, University at Buffalo, Buffalo, NY 14228 USA (e-mail: zhongcha@buffalo.edu; szou3@buffalo.edu).

Digital Object Identifier 10.1109/TSP.2023.3256275

[1] An algorithm is second-order asymptotically optimal if as the false alarm rate goes to zero, its detection delay is within an $O(1)$ term of the best possible detection delay.

We then focus on a general and more challenging setting with dynamic anomaly, where the sensor affected by the anomaly changes with time. Here, we refer to the sequence of sensors affected by the anomaly over time as the trajectory of the anomaly. We consider the detection delay under the worst-case trajectory. Compared to the static setting, the additional challenge is that the affected sensor changes with time, and thus the change is not persistent at any particular sensor. Therefore, estimating the identity of the affected sensor over time is not applicable. We then propose a weighted approach to address this challenge, and find the optimal weight to construct a weighted mixture CuSum algorithm. We prove that the weighted mixture CuSum algorithm is first-order asymptotically optimal.[2] We also discuss two computationally efficient approximations for large-scale networks.

We also conduct extensive numerical experiments to demonstrate the performance of our proposed algorithms. The numerical results show that for the static setting, our GM-CuSum algorithm outperforms a heuristic uniformly weighted mixture CuSum algorithm; the optimal weighted mixture CuSum algorithm also performs well for the static setting; and for the dynamic setting, our optimal weighted mixture CuSum algorithm outperforms an uniformly weighted one and the GM-CuSum algorithm. These numerical results validate our theoretical optimality results.

B. Related Work

The quickest change detection (QCD) problem in sensor networks with labeled samples was extensively studied in the literature, e.g., [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], where the fusion center knows the identity of each sample, i.e., knows which sensor that each sample is from. Therefore, one CuSum algorithm can be implemented at each sensor and then be combined to make the decision. This type of algorithms were shown to be asymptotically optimal for various settings. In this paper, we investigate the setting with unlabeled samples, where at each time step samples are arbitrarily permuted, and the permutation is time-varying. The fusion center does not know which sensor each sample comes from, and then cannot implement a CuSum algorithm for each sensor.

Various learning and inference problems with unlabeled data have been studied in the literature [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], which mainly focus on the offline setting with non-sequential data. Here we only review several closely related ones on detection problems. In [6], hypothesis testing with unlabeled samples was studied, where two practical algorithms, the unlabeled log-likelihood ratio test and the generalized likelihood ratio test were proposed. A more specific problem was studied in [7] where samples follow the Bernoulli distribution and an approximated log-likelihood test based on the central limit theorem was proposed. In [4], the binary hypothesis testing problem with unlabeled samples was studied, and an optimal mixture likelihood ratio test (MLRT) was developed. In [5], the bandwidth-constrained QCD problem with unlabeled samples was investigated, where each sensor sends 1-bit quantized feedback to the fusion center. In [17], the QCD problem with unlabeled samples was studied where the change affects all the sensors simultaneously. In this paper, we investigate a practical scenario where an anomaly may not affect all the sensors, which is of particular interest in the distributed setting, and the anomaly may also be dynamic and affect different sensors at different times, e.g., a moving target in surveillance systems.

Existing studies of quickly detecting a dynamic change mostly focus on the labeled setting, e.g., [31], [32], [33]. Our problem is similar to the one in [33] but we focus on the unlabeled setting. Our major challenge is due to the additional ambiguity of unknown labels. The QCD problem with a slowly changing post-change distribution was studied in [34], [35], whereas in this paper, the anomaly can move arbitrarily fast.

With unlabeled samples, our problem is also related to the composite QCD problem with unknown pre- and post-change parameters e.g., [22], [36], [37], [38]. Our work is different from the existing literature since the unknown parameters, i.e., the identity and the label of the affected sensor, are time-varying. Therefore, the generalized likelihood approach which estimates the unknown parameters using their MLEs may not perform well. Moreover, unlike studies in [36], [37], [38] where the distributions are assumed to belong to the exponential family, we do not have any assumptions on the distributions.

II. Problem Formulation

Consider a network monitored in real time by a set of n heterogeneous sensors. These sensors can be clustered into K types, and each type k has n_k sensors, $1 \le k \le K$. The data-generating distributions of samples from type k sensors are denoted by $p_{\theta,k}$, $\theta \in \{0, 1\}$, which are known to the fusion center. At some unknown time ν, an anomaly emerges in the network, and changes the data-generating distribution of one sensor. The fusion center does not know which type of sensor is affected. If a sensor of type k is affected by the anomaly, then its samples are generated by $p_{1,k}$, otherwise, by $p_{0,k}$. The goal is to detect the anomaly as quickly as possible subject to false alarm constraints. We focus on the case with unlabeled samples, where the data vector at each time step undergoes an unknown permutation of its entries, and the original position of each datum in the vector is unknown to the fusion center. In other words, the fusion center does not know which type of sensors that each sample comes from, and therefore, does not know the sample's exact data-generating distribution.

Based on whether the sensor is affected by the anomaly and the type of the sensor, we rearrange the sensors into $2K$ groups. The first K groups consists of sensors that are not affected by the anomaly; and the second K groups consists of sensors that are affected by the anomaly. Specifically, for sensors in group $1 \le k \le K$, their samples are generated by $p_{0,k}$, and for sensors in group $K < k \le 2K$, their samples are generated by $p_{1,k-K}$.

[2] An algorithm is first-order asymptotically optimal if as the false alarm rate goes to zero, the ratio between its detection delay and the best possible detection delay goes to 1.

In this paper, we use capital letters to denote random variables and lower case letters to denote their realizations. Denote by $X^n[t] = \{X_1[t], \ldots, X_n[t]\}$ the n arbitrarily permuted samples at time t received by the fusion center. We assume that $X_1[t], \ldots, X_n[t]$ are independent, and $X^n[t_1]$ is independent from $X^n[t_2]$ for any $t_1 \neq t_2$. Note that $X_i[t]$ is not necessarily the sample from sensor i since samples are permuted/unlabeled. The results in our paper hold for both continuous and discrete random variables $X_i[t]$.

Let $\mathcal{K} = \{1, 2, \ldots, K\}$. Denote by $s[t] \in \mathcal{K} \cup \{0\}$ the type of the affected sensor at t. For notational convenience, we use $s[t] = 0$ to denote the case when there is no anomaly, i.e., $t < \nu$. Let $\boldsymbol{s} \triangleq \{s[t]\}_{t=1}^{\infty}$ denote the trajectory of the anomaly. Here $\boldsymbol{s}$ is *unknown* to the decision maker. Even if the trajectory of the anomaly $\boldsymbol{s}$ is given, the distribution of $X^n[t]$ still cannot be fully specified due to lack of label information. To characterize the distribution of $X^n[t]$, we define a label function $\sigma_t^{s[t]} : \{1, \ldots, n\} \to \{1, \ldots, K, s[t] + K\}$. This function associates sample $X_i[t]$, $1 \leq i \leq n$, to group j for some $j \in \{1, 2, \ldots, K, K + s[t]\}$, i.e., specifies the probability distribution of $X_i[t]$. Specifically, if $\sigma_t^{s[t]}(i) = j$, then

$$X_i[t] \sim \begin{cases} p_{0,j}, & \text{if } 1 \leq j \leq K, \\ p_{1,j-K}, & \text{if } K < j \leq 2K. \end{cases} \tag{1}$$

Here $\sigma_t^{s[t]}$ can be interpreted as the inverse of the permutation applied to the data vector. We further note that $\sigma_t^{s[t]}$ is *unknown* to the decision maker.

Let $\Omega_{\boldsymbol{s}} = \left\{\sigma_1^{s[1]}, \ldots, \sigma_\infty^{s[\infty]}\right\}$ be the labels when the trajectory of the anomaly is $\boldsymbol{s}$, which is unknown. Let $\mathbb{P}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}$ and $\mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}$ denote the probability measure and the corresponding expectation when the change point is at ν and the samples received by the fusion center are permuted according to the label $\Omega_{\boldsymbol{s}}$ (see Appendix A for more details). We further let $\mathbb{P}_{\Omega}^{\infty}$ and $\mathbb{E}_{\Omega}^{\infty}$ denote the probability measure and the corresponding expectation when there is no change, i.e., $\nu = \infty$, where $\Omega = \Omega_{\boldsymbol{s}}$ with $s[t] = 0, \forall t \geq 1$.

We extend Lorden's criterion [39], and define the worst-case average detection delay (WADD) and the worst-case average running length (WARL) for any stopping time τ:

$$\text{WADD}(\tau) = \sup_{\nu \geq 1} \sup_{\boldsymbol{s}} \sup_{\Omega_{\boldsymbol{s}}} \text{esssup} \mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}[(\tau - \nu)^+ | \mathbf{X}^n[1, \nu - 1]],$$

$$\text{WARL}(\tau) = \inf_{\Omega} \mathbb{E}_{\Omega}^{\infty}[\tau], \tag{2}$$

where $\mathbf{X}^n[t_1, t_2] = \{X^n[t_1], \ldots, X^n[t_2]\}$, for any $t_1 \leq t_2$ and $(\tau - \nu)^+ = \max\{\tau - \nu, 0\}$. Let $f : X \to \mathbb{R}$ be a real-valued function and $(X, \mathcal{X}, \mu)$ be a probability space. The essential supremum is then defined as

$$\text{esssup} f = \inf\{a \in \mathbb{R} : \mu(\{x : f(x) > a\}) = 0\}. \tag{3}$$

The goal is to design a stopping rule that minimizes the WADD subject to a constraint on the WARL:

$$\inf_{\tau : \text{WARL}(\tau) \geq \gamma} \text{WADD}(\tau), \tag{4}$$

TABLE I
SUMMARY OF NOTATIONS

Symbol	Definition
n	Number of sensors
K	Number of different types
n_k	Number of sensors of type k
$s[t]$	Type of the affected sensor at time t
$\boldsymbol{s} \triangleq \{s[t]\}_{t=1}^{\infty}$	Trajectory of the anomaly
$\sigma_t^{s[t]}$	Label function that associates sample to group
$\Omega_{\boldsymbol{s}}$	Labels when the trajectory of the anomaly is $\boldsymbol{s}$
$\mathbb{P}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}$ ($\mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}$)	Probability measure (expectation) when the change point is at ν and the samples are permuted according to the label $\Omega_{\boldsymbol{s}}$
$\mathbb{P}_{\Omega}^{\infty}$ ($\mathbb{E}_{\Omega}^{\infty}$)	Probability measure (expectation) when there is no change
$D(P\|\|Q)$	Kullback-Leibler (KL) divergence between two distributions P and Q.

where $\gamma > 0$ is a pre-specified threshold. Here the false alarm constraint is to guarantee that under all possible sample permutations, the average running length to a false alarm is always lower bounded by γ.

A stopping time T is second-order asymptotically optimal if $\text{WARL}(T) \geq \gamma$ and for large γ

$$\text{WADD}(T) = \inf_{\tau : \text{WARL}(\tau) \geq \gamma} \text{WADD}(\tau) + O(1). \tag{5}$$

A stopping time T is first-order asymptotically optimal if $\text{WARL}(T) \geq \gamma$ and for large γ

$$\text{WADD}(T) = \inf_{\tau : \text{WARL}(\tau) \geq \gamma} \text{WADD}(\tau)\,(1 + o(1)). \tag{6}$$

In Table I, we summarize important notations in this paper.

III. Static Anomaly

We first investigate the case with static anomaly, i.e., the sensor affected by the anomaly does not change with time. In this case, for any $t \geq \nu$, $s[t] = k$ for some unknown type k. Then, for all $j \in \{1, 2, \ldots, K, k + K\}$, there are $\binom{n}{n_1, \ldots, n_k - 1, \ldots, n_K, 1}$ possible σ_t^k's to associate each sample with a data-generating distribution, and we denote the collection of all possible labels by $\mathcal{S}_{n,k}$ (see Appendix A for more details). Before the anomaly emerges, i.e., $t < \nu$, $X^n[t]$ follows the distribution

$$\mathbb{P}_{0,\sigma_t^0}(X^n[t]) \triangleq \prod_{i=1}^{n} p_{0,\sigma_t^0(i)}(X_i[t]), \tag{7}$$

for some unknown $\sigma_t^0 \in \mathcal{S}_{n,0}$. At time $t \geq \nu$, $s[t] = k$, $X^n[t]$ follows the distribution

$$\mathbb{P}_{\sigma_t^k}^{k}(X^n[t]) \triangleq \prod_{i : \sigma_t^k(i) \leq K} p_{0,\sigma_t^k(i)}(X_i[t]) \times \prod_{i : \sigma_t^k(i) > K} p_{1,\sigma_t^k(i)-K}(X_i[t]), \tag{8}$$

for some unknown $\sigma_t^k \in \mathcal{S}_{n,k}$. Let $\Omega_k = \{\sigma_1^0, \ldots, \sigma_{\nu-1}^0, \sigma_\nu^k, \ldots, \sigma_\infty^k\}$ be the labels over time, when the anomaly emerges at ν (similarly defined as $\Omega_{\boldsymbol{S}}$). Let $\mathbb{P}_{\Omega_k}^{k,\nu}$ denote the probability measure when the change point is at ν and the samples are generated according to (7), (8) and Ω_k. Further let $\mathbb{E}_{\Omega_k}^{k,\nu}$ denote the corresponding expectation.

The WADD for a stopping time τ can be written as

$$\text{WADD}(\tau) = \sup_{\nu\geq 1}\sup_{k}\sup_{\Omega_k} \text{esssup}\mathbb{E}^{k,\nu}_{\Omega_k}[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]].$$

The WARL is defined in the same way as in (2).

The goal is to design a stopping rule that minimizes the WADD subject to a constraint on the WARL:

$$\inf_{\tau:\text{WARL}(\tau)\geq\gamma} \text{WADD}(\tau). \tag{9}$$

A. Universal Lower Bound on WADD

We first derive a universal lower bound on WADD for any τ satisfying the false alarm constraint: $\inf_\Omega \mathbb{E}^\infty_\Omega[\tau]\geq\gamma$.

Let $I_k = D(\widetilde{\mathbb{P}}_k||\widetilde{\mathbb{P}}_0)$ denote the Kullback-Leibler (KL) divergence between two mixture distributions $\widetilde{\mathbb{P}}^k = \frac{1}{|\mathcal{S}_{n,k}|}\sum_{\sigma\in\mathcal{S}_{n,k}}\mathbb{P}^k_\sigma$ and $\widetilde{\mathbb{P}}_0 = \frac{1}{|\mathcal{S}_{n,0}|}\sum_{\sigma\in\mathcal{S}_{n,0}}\mathbb{P}_{0,\sigma}$. Here, $\widetilde{\mathbb{P}}^k$ is the uniform mixture of all possible labels when the affected sensor is type k. Let $I^* = \min_{1\leq k\leq K} I_k$. We then have the following theorem.

Theorem 1: As $\gamma\to\infty$,

$$\inf_{\tau:\text{WARL}(\tau)\geq\gamma}\text{WADD}(\tau)\geq\frac{\log\gamma}{I^*}+O(1). \tag{10}$$

The proof of Theorem 1 can be found in Appendix B. The main challenge in the proof of Theorem 1 is due to the worst-case over all labels and affected sensors in WADD and WARL. From Theorem 1, it can be seen that the WADD for problem (9) is lower bounded by $\frac{\log\gamma}{I^*}+O(1)$ for any stopping rule that satisfies the constraint on WARL. Theorem 1 motivates us to find the k that minimizes I_k, i.e., achieves I^*, and design an algorithm to achieve this universal lower bound.

B. Generalized Mixture CuSum Algorithm

In this section, we construct an algorithm that achieves the universal lower bound asymptotically.

When there are unknown parameters, MLE is commonly used to estimate the unknown parameters. In the static setting, k does not change with time, however, σ^k_t changes with time, thus a direct MLE for σ^k_t at each time t may not work well.

If k is known, then our problem is invariant under the group of transformations of all possible labels (permutations), that is, our problem is independent of the order of collected samples at each time. Therefore, our problem is related to the invariant theory in [40, Section 6]. This motivates us to take a mixture approach w.r.t. the unknown labels, and then take a MLE approach w.r.t. the unknown affected sensor.

Let $W[t] = \max_{k\in\mathcal{K}}\max_{1\leq j\leq t}\sum_{i=j}^t\log\frac{\widetilde{\mathbb{P}}^k(X^n[i])}{\widetilde{\mathbb{P}}_0(X^n[i])}$. We then define the GM-CuSum stopping time as follows:

$$T_G = \inf\{t: W[t]\geq b\}, \tag{11}$$

where $b>0$ is the threshold. Here $W[t]$ can be updated efficiently. We keep K CuSums in parallel. Note that this can be done recursively. Let $W_k[t] = \max_{1\leq j\leq t}\sum_{i=j}^t\log\frac{\widetilde{\mathbb{P}}^k(X^n[i])}{\widetilde{\mathbb{P}}_0(X^n[i])}$. The test statistic $W[t]$ has the following recursion:

$$W[t+1] = \max_{k\in\mathcal{K}}\left\{(W_k[t])^+ + \log\frac{\widetilde{\mathbb{P}}^k(X^n[t+1])}{\widetilde{\mathbb{P}}_0(X^n[t+1])}\right\}, \tag{12}$$

where $W_k[0]=0, \forall k$. We then take their maximum as $W[t]$.

In the following theorem, we show 1) the WARL lower bound of T_G and 2) the WADD upper bound of T_G.

Theorem 2: 1) Let $b=\log(K\gamma)$ in (11). Then $\text{WARL}(T_G)\geq\gamma$; and 2) As $\gamma\to\infty$, $\text{WADD}(T_G)\leq\frac{\log\gamma}{I^*}+O(1)$.

The proof of Theorem 2 can be found in Appendix C. The proof of the lower bound on WARL is based on Doob's submartingale inequality [41] and the optional sampling theorem [41]. The major challenge lies in that we consider the worst-case label. A key property we develop and use in the proof is that under the pre-change distribution $\mathbb{P}_{0,\sigma^0_t}$, for any $k\in\mathcal{K}$, the expectation of the mixture likelihood ratio $\mathbb{E}_{0,\sigma^0}\left[\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]$ is invariant for different σ^0's.

Theorem 2 suggests that to meet the WARL constraint, b should be chosen such that $b=\log K\gamma$.

Based on Theorems 1 and 2, we then establish the second-order asymptotic optimality of T_G.

Theorem 3: T_G is second-order asymptotically optimal for the problem in (9).

The asymptotic optimality of T_G can be derived similarly under the Pollak's criterion [42]. We omit the details here.

IV. Quickest Dynamic Anomaly Detection

In this section, we consider the general problem with a dynamic anomaly, where the sensor affected by the anomaly changes with time. The GM-CuSum algorithm designed for static anomaly may not work well anymore since the sensor affected by the anomaly changes with time.

A. Universal Lower Bound on WADD

Define the following weighted mixture distribution: $\widetilde{\mathbb{P}}^{\boldsymbol{\beta}}(X^n) = \sum_{k=1}^K\beta_k\widetilde{\mathbb{P}}^k(X^n)$, where $\boldsymbol{\beta}=\{\beta_k\}_{k=1}^K$, $0\leq\beta_k\leq 1$ and $\sum_{k=1}^K\beta_k=1$. Denote by $I_{\boldsymbol{\beta}}$ the KL divergence between $\widetilde{\mathbb{P}}^{\boldsymbol{\beta}}$ and $\widetilde{\mathbb{P}}_0$. Let $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} I_{\boldsymbol{\beta}}$.

For the universal lower bound on WADD, we have the following theorem.

Theorem 4: As $\gamma\to\infty$, we have that

$$\inf_{\tau:\text{WARL}(\tau)\geq\gamma}\text{WADD}(\tau)\geq\frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1)). \tag{13}$$

The proof of Theorem 4 can be found in Appendix D. From Theorem 4, the WADD for the problem in (4) is lower bounded by $\frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1))$ for large γ. This motivates us to apply the optimal weight $\boldsymbol{\beta}^*$ to design an algorithm that can achieves the WADD lower bound asymptotically. Moreover, we have that $I^*\geq I_{\boldsymbol{\beta}^*}$ which implies that a dynamic anomaly is more difficult to detect than a static anomaly.

B. Weighted Mixture CuSum

In the static setting, the unknown affected sensor can be estimated by its MLE. However, in the dynamic setting, the affected sensor changes with time, and the MLE approach may not work well. Theorem 4 motivates us to tackle the unknown anomaly trajectory using a weighted approach where the probability that the k-th group is affected by the anomaly is β_k^*. We then construct our optimal weighted mixture CuSum algorithm as follows. Define the log of weighted mixture likelihood ratio using $\boldsymbol{\beta}^*$:

$$\ell_{\boldsymbol{\beta}^*}(X^n) = \log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}. \tag{14}$$

It can be easily shown that $\ell_{\boldsymbol{\beta}^*}(X^n)$ is invariant to any permutations on X^n, i.e., for any permutation $\pi(X^n) = (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$, $\ell_{\boldsymbol{\beta}^*}(X^n) = \ell_{\boldsymbol{\beta}^*}(\pi(X^n))$. This is due to the fact that $\ell_{\boldsymbol{\beta}^*}(X^n)$ takes the sum over all possible labels thus is invariant to the actual permutation.

We then construct the following optimal weighted mixture CuSum algorithm:

$$T_{\boldsymbol{\beta}^*}(b) = \inf\left\{ t : \max_{1\le j\le t+1} \sum_{i=j}^{t} \ell_{\boldsymbol{\beta}^*}\left(X^n[i]\right) \ge b \right\}. \tag{15}$$

Let $\widehat{W}[t] = \max_{1\le j\le t+1} \sum_{i=j}^{t} \ell_{\boldsymbol{\beta}^*}(X^n[i])$, then $\widehat{W}[t+1] = (\widehat{W}[t])^+ + \ell_{\boldsymbol{\beta}^*}(X^n[t+1])$, and $\widehat{W}[0] = 0$.

Different from the way that we handle the unknown label σ, here, for the unknown type of the affected sensor, we take the mixture according to $\boldsymbol{\beta}^*$ instead of a uniform distribution over $\mathcal{K}$. As will be shown later both theoretically in Theorem 6 and numerically in Section VI, taking a uniform mixture over $\mathcal{K}$ may not lead to the optimal performance.

Let $\widetilde{\mathbb{E}}^k$ and $\widetilde{\mathbb{E}}_0$ denote the expectation under the probability $\widetilde{\mathbb{P}}^k$ and $\widetilde{\mathbb{P}}_0$ respectively. The following property of $\boldsymbol{\beta}^*$ plays an important role in developing the asymptotic optimality of the weighted mixture CuSum algorithm.

Lemma 1: For any $k \in \mathcal{K}$, $\widetilde{\mathbb{E}}^k\left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right] \ge I_{\boldsymbol{\beta}^*}$.

The proof of Lemma 1 can be found in Appendix E.

In the following, we provide a heuristic explanation of how $\widehat{W}[t]$ evolves in the pre- and post-change regimes. We first argue that $\mathbb{E}^k_{\sigma^k}[\ell_{\boldsymbol{\beta}^*}(X^n)]$ is invariant for different σ^k's. Specifically, let $\mathbb{E}^k_{\sigma^k}$ denote the expectation under $\mathbb{P}^k_{\sigma^k}$, where a sensor of type k is affected, and the data received is labeled according to σ^k. For any π, let $\hat{\sigma}^k = \sigma^k \circ \pi$. Then $\mathbb{E}^k_{\sigma^k}[\ell_{\boldsymbol{\beta}^*}(\pi(X^n))] = \mathbb{E}^k_{\sigma^k\circ\pi}[\ell_{\boldsymbol{\beta}^*}(X^n)] = \mathbb{E}^k_{\hat{\sigma}^k}[\ell_{\boldsymbol{\beta}^*}(X^n)]$. For any $\hat{\sigma}^k \in \mathcal{S}_{n,k}$, a π can always be found so that $\sigma^k \circ \pi = \hat{\sigma}^k$. Thus, for any $\sigma^k, \hat{\sigma}^k \in \mathcal{S}_{n,k}$, $\mathbb{E}^k_{\hat{\sigma}^k}[\ell_{\boldsymbol{\beta}^*}(X^n)] = \mathbb{E}^k_{\sigma^k}[\ell_{\boldsymbol{\beta}^*}(X^n)]$. Therefore, $\mathbb{E}^k_{\sigma^k}[\ell_{\boldsymbol{\beta}^*}(X^n)]$ is invariant for different σ^k's. Then, under the pre-change distribution $\mathbb{P}_{0,\sigma_t^0}$, the expectation of the weighted mixture likelihood ratio $\mathbb{E}_{0,\sigma_t^0}[\ell_{\boldsymbol{\beta}^*}(X^n)]$ is invariant for different σ_t^0's. This implies that

$$\mathbb{E}_{0,\sigma_t^0}\left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right] = -D\left(\widetilde{\mathbb{P}}_0 || \widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}\right) \le 0. \tag{16}$$

Therefore, before the change time ν, $\widehat{W}[t]$ has a negative drift. Similarly, from Lemma 1, after the change time ν, under any group assignment $\Omega_{\boldsymbol{s}}$ and trajectory $\boldsymbol{s}$, $\widehat{W}[t]$ has a positive drift whose expectation is no less than $I_{\boldsymbol{\beta}^*}$, and evolves towards ∞.

The following theorem establishes 1) the WARL lower bound of $T_{\boldsymbol{\beta}^*}$, and 2) the WADD upper bound of $T_{\boldsymbol{\beta}^*}$.

Theorem 5: 1) For $T_{\boldsymbol{\beta}^*}$ defined in (15), let $b = \log\gamma$, then $\mathrm{WARL}(T_{\boldsymbol{\beta}^*}) \ge \gamma$. 2) As $\gamma \to \infty$, we have that $\mathrm{WADD}(T_{\boldsymbol{\beta}^*}) \le \frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1))$.

The proof of Theorem 5 can be found in Appendix F. The proof is based on the Weak Law of Large Numbers for the weighted mixture likelihood ratio (similar to [37]). The major challenge lies in that here we are interested in the worst-case label and the worst-case anomaly trajectory. In our problem, the label and the affected sensor change with time. Therefore, it's challenging to explicitly characterize the worst-case label and anomaly trajectory for $T_{\boldsymbol{\beta}^*}$. To show the asymptotically optimal performance of $T_{\boldsymbol{\beta}^*}$, instead of finding the worst-case label and anomaly trajectory, we apply the symmetric property of $T_{\boldsymbol{\beta}^*}$ and Lemma 1 to show that the WADD and WARL of $T_{\boldsymbol{\beta}^*}$ are bounded under all possible labels and trajectories.

We then establish the first-order asymptotic optimality of $T_{\boldsymbol{\beta}^*}$ in the following theorem.

Theorem 6: $T_{\boldsymbol{\beta}^*}$ is first-order asymptotically optimal for problem (4).

Proof: Combining Theorems 4 and 5, we establish the first-order asymptotic optimality of $T_{\boldsymbol{\beta}^*}$. ■

The asymptotic optimality of $T_{\boldsymbol{\beta}^*}$ can be derived similarly under the Pollak's criterion [42]. We omit the details here.

If we apply $T_{\boldsymbol{\beta}^*}$ (designed for the dynamic setting) to the static setting, the WADD of $T_{\boldsymbol{\beta}^*}$ can also be upper bounded by $\frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1))$. However, $T_{\boldsymbol{\beta}^*}$ may not be asymptotically optimal. On the other hand, in the dynamic setting, the sensor affected by the anomaly changes with time, and thus the MLE may not work well. Therefore, the optimal weighted mixture CuSum algorithm works better than the GM-CuSum.

V. Computational Complexity and Efficient Approximation

In the previous sections, we proved that the GM-CuSum algorithm and the optimal weighted mixture CuSum algorithm are asymptotically optimal. However, the test statistic involves computing the mixture likelihood over all possible $\sigma_t^{s[t]}$, which is expensive when n is large.

At each time t, we have $\binom{n}{n_1,\ldots,n_{s[t]}-1,\ldots,n_K,1}$ possible $\sigma_t^{s[t]}$'s. Consider the case with large n, and assume that $\lim_{n\to\infty}\frac{n_k}{n} = \alpha_k$, which is a constant. Let $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^T$. Consider discrete distributions,[3] and denote by $\mathcal{X}$ the support set of the samples. From the exponential bound on the size of the type

[3]Samples in sensor networks are usually quantized before transmitting to fusion center to reduce the power consumption.

class [43], we have that

$$\frac{2^{nH\left(\left[\frac{n_1}{n},\ldots,\frac{n_{s[t]}-1}{n},\ldots,\frac{n_K}{n},\frac{1}{n}\right]\right)}}{(n+1)^{|\mathcal{X}|}} \leq \binom{n}{n_1,\ldots,n_{s[t]}-1,\ldots,n_K,1} \leq 2^{nH\left(\left[\frac{n_1}{n},\ldots,\frac{n_{s[t]}-1}{n},\ldots,\frac{n_K}{n},\frac{1}{n}\right]\right)},$$

where H denotes the Shannon entropy. We then have that

$$\lim_{n\to\infty} H\left(\left[\frac{n_1}{n},\ldots,\frac{n_{s[t]}-1}{n},\ldots,\frac{n_K}{n},\frac{1}{n}\right]\right) = H(\boldsymbol{\alpha}).$$

Therefore, the computational complexity of the GM-CuSum and the optimal weighted mixture CuSum increase exponentially with n, which is expensive for large n.

In this following, we discuss two computationally efficient methods to approximate the test statistics of the GM-CuSum and the optimal weighted mixture CuSum when n is large, and then evaluate their performance in Section VI-D.

The first method is based on the method of types [43]. Let Π_{X^n} denote the empirical distribution of X^n and let $\mathcal{T}(\Pi_{X^n})$ denote the type class of Π_{X^n}. For any k, we have that [4]

$$\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} = \log\frac{\widetilde{\mathbb{P}}^k(\mathcal{T}(\Pi_{X^n}))}{\widetilde{\mathbb{P}}_0(\mathcal{T}(\Pi_{X^n}))}. \tag{17}$$

From the generalized Sanov's theorem [4], [17], the probability of types $\log\widetilde{\mathbb{P}}^k(\mathcal{T}(\Pi_{X^n}))$ can further be approximated by the following optimization problem

$$-\inf_{\substack{\boldsymbol{U}=[U_1,\ldots,U_{K+1}]^T\in(\mathcal{P}_\mathcal{X})^{K+1}\\ \boldsymbol{\phi}^T\boldsymbol{U}=\Pi_{X^n}}} \sum_{j=1,j\neq k}^{K} n_j D(U_j||p_{0,j}) + (n_k-1)D(U_k||p_{0,k}) + D(U_{K+1}||p_{1,k}), \tag{18}$$

where $\boldsymbol{\phi} = \left[\frac{n_1}{n},\ldots,\frac{n_k-1}{n},\ldots,\frac{n_K}{n},\frac{1}{n}\right]^T$ and $\mathcal{P}_\mathcal{X}$ denotes the probability simplex on $\mathcal{X}$. Problem (18) is a convex optimization problem whose computational complexity is independent of n. Therefore, the computation of test statistics of the GM-CuSum algorithm and the optimal weighted mixture CuSum algorithm can be converted to solving convex optimization problems and the overall complexity at each time step is only linear in the number of sensors.

The second method is to estimate the unknown labels using the MLE and to use the generalized likelihood ratio test (GLRT) to approximate the mixture likelihood ratio [6]. Computing the GLRT is a special case of the assignment problem and efficient algorithms have been developed. In [6], two efficient greedy algorithms were proposed to solve the assignment problem approximately with complexity $O(n^2)$. Therefore, at each time t, the test statistics of the GM-CuSum and the optimal weighted mixture CuSum algorithm can be approximated with computational complexity $O(Kn^2)$.

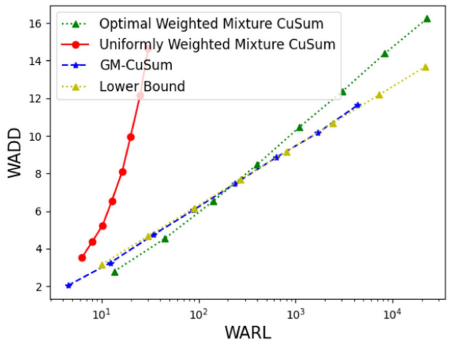

Fig. 1. Static setting: $n = 4, K = 4$.

VI. Numerical Results

A. Static Anomaly Detection

We first consider the static setting, and compare our GM-CuSum with a uniformly weighted mixture CuSum

$$T_B = \inf\left\{t : \max_{1\leq j\leq t}\sum_{i=j}^{t}\log\frac{\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\widetilde{\mathbb{P}}^k(X^n[i])}{\widetilde{\mathbb{P}}_0(X^n[i])} \geq b\right\}$$

and the optimal weighted mixture CuSum. We plot the WADD as a function of the WARL under the worst-case static trajectory. We also plot the asymptotic lower bound $\frac{\log\gamma}{I^*}$.

We first compare the three algorithms under the Gaussian distributions. There are four types of sensors and each type contains one sensor. For the type I sensor, the pre- and post-change distributions are $\mathcal{N}(-1,1)$ and $\mathcal{N}(2,1)$, for the type II sensor, the pre- and post-change distributions are $\mathcal{N}(1,1)$ and $\mathcal{N}(3,1)$, for the type III sensor, the pre- and post-change distributions are $\mathcal{N}(-1,1)$ and $\mathcal{N}(3,1)$, for the type IV sensor, the pre- and post-change distributions are $\mathcal{N}(1,1)$ and $\mathcal{N}(-1,1)$ respectively. The optimal weight for our weighted mixture CuSum algorithm is solved by Monte-Carlo. It can be seen from Fig. 1 that with the same false alarm rate, our GM-CuSum has the lowest WADD, which implies that it detects the anomaly with the smallest detection delay. Moreover, the slope of the GM-CuSum matches the lower bound, which validates that the GM-CuSum is asymptotically optimal. The optimal weighted mixture CuSum algorithm also has a good performance in the static setting.

We then compare the three algorithms under the binomial distributions. We consider two cases with different K. For the case where there are two types of sensors, for type I sensors, the pre- and post-change distributions are $\mathcal{B}(10,0.2)$ and $\mathcal{B}(10,0.5)$, for type II sensors, the pre- and post-change distributions are $\mathcal{B}(10,0.8)$ and $\mathcal{B}(10,0.6)$, respectively. Here $\mathcal{B}$ denotes binomial distribution, the first parameter denotes the number of trials and the second parameter denotes the success probability of each trial. We plot the results for the case where each type has four sensors in Fig. 2. For the case where there are four types of sensors, for type I sensors, the pre- and post-change distributions are $\mathcal{B}(10,0.2)$ and $\mathcal{B}(10,0.8)$, for type II sensors, the pre- and post-change distributions are $\mathcal{B}(10,0.3)$ and $\mathcal{B}(10,0.6)$, for type III sensors, the pre- and post-change distributions are $\mathcal{B}(10,0.5)$ and $\mathcal{B}(10,0.9)$, for type IV sensors, the pre- and post-change

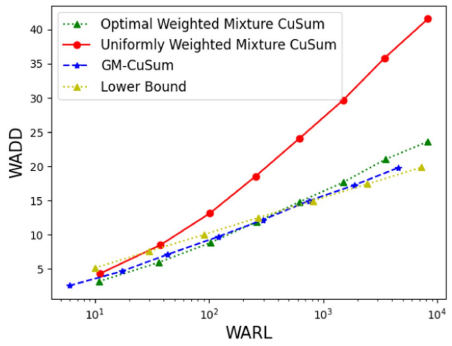

Fig. 2. Static setting: $n = 8, K = 2$.

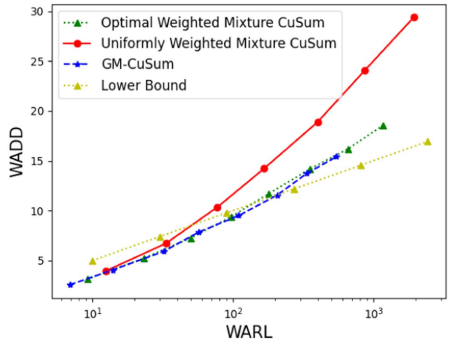

Fig. 3. Static setting: $n = 8, K = 4$.

distributions are $\mathcal{B}(10, 0.4)$ and $\mathcal{B}(10, 0.7)$ respectively. We plot the results for the case where each type has two sensors in Fig. 3. We use Monte-Carlo to obtain the optimal weight for our optimal weighted mixture CuSum algorithm. It can be seen from Figs. 2 and 3 that with the same false alarm rate, our GM-CuSum has the lowest WADD, which implies that it detects the anomaly with the smallest detection delay. Moreover, the relationship between the WADD and log of the WARL is linear, and the slope of the GM-CuSum matches with the one of the lower bound, validating its asymptotic optimality.

B. Dynamic Anomaly Detection

For detecting the dynamic anomaly, we use the same parameters of distributions as in the static setting.

We compare our optimal weighted mixture CuSum algorithm with a uniformly weighted mixture CuSum, i.e., replace $\boldsymbol{\beta}^*$ in (15) with $\boldsymbol{\beta} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and the GM-CuSum under the Gaussian distribution. We plot the average detection delay (ADD) and the average run length (ARL) for some randomly generated trajectories since the worst-case trajectory is difficult to simulate. It can be seen from Fig. 4 that the weighted mixture CuSum algorithm outperforms the uniformly weighted mixture CuSum algorithm and the GM-CuSum. Therefore, in the dynamic setting, the optimal weighted mixture CuSum algorithm detects the presence of the anomaly with the lowest detection delay. Moreover, the slope of our optimal weighted mixture CuSum algorithm matches with the one of the lower bound, which validates its asymptotic optimality.

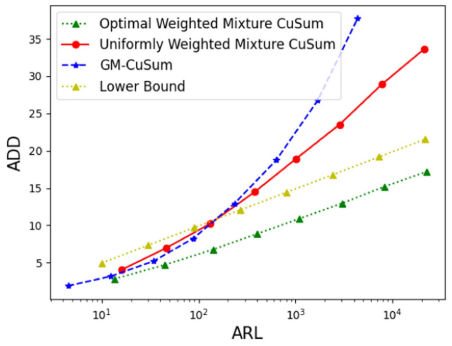

Fig. 4. Dynamic setting: $n = 4, K = 4$.

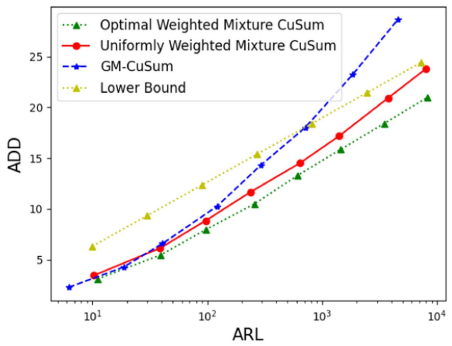

Fig. 5. Dynamic setting: $n = 8, K = 2$.

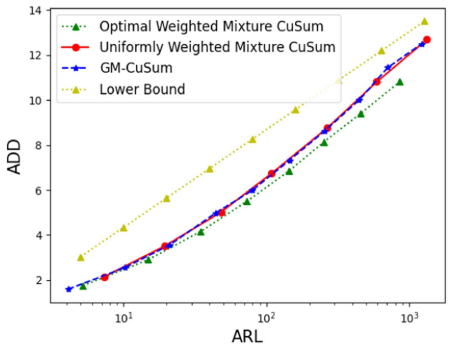

Fig. 6. Dynamic setting: $n = 8, K = 4$.

We then compare the three algorithms using the binomial distributions as in the static setting. It can be seen from Figs. 5 and 6 that our optimal weighted mixture CuSum algorithm outperforms the uniformly weighted mixture CuSum algorithm and the GM-CuSum since with the same false alarm rate, the optimal weighted mixture CuSum algorithm has the smallest detection delay. The relationship between the ADD and log of the ARL is linear. Moreover, the slope of our optimal weighted mixture CuSum algorithm matches the theoretical lower bound, which demonstrates its asymptotic optimality. It can also be observed that the GM-CuSum algorithm does not perform well under the dynamic setting.

C. Moving Target Detection With Unlabeled Samples

In this section, we consider a practical application of target detection [7], [44]. For simplicity, consider a 3×3 grid. One

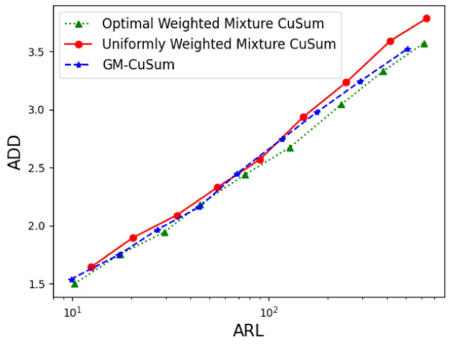

Fig. 7. Moving target detection.

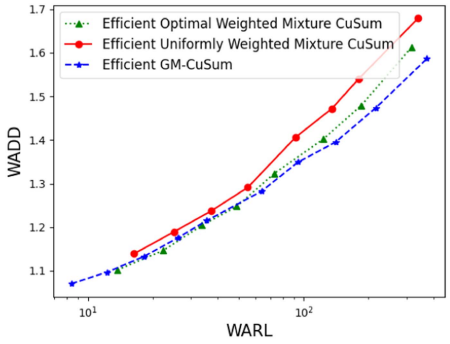

Fig. 8. Static setting: $n = 60, K = 3$.

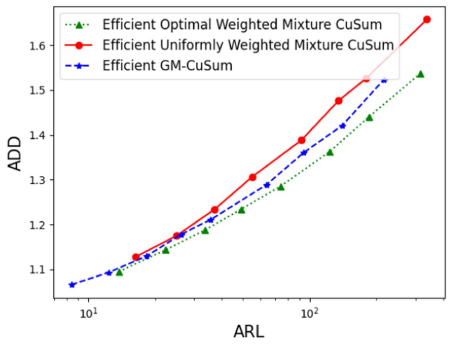

Fig. 9. Dynamic setting: $n = 60, K = 3$.

sensor is deployed at the center of each cell. The fusion center can only collect unlabeled data. Before the target presents, at each time, sensors will send exponentially distributed samples with mean 1 to the fusion center. When the target appears, the distribution of each sample changes to another exponential distribution with an elevated mean, depending on the distance between the cell and the target. The target moves across different cells at different time. In this case, the mixture likelihood $\widetilde{\mathbb{P}}^k(X^n)$ can be computed when the target lies in a specific cell and the GM-CuSum algorithm and the optimal weighted mixture CuSum algorithm can further be designed. We compare the GM-CuSum algorithm, the uniformly weighted mixture CuSum algorithm and the optimal weighted mixture CuSum algorithm. From Fig. 7, with the same false alarm rate, our optimal weighted mixture CuSum has the smallest detection delay which implies that it detects the presence of the target quickly, e.g., with an ARL of 10^3, it only takes about 3.5 samples to detect the target.

D. Computationally Efficient Approximation

In this section, we implement the computationally efficient approximation in Section V by replacing the test statistics in uniformly weighted mixture CuSum algorithm, GM-CuSum algorithm and optimal weighted mixture CuSum with the value of the optimization problem in (18) and demonstrate their performance when n is large.

Let $n = 60, K = 3$ and $n_1 = n_2 = n_3 = 20$. After the anomaly emerges, the distribution of all the sensors of an unknown type changes. Our GM-CuSum algorithm and optimal weighted mixture CuSum algorithm can be designed with a slight modification of the mixture likelihood $\widetilde{\mathbb{P}}^k(X^n)$. For type I sensors, the pre- and post-change distributions are $\mathcal{B}(10, 0.45)$ and $\mathcal{B}(10, 0.6)$, for type II sensors, the pre- and post-change distributions are $\mathcal{B}(10, 0.4)$ and $\mathcal{B}(10, 0.6)$, and for type III sensors, the pre- and post-change distributions are $\mathcal{B}(10, 0.7)$ and $\mathcal{B}(10, 0.55)$ respectively. The optimal weight for our optimal weighted mixture CuSum algorithm is obtained by Monte-Carlo. We consider the static setting, and plot the WADD as a function of the WARL. It can be seen from Fig. 8 that the GM-CuSum has the smallest WADD given the same WARL. Also the WADD is small (it only takes around 1.6 samples to detect the target for a given a WARL at 4×10^2). Moreover, our GM-CuSum has the best performance for detecting the static anomaly when n is large.

We then consider the dynamic setting, we plot the ADD as a function of ARL for some random trajectories. It can be seen from Fig. 9 that with the same ARL, our optimal weighted mixture CuSum algorithm has the smallest detection delay and thus has the best performance for detecting the dynamic anomaly when n is large.

VII. Conclusion

In this paper, we studied the problem of quickest detection of an anomaly in networks with unlabeled samples. We first investigated the case with a static anomaly. A GM-CuSum algorithm was proposed and shown to be second-order asymptotically optimal. We then extended our study to the case with a dynamic anomaly, that is, the affected sensor changes with time. We proposed an optimal weighted mixture CuSum algorithm, and proved that it is first-order asymptotically optimal. Our approaches provide useful insights for general (sequential) statistical inference problems with unlabeled samples.

Appendix A

Before the anomaly emerges, i.e., $t < \nu$, there are n_k sensors in group k, $\forall 1 \le k \le K$, and 0 sensors in group k, $\forall K < k \le 2K$. Then, there are in total $\binom{n}{n_1,\ldots,n_K}$ possible $\sigma_t^{s[t]}$: $\{1,\ldots,n\} \to \{1,\ldots,K\}$ satisfying $|\{i : \sigma_t^{s[t]}(i) = k\}| = n_k$, for any $k = 1,\ldots,K$. We denote the collection of all such labels by $\mathcal{S}_{n,0}$. After the anomaly emerges, i.e., $t \ge \nu$, one sensor of type $s[t] \ne 0$ is affected by anomaly. Therefore, the

number of sensors in group $s[t]$ and $s[t]+K$ are $n_{s[t]}-1$ and 1 respectively. Then, there are $\binom{n}{n_1,\ldots,n_{s[t]}-1,\ldots,n_K,1}$ possible $\sigma_t^{s[t]}$: $\{1,\ldots,n\}\to\{1,\ldots,K,s[t]+K\}$ satisfying

$$\left|\left\{i:\sigma_t^{s[t]}(i)=k\right\}\right|=\begin{cases}n_k, & \text{if } 1\le k\le K \text{and } k\ne s[t],\\ n_k-1, & \text{if } k=s[t],\\ 1, & \text{if } k=s[t]+K,\\ 0, & \text{otherwise.}\end{cases}$$

We then denote the collection of all such labels by $\mathcal{S}_{n,s[t]}$.

Before the anomaly emerges, i.e., $t<\nu$, the samples $X^n[t]$ follows the distribution

$$\mathbb{P}_{0,\sigma_t^0}\left(X^n[t]\right)=\prod_{i=1}^{n}p_{0,\sigma_t^0(i)}(X_i[t]), \tag{19}$$

for some unknown $\sigma_t^0\in\mathcal{S}_{n,0}$. At time $t\ge\nu$, $X^n[t]$ follows the distribution

$$\mathbb{P}^{s[t]}_{\sigma_t^{s[t]}}\left(X^n[t]\right)\triangleq\prod_{i:\sigma_t^{s[t]}(i)\le K}p_{0,\sigma_t^{s[t]}(i)}(X_i[t])\times\prod_{i:\sigma_t^{s[t]}(i)>K}p_{1,\sigma_t^{s[t]}(i)-K}(X_i[t]), \tag{20}$$

for some unknown $\sigma_t^{s[t]}\in\mathcal{S}_{n,s[t]}$.

Appendix B
Proof of Theorem 1

Consider a simple QCD problem with a pre-change distribution $\widetilde{\mathbb{P}}_0$ and a post-change distribution $\widetilde{\mathbb{P}}^k$, respectively. Let

$$\widetilde{\mathrm{WADD}}_k(\tau)=\sup_{\nu\ge1}\operatorname{esssup}\widetilde{\mathbb{E}}^{k,\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right],$$
$$\widetilde{\mathrm{ARL}}(\tau)=\widetilde{\mathbb{E}}^{\infty}[\tau], \tag{21}$$

where $\widetilde{\mathbb{E}}^{k,\nu}$ denotes the expectation when the change is at ν, the pre- and post-change distributions are $\widetilde{\mathbb{P}}_0$ and $\widetilde{\mathbb{P}}^k$, and $\widetilde{\mathbf{X}}^n[t]$ for $1\le t\le\nu-1$ are i.i.d. from $\widetilde{\mathbb{P}}_0$, $\widetilde{\mathbb{E}}^{\infty}$ denotes the expectation when samples are generated according to $\widetilde{\mathbb{P}}_0$.

For any $1\le k\le K$, consider another QCD problem with a pre-change distribution $\mathbb{P}_{0,\sigma_t^0}$ and a post-change distribution $\mathbb{P}^k_{\sigma_t^k}$, respectively. For this pair of pre- and post-change distributions, define

$$\mathrm{WADD}_k(\tau)=\sup_{\nu\ge1}\sup_{\Omega_k}\operatorname{esssup}\mathbb{E}^{k,\nu}_{\Omega_k}\left[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]\right],$$
$$\mathrm{WARL}(\tau)=\inf_{\Omega}\mathbb{E}^{\infty}_{\Omega}[\tau]. \tag{22}$$

For any $1\le k\le K$ and any τ satisfying $\mathrm{WARL}(\tau)\ge\gamma$, it can be shown that

$$\begin{aligned}\mathrm{WADD}(\tau)&=\sup_{k\in\mathcal{K}}\mathrm{WADD}_k(\tau)\\&\ge\sup_{\nu\ge1}\sup_{\Omega_k}\operatorname{esssup}\mathbb{E}^{k,\nu}_{\Omega_k}\left[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]\right]\\&\ge\sup_{\nu\ge1}\operatorname{esssup}\widetilde{\mathbb{E}}^{k,\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right]\\&=\widetilde{\mathrm{WADD}}_k(\tau).\end{aligned} \tag{23}$$

The second inequality is due to the fact that for any τ, $\mathrm{WADD}_k(\tau)\ge\widetilde{\mathrm{WADD}}_k(\tau)$ [17, eq. (18)]. Similarly, we have that for any τ, $\mathrm{WARL}(\tau)\le\widetilde{\mathrm{ARL}}(\tau)$ [17, eq. (18)]. It then follows that for any $k\in\mathcal{K}$,

$$\begin{aligned}\inf_{\tau:\mathrm{WARL}(\tau)\ge\gamma}\mathrm{WADD}(\tau)&\ge\inf_{\tau:\widetilde{\mathrm{ARL}}(\tau)\ge\gamma}\widetilde{\mathrm{WADD}}_k(\tau)\\&\ge\frac{\log\gamma}{I_k}+O(1),\ \text{as }\gamma\to\infty.\end{aligned} \tag{24}$$

The last inequality is due to the universal lower bound on WADD for a simple QCD problem [37]. We then have that

$$\inf_{\tau:\mathrm{WARL}(\tau)\ge\gamma}\mathrm{WADD}(\tau)\ge\frac{\log\gamma}{I^*}+O(1),\ \text{as }\gamma\to\infty. \tag{25}$$

Appendix C
Proof of Theorem 2

For any $m\ge0$, let $r_0=0$ and define the stopping time

$$r_{m+1}=\inf\left\{t>r_m:\sup_k\sum_{i=r_m+1}^{t}\log\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\le0\right\}. \tag{26}$$

For any permutation $\pi(X^n)=(X_{\pi(1)},X_{\pi(2)},\ldots,X_{\pi(n)})$, we have that $\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}=\log\frac{\widetilde{\mathbb{P}}^k(\pi(X^n))}{\widetilde{\mathbb{P}}_0(\pi(X^n))}$. For any π, let $\hat\sigma^0=\sigma^0\circ\pi$, where "$\circ$" denotes the composition of two functions. Then $\mathbb{E}_{0,\sigma^0}\left[\log\frac{\widetilde{\mathbb{P}}^k(\pi(X^n))}{\widetilde{\mathbb{P}}_0(\pi(X^n))}\right]=\mathbb{E}_{0,\sigma^0\circ\pi}\left[\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]=\mathbb{E}_{0,\hat\sigma^0}\left[\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]$. For any $\hat\sigma^0\in\mathcal{S}_{n,0}$, a π can always be found so that $\sigma^0\circ\pi=\hat\sigma^0$. Thus, for any $\sigma^0,\hat\sigma^0\in\mathcal{S}_{n,0}$, $\mathbb{E}_{0,\hat\sigma^0}\left[\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]=\mathbb{E}_{0,\sigma^0}\left[\log\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]$.

We then have that for any $\sigma^0\in\mathcal{S}_{n,0}$,

$$\begin{aligned}\mathbb{E}_{0,\sigma^0}\left[\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]&=\frac{1}{|\,\mathcal{S}_{n,0}\,|}\sum_{\sigma^0\in\mathcal{S}_{n,0}}\mathbb{E}_{0,\sigma^0}\left[\frac{\widetilde{\mathbb{P}}^k(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}\right]\\&=\int\frac{\widetilde{\mathbb{P}}^k(x^n)}{\widetilde{\mathbb{P}}_0(x^n)}\cdot\widetilde{\mathbb{P}}_0(x^n)\mathrm{d}x^n=1.\end{aligned} \tag{27}$$

Therefore, for any Ω and $t>r_m$,

$$\begin{aligned}&\mathbb{E}^{\infty}_{\Omega}\left[\prod_{i=r_m+1}^{t+1}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\middle|\mathcal{F}_t\right]\\&=\mathbb{E}^{\infty}_{\Omega}\left[\prod_{i=r_m+1}^{t}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\middle|\mathcal{F}_t\right]\cdot\mathbb{E}_{0,\sigma^0}\left[\frac{\widetilde{\mathbb{P}}^k(X_{t+1}^n)}{\widetilde{\mathbb{P}}_0(X_{t+1}^n)}\middle|\mathcal{F}_t\right]\\&=\prod_{i=r_m+1}^{t}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\cdot\mathbb{E}_{0,\sigma^0}\left[\frac{\widetilde{\mathbb{P}}^k(X_{t+1}^n)}{\widetilde{\mathbb{P}}_0(X_{t+1}^n)}\right]\\&=\prod_{i=r_m+1}^{t}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}.\end{aligned} \tag{28}$$

Therefore, $\left\{\prod_{i=r_m+1}^{t}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}, \mathcal{F}_t, t>r_m\right\}$ is a martingale under $\mathbb{P}_\Omega^\infty$ for any Ω with mean 1.

We then have that for any Ω,

$$\begin{aligned}&\mathbb{P}_\Omega^\infty\left\{\sup_k\sum_{i=r_m+1}^{t}\log\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\geq b \text{ for some } t>r_m\middle|\mathcal{F}_{r_m}\right\}\\&\leq\sum_{k=1}^{K}\mathbb{P}_\Omega^\infty\left\{\prod_{i=r_m+1}^{t}\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\geq e^b \text{ for some } t>r_m\middle|\mathcal{F}_{r_m}\right\}\\&\leq K\frac{\mathbb{E}_{0,\sigma^0}\left[\frac{\widetilde{\mathbb{P}}^k(X_{r_m+1}^n)}{\widetilde{\mathbb{P}}_0(X_{r_m+1}^n)}\right]}{e^b}=Ke^{-b},\end{aligned}\tag{29}$$

where the last inequality is due to Doob's submartingale inequality [41] and the optional sampling theorem [41].

Let $M=\inf\left\{m\geq 0: r_m<\infty, \sup_k\sum_{i=r_m+1}^{t}\log\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}\geq b \text{ for some } t>r_m\right\}$. We have that for any Ω,

$$\begin{aligned}&\mathbb{P}_\Omega^\infty\left(M\geq m+1|\mathcal{F}_{r_m}\right)\\&\geq\mathbb{P}_\Omega^\infty\left\{\sup_k\sum_{i=r_m+1}^{t}\log\frac{\widetilde{\mathbb{P}}^k(X_i^n)}{\widetilde{\mathbb{P}}_0(X_i^n)}<b \text{ for all } t>r_m\middle|\mathcal{F}_{r_m}\right\}\\&\geq 1-Ke^{-b}.\end{aligned}\tag{30}$$

We then have that for any Ω,

$$\begin{aligned}\mathbb{P}_\Omega^\infty(M>m)&=\mathbb{E}_\Omega^\infty\left[\mathbb{P}_\Omega^\infty\left(M\geq m+1|\mathcal{F}_{r_m}\right)\cdot\mathbb{1}_{\{M\geq m\}}\right]\\&\geq\left(1-Ke^{-b}\right)\mathbb{P}_\Omega^\infty(M>m-1)\\&\geq\left(1-Ke^{-b}\right)^m.\end{aligned}\tag{31}$$

It then follows that

$$\begin{aligned}&\mathrm{WARL}(T_G)=\inf_\Omega\mathbb{E}_\Omega^\infty[T_G]\geq\inf_\Omega\mathbb{E}_\Omega^\infty[M]\\&\geq\inf_\Omega\sum_{m=0}^{\infty}\mathbb{P}_\Omega^\infty(M>m)\geq\sum_{m=0}^{\infty}\left(1-Ke^{-b}\right)^m=\frac{e^b}{K}.\end{aligned}\tag{32}$$

Let $b=\log K\gamma$, we have that $\mathrm{WARL}(T_G)\geq\gamma$. Let T_k be the mixture CuSum algorithm for problem in (22):

$$T_k=\inf\left\{t:\max_{1\leq j\leq t}\sum_{i=j}^{t}\log\frac{\widetilde{\mathbb{P}}^k(X^n[i])}{\widetilde{\mathbb{P}}_0(X^n[i])}\geq b\right\}.\tag{33}$$

It then follows that for any $1\leq k\leq K$,

$$\begin{aligned}\mathrm{WADD}_k(T_G)&=\sup_{\nu\geq 1}\sup_{\Omega_k}\mathrm{esssup}\mathbb{E}_{\Omega_k}^{k,\nu}[(T_G-\nu)^+|\mathbf{X}^n[1,\nu-1]]\\&\leq\sup_{\nu\geq 1}\sup_{\Omega_k}\mathrm{esssup}\mathbb{E}_{\Omega_k}^{k,\nu}[(T_k-\nu)^+|\mathbf{X}^n[1,\nu-1]]\\&\leq\frac{\log b}{I_k}+O(1),\end{aligned}\tag{34}$$

where the last equality is because of the exact optimality of the mixture CuSum algorithm (see Theorem 1 in [17]).

To satisfy the WARL constraint, set $b=\log K\gamma$, we have

$$\begin{aligned}&\mathrm{WADD}(T_G)=\sup_{k\in\mathcal{K}}\mathrm{WADD}_k(T_G)\leq\sup_{k\in\mathcal{K}}\mathrm{WADD}_k(T_k)\\&=\sup_{k\in\mathcal{K}}\frac{\log K\gamma}{I_k}+O(1)=\frac{\log\gamma}{I^*}+O(1),\ \text{as}\ \gamma\to\infty.\end{aligned}\tag{35}$$

Appendix D
Proof of Theorem 4

For any trajectory $\boldsymbol{s}$ and stopping time τ, define the WADD and WARL

$$\begin{aligned}\mathrm{WADD}_{\boldsymbol{s}}(\tau)&=\sup_{\nu\geq 1}\sup_{\Omega_{\boldsymbol{s}}}\mathrm{esssup}\mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},\nu}[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]],\\\mathrm{ARL}(\tau)&=\inf_\Omega\mathbb{E}_\Omega^\infty[\tau].\end{aligned}\tag{36}$$

Consider QCD problem with a pre-change distribution $\widetilde{\mathbb{P}}_0=\frac{1}{|\mathcal{S}_{n,0}|}\sum_{\sigma^0\in\mathcal{S}_{n,0}}\mathbb{P}_{0,\sigma^0}$ and a post-change distribution $\widetilde{\mathbb{P}}^{s[t]}=\frac{1}{|\mathcal{S}_{n,s[t]}|}\sum_{\sigma^{s[t]}\in\mathcal{S}_{n,s[t]}}\mathbb{P}_{\sigma^{s[t]}}^{s[t]}$, respectively. For this pair of pre- and post-change distributions and any trajectory $\boldsymbol{s}$, define

$$\begin{aligned}\widetilde{\mathrm{WADD}}_{\boldsymbol{s}}(\tau)&=\sup_{\nu\geq 1}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right],\\\widetilde{\mathrm{ARL}}(\tau)&=\widetilde{\mathbb{E}}^\infty[\tau].\end{aligned}\tag{37}$$

where $\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}$ denotes the expectation when change point is ν, before the change point, the data follows distribution $\widetilde{\mathbb{P}}_0$ and after the change point, at time t, the data follows the distribution $\widetilde{\mathbb{P}}^{s[t]}$, and $\widetilde{\mathbf{X}}^n[1,\nu-1]$ are i.i.d. from $\widetilde{\mathbb{P}}_0$; and $\widetilde{\mathbb{E}}^\infty$ denote the expectation when the data follows distribution $\widetilde{\mathbb{P}}_0$.

Consider another QCD problem with pre-change distribution $\widetilde{\mathbb{P}}_0$ and post-change distribution $\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}$. Under this pair of pre- and post-change distributions, define

$$\begin{aligned}\widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}^*}(\tau)&=\sup_{\nu\geq 1}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{\beta}^*,\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right],\\\widetilde{\mathrm{ARL}}(\tau)&=\widetilde{\mathbb{E}}^\infty[\tau].\end{aligned}\tag{38}$$

In QCD problems, ARL only depends on the pre-change distribution. Therefore, for any stopping time τ, problems in (2) and (36) have the same ARL, problems in (37) and (38) have the same ARL. Let $\mathcal{C}_\gamma$ denotes the collection of all stopping times τ that satisfy $\mathrm{ARL}(\tau)\geq\gamma$ and $\widetilde{\mathcal{C}}_\gamma$ denotes the collection of all stopping times τ that satisfy $\widetilde{\mathrm{ARL}}(\tau)\geq\gamma$. Our goal is to prove that

$$\inf_{\tau\in\mathcal{C}_\gamma}\mathrm{WADD}(\tau)\geq\inf_{\tau\in\widetilde{\mathcal{C}}_\gamma}\widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}^*}(\tau)\sim\frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1)).\tag{39}$$

Construct a new sequence of random variables $\left\{\widehat{X}^n[t]\right\}_{t=1}^{\infty}$. Before the change point, $\widehat{X}^n[t]$ are i.i.d. according to the mixture distribution $\widetilde{\mathbb{P}}_0=\frac{1}{|\mathcal{S}_{n,0}|}\sum_{\sigma^0\in\mathcal{S}_{n,0}}\mathbb{P}_{0,\sigma^0}$. After the change point, i.e., $t\geq\nu$, $\widehat{X}^n[t]$ follows the distribution $\mathbb{P}_{\sigma_t^{s[t]}}^{s[t]}$ for some $\sigma_t^{s[t]}\in\mathcal{S}_{n,s[t]}$. Specifically,

$$\widehat{X}^n[t]\sim\begin{cases}\widetilde{\mathbb{P}}_0, & \text{if } t<\nu,\\ \mathbb{P}_{\sigma_t^{s[t]}}^{s[t]}, & \text{if } t\geq\nu.\end{cases}\tag{40}$$

For any stopping time τ and any $\boldsymbol{s}$, let

$$\widehat{\text{WADD}}_{\boldsymbol{s}}(\tau) = \sup_{\nu\geq 1} \sup_{\sigma_\nu^{s[\nu]},\ldots,\sigma_\infty^{s[\infty]}} \text{esssup} \widehat{\mathbb{E}}^{\boldsymbol{s},\nu}_{\sigma_\nu^{s[\nu]},\ldots,\sigma_\infty^{s[\infty]}} \left[(\tau-\nu)^+ | \widehat{\mathbf{X}}^n[1,\nu-1]\right], \quad (41)$$

where $\widehat{\mathbb{E}}^{\boldsymbol{s},\nu}_{\sigma_\nu^{s[\nu]},\ldots,\sigma_\infty^{s[\infty]}}$ denotes the expectation when the data is distributed according to (40).

Let $\widetilde{\text{WADD}}(\tau) = \sup_{\boldsymbol{s}} \widetilde{\text{WADD}}_{\boldsymbol{s}}(\tau)$. To prove (39), we first show that for any $\boldsymbol{s}$, $\text{WADD}_{\boldsymbol{s}}(\tau) = \widehat{\text{WADD}}_{\boldsymbol{s}}(\tau)$, and then show that $\widehat{\text{WADD}}_{\boldsymbol{s}}(\tau) \geq \widetilde{\text{WADD}}_{\boldsymbol{s}}(\tau)$. We complete our proof by showing that for any τ and $\boldsymbol{\beta}$, $\widetilde{\text{WADD}}(\tau) \geq \widetilde{\text{WADD}}_{\boldsymbol{\beta}}(\tau)$.

Step 1: Denote by $\mathcal{M}$ the collection of all $\{\sigma_1^0,\ldots,\sigma_{\nu-1}^0\}$, and μ is an element in $\mathcal{M}$. When the trajectory is $\boldsymbol{s}$, denote by $\mathcal{N}_{\boldsymbol{s}}$ the collection of all $\{\sigma_\nu^{s[\nu]},\ldots,\sigma_\infty^{s[\infty]}\}$, and ω is an element in $\mathcal{N}_{\boldsymbol{s}}$. Then, the $\text{WADD}_{\boldsymbol{s}}$ can be written as

$$\begin{aligned}\text{WADD}_{\boldsymbol{s}}(\tau) &= \sup_{\nu\geq 1} \sup_{\Omega_{\boldsymbol{s}}} \text{esssup}\mathbb{E}^{\boldsymbol{s},\nu}_{\Omega_{\boldsymbol{s}}}[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]] \\ &= \sup_{\nu\geq 1} \sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}} \sup_{\mu\in\mathcal{M}} \text{esssup}\mathbb{E}^{\boldsymbol{s},\nu}_{\omega}[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]],\end{aligned}$$

where $\mathbb{E}^{\boldsymbol{s},\nu}_{\omega}$ denotes the expectation when change point is ν, the trajectory is $\boldsymbol{s}$, and after the change point, the data follows distribution $\prod_{t=\nu}^{\infty}\mathbb{P}^{s[t]}_{\sigma_t^{s[t]}}$. We note that $\widehat{X}^n[t]$ and $X^n[t]$, for $t\geq\nu$, have the same distribution $\mathbb{P}^{s[t]}_{\sigma_t^{s[t]}}$. Therefore, the difference between $\text{WADD}_{\boldsymbol{s}}$ and $\widehat{\text{WADD}}_{\boldsymbol{s}}$ lies in that they take esssup with respect to different distributions, i.e., the distributions of $\mathbf{X}^n[1,\nu-1]$ and $\widehat{\mathbf{X}}^n[1,\nu-1]$ are different. Let $f_\omega(\mathbf{X}^n[1,\nu-1])$ denote $\mathbb{E}^{\boldsymbol{s},\nu}_{\omega}[(\tau-\nu)^+|\mathbf{X}^n[1,\nu-1]]$. Then, $\text{WADD}_{\boldsymbol{s}}$ and $\widehat{\text{WADD}}_{\boldsymbol{s}}$ can be written as

$$\begin{aligned}\text{WADD}_{\boldsymbol{s}}(\tau) &= \sup_{\nu\geq 1}\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}}\sup_{\mu\in\mathcal{M}} \text{esssup} f_\omega\left(\mathbf{X}^n[1,\nu-1]\right), \\ \widehat{\text{WADD}}_{\boldsymbol{s}}(\tau) &= \sup_{\nu\geq 1}\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}} \text{esssup} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right). \quad (42)\end{aligned}$$

It then suffices to show that for any $\omega\in\mathcal{N}_s$,

$$\sup_{\mu\in\mathcal{M}} \text{esssup} f_\omega\left(\mathbf{X}^n[1,\nu-1]\right) = \text{ess sup}\, f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right).$$

For any $\omega\in\mathcal{N}_{\boldsymbol{s}}$ and $\mu\in\mathcal{M}$, let

$$\begin{aligned}b_{\omega,\mu} &= \text{esssup} f_\omega\left(\mathbf{X}^n[1,\nu-1]\right) \\ &= \inf\left\{b : \mathbb{P}_\mu\left(f_\omega(\mathbf{X}^n[1,\nu-1]) > b\right) = 0\right\}, \quad (43)\end{aligned}$$

where $\mathbb{P}_\mu$ denotes the probability measure when the data is generated from $\mathbb{P}_{0,\sigma_1^0},\ldots,\mathbb{P}_{0,\sigma_{\nu-1}^0}$ before change point ν.

Let $b^*_\omega = \text{esssup} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right)$. It can be shown that

$$\begin{aligned}b^*_\omega &= \inf\left\{b : \int_{\mathbf{x}^n[1,\nu-1]} \mathbb{1}_{\{f_\omega(\mathbf{x}^n[1,\nu-1])>b\}} \times \mathrm{d}\prod_{t=1}^{\nu-1}\widetilde{\mathbb{P}}_0(x^n(t)) = 0\right\} \\ &= \inf\left\{b : \int_{\mathbf{x}^n[1,\nu-1]} \mathbb{1}_{\{f_\omega(\mathbf{x}^n[1,\nu-1])>b\}} \times \mathrm{d}\frac{1}{|\mathcal{M}|}\sum_{\mu\in\mathcal{M}}\mathbb{P}_\mu(\mathbf{x}^n[1,\nu-1]) = 0\right\} \\ &= \inf\left\{b : \frac{1}{|\mathcal{M}|}\sum_{\mu\in\mathcal{M}}\mathbb{P}_\mu\left(f_\omega\left(\mathbf{X}^n[1,\nu-1]\right) > b\right) = 0\right\}.\end{aligned}$$

It then follows that for any $\mu\in\mathcal{M}$, and $\omega\in\mathcal{N}_s$, $\mathbb{P}_\mu(f_\omega(\mathbf{X}^n[1,\nu-1]) > b^*_\omega) = 0$. Therefore, for any $\mu\in\mathcal{M}$, we have that $b_{\omega,\mu}\leq b^*_\omega$. Then

$$\sup_{\mu\in\mathcal{M}} b_{\omega,\mu} \leq b^*_\omega. \quad (44)$$

Conversely, for any $\mu\in\mathcal{M}$, we have that $\mathbb{P}_\mu(f_\omega(\mathbf{X}^n[1,\nu-1]) > \sup_{\mu\in\mathcal{M}} b_{\omega,\mu}) = 0$. Then, $\frac{1}{|\mathcal{M}|}\sum_{\mu\in\mathcal{M}}\mathbb{P}_\mu(f_\omega(\mathbf{X}^n[1,\nu-1]) > \sup_{\mu\in\mathcal{M}} b_{\omega,\mu}) = 0$. This further implies that

$$b^*_\omega \leq \sup_{\mu\in\mathcal{M}} b_{\omega,\mu}. \quad (45)$$

Combining (44) and (45), we have that $\sup_{\mu\in\mathcal{M}} b_{\omega,\mu} = b^*_\omega$, and thus $\sup_{\mu\in\mathcal{M}}\text{esssup} f_\omega(\mathbf{X}^n[1,\nu-1]) = \text{esssup} f_\omega(\widehat{\mathbf{X}}^n[1,\nu-1])$. This implies that for any τ,

$$\text{WADD}_{\boldsymbol{s}}(\tau) = \widehat{\text{WADD}}_{\boldsymbol{s}}(\tau). \quad (46)$$

Step 2: The next step is to show $\widehat{\text{WADD}}_{\boldsymbol{s}}(\tau) \geq \widetilde{\text{WADD}}_{\boldsymbol{s}}(\tau)$. We first show that $\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}}\text{esssup} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right) \geq \text{esssup}\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right)$. Denote by $\widetilde{\mathbb{P}}^\nu$ the probability measure when the change is at ν, the pre- and post-change distributions are $\widetilde{\mathbb{P}}_0$ and $\widetilde{\mathbb{P}}^{s[t]}$ at time t, respectively. Let $\hat{b} = \sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}}\text{esssup} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right)$. For any $\omega\in\mathcal{N}_{\boldsymbol{s}}$, we have that $\widetilde{\mathbb{P}}^\nu\left(f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right) > \hat{b}\right) = 0$. Since $\mathcal{N}_{\boldsymbol{s}}$ is countable, and a countable union of sets of measure zero has measure zero, we then have that

$$\begin{aligned}&\widetilde{\mathbb{P}}^\nu\left(\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right) > \hat{b}\right) \\ &\leq \widetilde{\mathbb{P}}^\nu\left(\cup_{\omega\in\mathcal{N}_{\boldsymbol{s}}}\left\{f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right) > \hat{b}\right\}\right) = 0. \quad (47)\end{aligned}$$

Therefore,

$$\hat{b} \geq \text{esssup}\sup_{\omega\in\mathcal{N}_{\boldsymbol{s}}} f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right). \quad (48)$$

Before the change point ν, $\widehat{X}^n[t]$ and $\widetilde{X}^n[t]$ follow the same distribution. For any $T\geq\nu+1$, we have that

$$\begin{aligned}&\sup_{\substack{\{\sigma_\nu^{s[\nu]},\ldots,\sigma_T^{s[T]}\} \\ \in\mathcal{S}_{n,s[\nu]}\times,\ldots,\times\mathcal{S}_{n,s[T]}}} \sum_{t=\nu+1}^{T}(t-\nu) \\ &\times \mathbb{P}^{\boldsymbol{s},\nu}_{\sigma_\nu^{s[\nu]},\ldots,\sigma_T^{s[T]}}(\tau = t|\widehat{\mathbf{X}}^n[1,\nu-1])\end{aligned}$$

$$\geq \sum_{t=\nu+1}^{T}(t-\nu)\frac{1}{\mid \mathcal{S}_{n,s[\nu]} \mid \times \cdots \times \mid \mathcal{S}_{n,s[T]} \mid} \times \sum_{\substack{\{\sigma_\nu^{s[\nu]},\ldots,\sigma_T^{s[T]}\} \\ \in \mathcal{S}_{n,s[\nu]}\times,\ldots,\times \mathcal{S}_{n,s[T]}}} \mathbb{P}^{\boldsymbol{s},\nu}_{\sigma_\nu^{s[\nu]},\ldots,\sigma_T^{s[T]}}\left(\tau = t|\widehat{\mathbf{X}}^n[1,\nu-1]\right)$$

$$= \sum_{t=\nu+1}^{T}(t-\nu)\widetilde{\mathbb{P}}^{\boldsymbol{s},\nu}\left(\tau = t|\widetilde{\mathbf{X}}^n[1,\nu-1]\right), \tag{49}$$

where $\mathbb{P}^{\boldsymbol{s},\nu}_{\sigma_\nu^{s[\nu]},\ldots,\sigma_T^{s[T]}}$ denotes the probability measure when change point is ν, the trajectory is $\boldsymbol{s}$, the observations from time ν to time T are generated according to $\mathbb{P}^{s[\nu]}_{\sigma_\nu^{s[\nu]}},\ldots,\mathbb{P}^{s[T]}_{\sigma_T^{s[T]}}$. As $T\to\infty$, we have that

$$f_\omega\left(\widehat{\mathbf{X}}^n[1,\nu-1]\right) \geq \widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right], \tag{50}$$

From (48) and (50), we have that

$$\widehat{\mathrm{WADD}}_{\boldsymbol{s}}(\tau) \geq \mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] = \widetilde{\mathrm{WADD}}_{\boldsymbol{s}}(\tau). \tag{51}$$

Combining (46) and (51), it follows that

$$\mathrm{WADD}_{\boldsymbol{s}}(\tau) = \widehat{\mathrm{WADD}}_{\boldsymbol{s}}(\tau) \geq \widetilde{\mathrm{WADD}}_{\boldsymbol{s}}(\tau). \tag{52}$$

This holds for any trajectory $\boldsymbol{s}$. It then follows that

$$\mathrm{WADD}(\tau) \geq \sup_{\nu\geq 0}\sup_{\boldsymbol{s}}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] = \widetilde{\mathrm{WADD}}(\tau). \tag{53}$$

Step 3: The last step is to show that for any τ and any $\boldsymbol{\beta}$, $\widetilde{\mathrm{WADD}}(\tau) \geq \widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}}(\tau)$. Firstly, we will show that

$$\sup_{\boldsymbol{s}}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] \geq \mathrm{esssup}\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right]. \tag{54}$$

Let $c = \sup_{\boldsymbol{s}}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]]$. Denote by $\Lambda_{\boldsymbol{s}}$ the collection of all trajectory $\boldsymbol{s}$. For any $\boldsymbol{s}$, we have that

$$\widetilde{\mathbb{P}}^{\nu}\left(\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] > c\right) = 0. \tag{55}$$

Since $\Lambda_{\boldsymbol{s}}$ is countable, it then follows that

$$\widetilde{\mathbb{P}}^{\nu}\left(\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] > c\right) \leq \widetilde{\mathbb{P}}^{\nu}\left(\cup_{\boldsymbol{s}\in\Lambda_{\boldsymbol{s}}}\left\{\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] > c\right\}\right) = 0.$$

Therefore,

$$c = \sup_{\boldsymbol{s}}\mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] \geq \mathrm{esssup}\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right]. \tag{56}$$

For any $T\geq\nu+1$, we have that

$$\sup_{\boldsymbol{s}}\sum_{t=\nu+1}^{T}(t-\nu)\widetilde{\mathbb{P}}^{s[\nu],\ldots,s[T]}\left(\tau=t|\widetilde{\mathbf{X}}^n[1,\nu-1]\right) \geq \sum_{t=\nu+1}^{T}(t-\nu)\sum_{\{s[\nu],\ldots,s[T]\}\in\Lambda_{\boldsymbol{s}}^{\otimes(T-\nu+1)}}\beta_{s[\nu]}\times\cdots\times\beta_{s[T]} \times\widetilde{\mathbb{P}}^{s[\nu],\ldots,s[T]}\left(\tau=t|\widetilde{\mathbf{X}}^n[1,\nu-1]\right)$$

$$= \sum_{t=\nu+1}^{T}(t-\nu)\widetilde{\mathbb{P}}^{\boldsymbol{\beta},\nu}\left(\tau=t|\widetilde{\mathbf{X}}^n[1,\nu-1]\right), \tag{57}$$

where $\widetilde{\mathbb{P}}^{s[\nu],\ldots,s[T]}$ denotes the probability measure when the trajectory is $\boldsymbol{s}$, the observations from time ν to time T are generated according to $\widetilde{\mathbb{P}}^{s[\nu]},\ldots,\widetilde{\mathbb{P}}^{s[T]}$. As $T\to\infty$, we have

$$\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] \geq \widetilde{\mathbb{E}}^{\boldsymbol{\beta},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right]. \tag{58}$$

From (56) and (58), we have that

$$\widetilde{\mathrm{WADD}}(\tau) \geq \mathrm{esssup}\widetilde{\mathbb{E}}^{\boldsymbol{\beta},\nu}\left[(\tau-\nu)^+|\widetilde{\mathbf{X}}^n[1,\nu-1]\right] = \widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}}(\tau). \tag{59}$$

Combining (53) and (59), we have that for any τ and $\boldsymbol{\beta}$

$$\mathrm{WADD}(\tau) \geq \widetilde{\mathrm{WADD}}(\tau) \geq \widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}}(\tau). \tag{60}$$

For any $T\geq 1$, we have that

$$\inf_{\{\sigma_1^0,\ldots,\sigma_T^0\}\in\mathcal{S}_{n,0}{}^{\otimes T}}\sum_{t=1}^{T}t\mathbb{P}^{\infty}_{\sigma_1^0,\ldots,\sigma_T^0}(\tau=t) \leq \sum_{t=1}^{T}t\frac{1}{\mid\mathcal{S}_{n,0}\mid^T}\sum_{\{\sigma_1^0,\ldots,\sigma_T^0\}\in\mathcal{S}_{n,0}{}^{\otimes T}}\mathbb{P}^{\infty}_{\sigma_1^0,\ldots,\sigma_T^0}(\tau=t) = \sum_{t=1}^{T}t\widetilde{\mathbb{P}}^{\infty}(\tau=t). \tag{61}$$

As $T\to\infty$, we have that $\mathrm{ARL}(\tau)\leq\widetilde{\mathrm{ARL}}(\tau)$.

Therefore, for any stopping time τ satisfying $\mathrm{ARL}(\tau)\geq\gamma$, it will also satisfy $\widetilde{\mathrm{ARL}}(\tau)\geq\gamma$. We then have that $\mathcal{C}_\gamma\subseteq\widetilde{\mathcal{C}}_\gamma$.

Since (60) holds for any $\boldsymbol{\beta}$, it holds for $\boldsymbol{\beta}^*$. Problem (38) is a classical QCD problem. We have that for large γ [37],

$$\inf_{\tau\in\mathcal{C}_\gamma}\mathrm{WADD}(\tau) \geq \inf_{\tau\in\widetilde{\mathcal{C}}_\gamma}\widetilde{\mathrm{WADD}}_{\boldsymbol{\beta}^*}(\tau) \sim \frac{\log\gamma}{I_{\boldsymbol{\beta}^*}}(1+o(1)).$$

Appendix E
Proof of Lemma 1

The minimization of $I_{\boldsymbol{\beta}}$ is to solve the following problem:

$$\inf_{\boldsymbol{\beta}} \quad I_{\boldsymbol{\beta}}$$

$$\text{s.t.} \quad -\beta_k \le 0, \text{ for } k \in [1,K], \sum_{k=1}^{K} \beta_k - 1 = 0. \tag{62}$$

This is a convex optimization problem with linear constraints. Define the Lagrange function $L(\boldsymbol{\beta}, \eta, \boldsymbol{\mu})$:

$$L(\boldsymbol{\beta}, \eta, \boldsymbol{\mu}) = I_{\boldsymbol{\beta}} + \eta \left(\sum_{k=1}^{K} \beta_k - 1 \right) - \sum_{k=1}^{K} \mu_k \beta_k. \tag{63}$$

The minimizer $\boldsymbol{\beta}^*$ satisfies the Karush–Kuhn–Tucker(KKT) conditions: $\mu_k,\ \beta_k^* \ge 0,\ \mu_k \beta_k^* = 0,\ \sum_{k=1}^{K} \beta_k^* - 1 = 0$ and

$$\frac{\partial L}{\partial \beta_k}|_{\boldsymbol{\beta}^*} = \widetilde{\mathbb{E}}^k \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right] + 1 + \eta - \mu_k = 0, \tag{64}$$

where $\widetilde{\mathbb{E}}^k$ denotes the expectation under the distribution $\widetilde{\mathbb{P}}^k$.

When $\beta_k^* > 0$, we have $\mu_k = 0$. Therefore, for any $k, k' \in \mathcal{K}$ with $\beta_k^*, \beta_{k'}^* > 0$, we have

$$\widetilde{\mathbb{E}}^k \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right] = \widetilde{\mathbb{E}}^{k'} \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right] = -(1+\eta).$$

The set $\mathcal{K}$ can be divided into two disjoint parts $\mathcal{K}_1$ and $\mathcal{K}_2$: $\beta_k^* > 0$ if $k \in \mathcal{K}_1$ and $\beta_k^* = 0$ if $k \in \mathcal{K}_2$. We have that

$$\begin{aligned} I_{\boldsymbol{\beta}^*} &= \sum_{k \in K_1} \beta_k^* \widetilde{\mathbb{E}}^k \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right] \\ &= \widetilde{\mathbb{E}}^k \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right], \text{ for all } k \in \mathcal{K}_1. \end{aligned} \tag{65}$$

For all $k \in \mathcal{K}_2$, $\beta_k^* = 0$. By the KKT conditions, we have that $\mu_k \ge 0$. Therefore, for any $k \in \mathcal{K}_2, \widetilde{\mathbb{E}}^k[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)}] + 1 + \eta = \mu_k \ge 0$. We then have that for any $k \in \mathcal{K}_2$,

$$\widetilde{\mathbb{E}}^k \left[\log \frac{\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}(X^n)}{\widetilde{\mathbb{P}}_0(X^n)} \right] \ge I_{\boldsymbol{\beta}^*}. \tag{66}$$

Appendix F
Proof of Theorem 5

Due to the fact that $\max_{1 \le k \le t+1} \sum_{i=k}^{t} \ell_{\boldsymbol{\beta}^*}(X_i^n)$ has initial value 0 and remains non-negative, the delay is the largest when the change happens at $\nu = 0$. Therefore, for any $\boldsymbol{s}$, we have

$$\text{WADD}_{\boldsymbol{s}}(T_{\boldsymbol{\beta}^*}) = \sup_{\Omega_{\boldsymbol{s}}} \mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}]. \tag{67}$$

For any $T \ge \nu + 1$, we have that

$$\begin{aligned} &\sup_{\substack{\{\sigma_1^{s[1]}, \ldots, \sigma_T^{s[T]}\} \\ \in \mathcal{S}_{n,s[1]} \times, \ldots, \times \mathcal{S}_{n,s[T]}}} \sum_{t=1}^{T} t \mathbb{P}_{\sigma_1^{s[1]}, \ldots, \sigma_T^{s[T]}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} = t) \\ &= \sum_{t=1}^{T} t \frac{1}{| \mathcal{S}_{n,s[1]} | \times \cdots \times | \mathcal{S}_{n,s[T]} |} \\ &\quad \times \sum_{\substack{\{\sigma_1^{s[1]}, \ldots, \sigma_T^{s[T]}\} \\ \in \mathcal{S}_{n,s[1]} \times, \ldots, \times \mathcal{S}_{n,s[T]}}} \mathbb{P}_{\sigma_1^{s[1]}, \ldots, \sigma_T^{s[T]}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} = t) \\ &= \sum_{t=1}^{T} t \widetilde{\mathbb{P}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} = t). \end{aligned} \tag{68}$$

As $T \to \infty$, we have that $\sup_{\Omega_{\boldsymbol{s}}} \mathbb{E}_{\Omega_{\boldsymbol{s}}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}] = \widetilde{\mathbb{E}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}] = \widetilde{\text{WADD}}_{\boldsymbol{s}}(T_{\boldsymbol{\beta}^*})$. For any $\boldsymbol{s}$, we have $\text{WADD}_{\boldsymbol{s}}(T_{\boldsymbol{\beta}^*}) = \widetilde{\text{WADD}}_{\boldsymbol{s}}(T_{\boldsymbol{\beta}^*})$. Therefore, $\text{WADD}(T_{\boldsymbol{\beta}^*}) = \widetilde{\text{WADD}}(T_{\boldsymbol{\beta}^*})$ by taking sup over $\boldsymbol{s}$ on both sides. It then follows that

$$\text{WADD}(T_{\boldsymbol{\beta}^*}) = \widetilde{\text{WADD}}(T_{\boldsymbol{\beta}^*}) = \sup_{\boldsymbol{s}} \widetilde{\mathbb{E}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}]. \tag{69}$$

Let $0 < \epsilon < I_{\boldsymbol{\beta}^*}$ and $n_b = \frac{b}{I_{\boldsymbol{\beta}^*} - \epsilon}$. For any trajectory $\boldsymbol{s}$, from the sum-integral inequality, we have that

$$\begin{aligned} \widetilde{\mathbb{E}}^{\boldsymbol{s},0} \left[\frac{T_{\boldsymbol{\beta}^*}}{n_b} \right] &= \int_0^{\infty} \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{T_{\boldsymbol{\beta}^*}}{n_b} > x \right) \mathrm{d}x \\ &\le \sum_{t=1}^{\infty} \widetilde{\mathbb{P}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} > t n_b) + 1. \end{aligned} \tag{70}$$

For any $\boldsymbol{s}$, we have that

$$\begin{aligned} &\widetilde{\mathbb{P}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} > t n_b) = \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\max_{1 \le k \le t n_b} \max_{1 \le i \le k} \sum_{j=i}^{k} \ell_{\boldsymbol{\beta}^*}(X_j^n) < b \right) \\ &\le \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\max_{1 \le i \le m n_b} \sum_{j=i}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n) < b, \forall m \in [t] \right) \\ &\le \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n) < b, \forall m \in [t] \right) \\ &= \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{\beta}^*} - \epsilon, \forall m \in [t] \right) \\ &= \prod_{m=1}^{t} \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{\beta}^*} - \epsilon \right). \end{aligned} \tag{71}$$

It then follows that

$$\begin{aligned} &\sup_{\boldsymbol{s}} \sum_{t=1}^{\infty} \widetilde{\mathbb{P}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*} > t n_b) \\ &\le \sup_{\boldsymbol{s}} \sum_{t=1}^{\infty} \prod_{m=1}^{t} \widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{\beta}^*} - \epsilon \right). \end{aligned}$$

Then we will bound $\widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{\beta}^*} - \epsilon \right)$.

Let $I_{\boldsymbol{s}_m} = \widetilde{\mathbb{E}}^{\boldsymbol{s},0} \left[\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} \right]$. From (65) and (66),

$$I_{\boldsymbol{s}_m} = \frac{1}{n_b} \sum_{j=(m-1)n_b+1}^{m n_b} \widetilde{\mathbb{E}}^{s[j]} \left[\ell_{\boldsymbol{\beta}^*}(X_j^n) \right] \ge I_{\boldsymbol{\beta}^*}. \tag{72}$$

It then follows that for any $\boldsymbol{s}$ and m

$$\widetilde{\mathbb{P}}^{\boldsymbol{s},0} \left(\frac{\sum_{j=(m-1)n_b+1}^{m n_b} \ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{\beta}^*} - \epsilon \right)$$

$$\leq \widetilde{\mathbb{P}}^{\boldsymbol{s},0}\left(\frac{\sum_{j=(m-1)n_b+1}^{mn_b}\ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} < I_{\boldsymbol{s}_m}-\epsilon\right)$$
$$\leq \widetilde{\mathbb{P}}^{\boldsymbol{s},0}\left(\left|\frac{\sum_{j=(m-1)n_b+1}^{mn_b}\ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} - I_{\boldsymbol{s}_m}\right| > \epsilon\right). \quad (73)$$

Assume that $\max_{k\in[1,K]}\widetilde{\mathbb{E}}^k[\ell_{\boldsymbol{\beta}^*}(X^n)^2]<\infty$. Let $\sigma^2 = \max_{k\in[1,K]}\mathrm{Var}_{\widetilde{\mathbb{P}}^k}(\ell_{\boldsymbol{\beta}^*}(X^n))$ where $\mathrm{Var}_{\widetilde{\mathbb{P}}^k}$ denotes the variance under the distribution $\widetilde{\mathbb{P}}^k$. By Chebychev's inequality,

$$\widetilde{\mathbb{P}}^{\boldsymbol{s},0}\left(\left|\frac{\sum_{j=(m-1)n_b+1}^{mn_b}\ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b} - I_{\boldsymbol{s}_m}\right| > \epsilon\right)$$
$$\leq \mathrm{Var}_{\widetilde{\mathbb{P}}^{\boldsymbol{s}}}\left(\frac{\sum_{j=(m-1)n_b+1}^{mn_b}\ell_{\boldsymbol{\beta}^*}(X_j^n)}{n_b}\right)\frac{1}{\epsilon^2}$$
$$= \frac{1}{\epsilon^2 n_b^2}\sum_{j=(m-1)n_b+1}^{mn_b}\mathrm{Var}_{\widetilde{\mathbb{P}}^{s[j]}}(\ell_{\boldsymbol{\beta}^*}(X_j^n))$$
$$\leq \frac{\sum_{j=(m-1)n_b+1}^{mn_b}\sigma^2}{n_b^2\epsilon^2} = \frac{\sigma^2}{n_b\epsilon^2}. \quad (74)$$

Let $\delta = \frac{\sigma^2}{n_b\epsilon^2}$. From (70) and (74), we have that

$$\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},0}\left[\frac{T_{\boldsymbol{\beta}^*}}{n_b}\right] \leq 1+\sup_{\boldsymbol{s}}\sum_{t=1}^{\infty}\widetilde{\mathbb{P}}^{\boldsymbol{s},0}(T_{\boldsymbol{\beta}^*}>tn_b)$$
$$\leq 1+\sum_{t=1}^{\infty}\left(\frac{\sigma^2}{n_b\epsilon^2}\right)^t = 1+\sum_{t=1}^{\infty}\delta^t = \frac{1}{1-\delta}. \quad (75)$$

Therefore, we have

$$\sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}] \leq \frac{b}{(I_{\boldsymbol{\beta}^*}-\epsilon)(1-\delta)}. \quad (76)$$

(76) holds for all ϵ. It then follows that as $b\to\infty$,

$$\mathrm{WADD}(T_{\boldsymbol{\beta}^*}) = \sup_{\boldsymbol{s}}\widetilde{\mathbb{E}}^{\boldsymbol{s},0}[T_{\boldsymbol{\beta}^*}] \leq \frac{b}{I_{\boldsymbol{\beta}^*}}(1+o(1)). \quad (77)$$

For the ARL lower bound, for any $T\geq 1$, we have that

$$\inf_{\{\sigma_1^0,\ldots,\sigma_T^0\}\in\mathcal{S}_{n,0}{}^{\otimes T}}\sum_{t=1}^{T}t\mathbb{P}^{\infty}_{\sigma_1^0,\ldots,\sigma_T^0}(T_{\boldsymbol{\beta}^*}=t)$$
$$= \sum_{t=1}^{T}t\frac{1}{|\mathcal{S}_{n,0}|^T}\sum_{\{\sigma_1^0,\ldots,\sigma_T^0\}\in\mathcal{S}_{n,0}{}^{\otimes T}}\mathbb{P}^{\infty}_{\sigma_1^0,\ldots,\sigma_T^0}(T_{\boldsymbol{\beta}^*}=t)$$
$$= \sum_{t=1}^{T}t\widetilde{\mathbb{P}}^{\infty}(T_{\boldsymbol{\beta}^*}=t). \quad (78)$$

As $T\to\infty$, we have that $\mathrm{WARL}(T_{\boldsymbol{\beta}^*}) = \widetilde{\mathrm{ARL}}(T_{\boldsymbol{\beta}^*})$. $T_{\boldsymbol{\beta}^*}$ is the CuSum algorithm for a simple QCD problem with pre-change distribution $\widetilde{\mathbb{P}}_0$ and post-change distribution $\widetilde{\mathbb{P}}^{\boldsymbol{\beta}^*}$. From the optimal property of CuSum algorithm in [39] and [45], we have that when $b=\log\gamma$, $\mathrm{WARL}(T_{\boldsymbol{\beta}^*}) = \widetilde{\mathrm{ARL}}(T_{\boldsymbol{\beta}^*}) \geq \gamma$.

REFERENCES

[1] Z. Sun and S. Zou, "Quickest dynamic anomaly detection in anonymous heterogeneous sensor networks," in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 106–111.

[2] T. E. Humphreys et al., "Assessing the spoofing threat: Development of a portable GPS civilian spoofer," in *Proc. 21st Int. Tech. Meeting Satell. Division Inst. Navigation*, 2008, pp. 2314–2325.

[3] L. Keller, M. Jafari Siavoshani, C. Fragouli, K. Argyraki, and S. Diggavi, "Identity aware sensor networks," in *Proc. IEEE Int. Conf. Commun. Comput. Control Appl.*, 2009, pp. 2177–2185.

[4] W. N. Chen and I. H. Wang, "Anonymous heterogeneous distributed detection: Optimal decision rules, error exponents, and the price of anonymity," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7390–7406, Nov. 2019.

[5] W. Li and Y. Huang, "Bandwidth-constrained distributed quickest change detection in heterogeneous sensor networks: Anonymous vs non-anonymous settings," 2022, *arXiv:2202.02697*.

[6] S. Marano and P. K. Willett, "Algorithms and fundamental limits for unlabeled detection using types," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2022–2035, Apr. 2019.

[7] S. Marano and P. Willett, "Making decisions by unlabeled bits," *IEEE Trans. Signal Process.*, vol. 68, pp. 2935–2947, 2020.

[8] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3237–3253, May 2018.

[9] S. Haghighatshoar and G. Caire, "Signal recovery from unlabeled samples," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1242–1257, May 2018.

[10] A. Abid, A. Poon, and J. Zou, "Linear regression with shuffled labels," 2017, *arXiv:1705.01342*.

[11] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 1040–1044.

[12] Z. Liu and J. Zhu, "Signal detection from unlabeled ordered samples," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2431–2434, Dec. 2018.

[13] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with shuffled data: Statistical and computational limits of permutation recovery," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3286–3300, May 2018.

[14] G. Elhami, A. Scholefield, B. Bejar Haro, and M. Vetterli, "Unlabeled sensing: Reconstruction algorithm and theoretical guarantees," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 4566–4570.

[15] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334–2345, Jun. 2008.

[16] G. Wang et al., "Signal amplitude estimation and detection from unlabeled binary quantized samples," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4291–4303, Aug. 2018.

[17] Z. Sun, S. Zou, R. Zhang, and Q. Li, "Quickest change detection in anonymous heterogeneous sensor networks," *IEEE Trans. Signal Process.*, vol. 70, pp. 1041–1055, 2022.

[18] G. Rovatsos, G. V. Moustakides, and V. V. Veeravalli, "Quickest detection of moving anomalies in sensor networks," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 762–773, Jun. 2021.

[19] A. G. Tartakovsky and V. V. Veeravalli, "Change-point detection in multichannel and distributed systems," *Appl. Sequential Methodol.: Real-World Examples Data Anal.*, vol. 173, pp. 339–370, 2004.

[20] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blazek, and H. Kim, "A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3372–3382, Sep. 2006.

[21] Y. Mei, "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[22] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *Ann. Statist.*, vol. 41, pp. 670–692, 2013.

[23] G. Fellouris and G. Sokolov, "Second-order asymptotic optimality in multisensor sequential change detection," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3662–3675, Jun. 2016.

[24] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a Markov process across a sensor array," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1961–1981, Apr. 2010.

[25] O. Hadjiliadis, H. Zhang, and H. V. Poor, "One shot schemes for decentralized quickest change detection," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3346–3359, Jul. 2009.

[26] M. Ludkovski, "Bayesian quickest detection in sensor arrays," *Sequential Anal.*, vol. 31, no. 4, pp. 481–504, 2012.

[27] V. V. Veeravalli, "Decentralized quickest change detection," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1657–1665, May 2001.
[28] A. G. Tartakovsky and V. V. Veeravalli, "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Anal.*, vol. 27, no. 4, pp. 441–475, 2008.
[29] S. Zou, V. V. Veeravalli, J. Li, D. Towsley, and A. Swami, "Distributed quickest detection of significant events in networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 8454–8458.
[30] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 494–514, Jun. 2021.
[31] S. Zou, V. V. Veeravalli, J. Li, and D. Towsley, "Quickest detection of dynamic events in networks," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2280–2295, Apr. 2020.
[32] G. Rovatsos, G. V. Moustakides, and V. V. Veeravalli, "Quickest detection of a dynamic anomaly in a sensor network," in *Proc. IEEE 53rd Asilomar Conf. Signals Syst. Comput.*, 2019, pp. 98–102.
[33] G. Rovatsos, S. Zou, and V. V. Veeravalli, "Sequential algorithms for moving anomaly detection in networks," *Sequential Anal.*, vol. 39, no. 1, pp. 6–31, 2020.
[34] S. Zou, G. Fellouris, and V. V. Veeravalli, "Quickest change detection under transient dynamics: Theory and asymptotic analysis," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1397–1412, Mar. 2019.
[35] R. Zhang, Y. Xie, R. Yao, and F. Qiu, "Online detection of cascading change-points," in *Proc. 58th Annu. Allerton Conf. Commun., Control, Comput.*, 2022, pp. 1–6, doi: 10.1109/Allerton49937.2022.9929381.
[36] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *Ann. Statist.*, vol. 23, pp. 255–271, 1995.
[37] T. Leung Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.
[38] T. Banerjee and V. V. Veeravalli, "Data-efficient minimax quickest change detection with composite post-change distribution," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 5172–5184, Sep. 2015.
[39] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, 1971.
[40] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing Statistical Hypotheses*. Berlin, Germany: Springer, 2005.
[41] D. Williams, *Probability With Martingales*. New York, NY, USA: Cambridge Univ. Press, 1991.
[42] M. Pollak et al., "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, no. 1, pp. 206–227, 1985.
[43] T. M. Cover, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.
[44] Z. Sutton, P. Willet, and S. Marano, "Sensor network target detection with unlabeled observations," in *Proc. IEEE Aerosp. Conf.*, 2020, pp. 1–10.
[45] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, 1986.

Zhongchang Sun (Student, IEEE) received the B.S. degree from the Beijing Institute of Technology, Beijing, China in 2019. He is working toward the Ph.D. degree with the Department of Electrical Engineering, University at Buffalo, the State University of New York, Buffalo, NY, USA. His research interests include hypothesis testing, quickest change detection, and distributionally robust optimization.

Shaofeng Zou received the B.E. degree (with honors) from Shanghai Jiao Tong University, Shanghai, China, in 2011, the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2016. He is currently an Assistant Professor with the Department of Electrical Engineering, University at Buffalo, the State University of New York, Buffalo, NY, USA. During 2016–2018, he was a Postdoctoral Research Associate with the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Champaign, IL, USA. His research interests include reinforcement learning, machine learning, statistical signal processing and information theory. He was the recipient of the National Science Foundation CRII Award in 2019 and the 2023 AAAI Distinguished Paper Award.