Taylor & Francis
Taylor & Francis Group

Sequential Analysis
Design Methods and Applications

ISSN: 0747-4946 (Print) 1532-4176 (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

Sequential algorithms for moving anomaly detection in networks

Georgios Rovatsos, Shaofeng Zou & Venugopal V. Veeravalli

SEQUENTIAL ANALYSIS
2020, VOL. 39, NO. 1, 6–31
https://doi.org/10.1080/07474946.2020.1726678

Sequential algorithms for moving anomaly detection in networks

Georgios Rovatsos[a], Shaofeng Zou[b], and Venugopal V. Veeravalli[a]

[a]Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA; [b]Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, New York, USA

ABSTRACT

The problem of quickest moving anomaly detection in networks is studied. Initially, the observations are generated according to a pre-change distribution. At some unknown but deterministic time, an anomaly emerges in the network. At each time instant, one node is affected by the anomaly and receives data from a post-change distribution. The anomaly moves across the network, and the node that it affects changes with time. However, the trajectory of the moving anomaly is assumed to be unknown. A discrete-time Markov chain is employed to model the unknown trajectory of the moving anomaly in the network. A windowed generalized likelihood ratio–based test is constructed and is shown to be asymptotically optimal. Other detection algorithms including the dynamic Shiryaev-Roberts test, a quickest change detection algorithm with recursive change point estimation, and a mixture cumulative sum (CUSUM) algorithm are also developed for this problem. Lower bounds on the mean time to false alarm are developed. Numerical results are further provided to compare their performances.

ARTICLE HISTORY
Received 21 August 2019
Revised 15 September 2019, 28 October 2019
Accepted 21 November 2019

KEYWORDS
Dynamic anomaly; generalized likelihood ratio; hidden Markov model; quickest change detection; sensor networks

SUBJECT CLASSIFICATIONS
62L10; 62M02; 60G40; 62F05

1. Introduction

Consider a system monitored in real time by a set of interconnected sensor nodes. At some unknown time, an anomaly appears in the network and changes the data generating distribution of some unknown node. As the anomaly moves around in the network randomly, the node that it affects also changes with time. At each time instant, if a node is not affected by the anomaly, then it receives independent and identically distributed (i.i.d.) samples from a prechange distribution; if a node is affected by the anomaly, then it receives i.i.d. samples from a postchange distribution. Observations are taken sequentially from all of the nodes. The goal here is to detect the appearance of the anomaly as quickly as possible, subject to false alarm constraints. In this article, we assume that the anomaly moves around in the network along the network edges (see Figure. 1), and the trajectory of the moving anomaly is unknown. After the anomaly emerges in the network, the data generating distribution of the network dynamically

CONTACT Venugopal V. Veeravalli ✉ vvv@illinois.edu ECE Department, University of Illinois at Urbana–Champaign, 315 Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801, USA.

Recommended by Nitis Mukhopadhyay
This work was presented in part in the International Symposium on Information Theorem (ISIT), 2019, Paris, France (see Rovatsos et al., 2019).

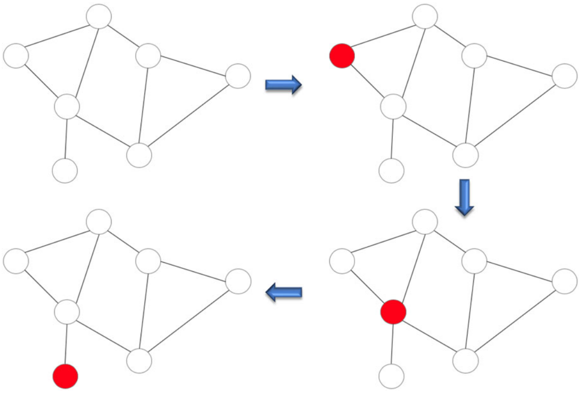

Figure 1. Dynamic anomaly in a network.

changes with time, as the node affected by the anomaly changes with time. In this article, we focus on the case where the anomaly only affects one node at each time.

The problem studied in this article fits into the framework of quickest change detection (QCD; for a detailed review of QCD theory, see Poor and Hadjiliadis, 2009; Tartakovsky et al., 2014; Veeravalli and Banerjee, 2013), which models a wide range of applications, including critical infrastructure monitoring, environmental monitoring, fraud detection, financial surveillance, cognitive radio, detection of bioterrorist attacks, and intrusion detection in computer networks (see Fienberg and Shmueli, 2005; Frisn, 2009; Lai et al., 2008; Mechitov et al., 2004; Rice et al., 2010; Rovatsos et al., 2016; Rovatsos, Jiang, et al. 2017; Tartakovsky et al., 2006). In the QCD problem, observations are sampled sequentially, and initially follow a nominal distribution. At some unknown time (change point), an event occurs and leads to a change in the data generating distribution of the observations. The goal is to detect this change as quickly as possible, subject to false alarm constraints.

In this article, we model the trajectory of the moving anomaly using a discrete-time Markov chain (DTMC), where each state of the DTMC corresponds to a distinct node being affected. However, the state of the DTMC is not directly observable. Instead, noisy samples whose distribution depends on the state of the DTMC are observed. Thus, the observed samples follow a hidden Markov model (HMM). Detecting the emerging of a moving anomaly can be viewed as detecting a change from an i.i.d. model to a HMM.

1.1. Related work

The problem of moving anomaly detection studied in this article is related to the multichannel QCD problem studied in Zou and Veeravalli (2018), Tartakovsky and Veeravalli (2004), Mei (2010), Xie and Siegmund (2013), Fellouris and Sokolov (2016), Raghavan and Veeravalli (2010), Ludkovski (2012), and Hadjiliadis et al. (2009), where some event leads to a persistent change in the data generating distributions of a subset of the nodes in the network. The difference between the multi-channel QCD problem and our problem is that in our problem, the anomaly moves around in the network, and thus the change is not persistent at any particular node, but it is persistent if we view the entire network as a whole. As a result, employing a cumulative sum (CUSUM) test for each node and declaring a change by combining them is not applicable in our setting.

Our work is also related to the QCD problem under transient dynamics studied in Zou et al. (2019), Rovatsos, Zou, and Veeravalli (2017), Rovatsos, Jiang, et al. (2017), and Moustakides and Veeravalli (2016), where the change in the probability distribution of the observations does not happen instantaneously but through a sequence of transient phases each corresponding to a distinct data generating distribution. In contrast, in the current article the trajectory of the moving anomaly is unknown, and the number of possible trajectories scales exponentially with time; therefore, the algorithms and analysis developed in Zou et al. (2019), Rovatsos, Zou, and Veeravalli (2017), Rovatsos, Jiang, et al. (2017), and Moustakides and Veeravalli (2016) cannot be directly applied.

The problem of QCD in HMMs has been studied in prior work; for example, see Fuh (2003, 2004), Fuh and Mei (2015), Fuh and Tartakovsky (2019), and Chen and Willett (1997). In Fuh (2003), the minimax setting was studied, where the change point is assumed to be deterministic but unknown. For this problem, the generalized likelihood ratio (GLR)-based test does not have a recursion and is thus not computationally efficient. In Fuh (2003), instead of using the GLR approach, a recursive test was designed using an approximate conditional probability distribution, and was further shown to be first-order asymptotically optimal. This recursive test was further studied in Fuh and Mei (2015) for two-state HMMs, and it was shown to be equivalent to a quasi-GLR scheme with respect to a pseudo postchange measure. Recently, the Bayesian setting was investigated in Fuh and Tartakovsky (2019), where the change point was modeled as a random variable with known distribution. A different formulation of QCD in HMMs was proposed in Moustakides (2019), and Shewhart-type tests (the Shewhart test was intially introduced in Shewhart, 1925) were constructed and were shown to exactly maximize the worst-case detection probability subject to false alarm constraints.

The main differences between our work and the work in Fuh (2003) and Fuh and Tartakovsky (2019) are as follows: (i) we focus on the application of sequential moving anomaly detection in networks and formulate it as the problem of QCD in HMMs; (ii) the work in Fuh (2003) and Fuh and Tartakovsky (2019) considers the setting where the observations are generated according to a HMM, and at some unknown but deterministic time, the parameters of the HMM change abruptly, whereas in our problem, the data before the change point are i.i.d. distributed, the data after the change are generated by an HMM, and the prechange data are independent from the post-change data; (iii) we construct a windowed GLR test and establish its first-order asymptotic optimality using a technique introduced in Lai (1998); (iv) we also construct several alternative algorithms, including the dynamic Shiryaev-Roberts test, the QCD test with change point estimation, and the mixture CUSUM algorithm; and (v) we comprehensively compare these algorithms numerically, and investigate the conditions under which each of these tests should be preferred.

1.2. Main contributions

We summarize our main contributions in this article as follows:

1. We study the problem of quickest moving anomaly detection in the networks. We model the trajectory of the moving anomaly using a discrete-time Markov chain, and formulate the quickest detection problem as a quickest detection problem in HMMs.
2. We first construct the windowed GLR algorithm and show that it is first-order asymptotically optimal. However, this approach, although it scans over only a finite window, is not computationally efficient. We therefore develop a number of alternative approaches to address this challenge.

   For the first alternative, we use a Bayesian approach, where the change point is modeled as a geometric random variable with parameter ρ. Under this setting, we obtain a test that can be updated recursively. We then let $\rho \to 0$, so that the test does not depend on ρ.

   The second alternative algorithm is motivated by the idea of recursive change point estimation used in Lau et al. (2019) and Lorden and Pollak (2008).

   The third alternative algorithm, a mixture CUSUM algorithm, can be applied if the transition probabilities are not available in practice. The algorithm tests a change from the pre-change distribution to a mixture of postchange distributions. Such an algorithm is computationally efficient and is numerically shown to perform as well as our windowed GLR test.

   For all of the alternative algorithms, we also develop the lower bounds on the mean times to false alarms (MTFAs) for practical false alarm control.
3. We conduct comprehensive numerical experiments to compare the proposed algorithms.

1.3. Organization of the article

The rest of the article is organized as follows. In Section 2, we introduce the problem model. In Section 3 we present the universal asymptotic lower bound on the worst-case average detection delay. In Section 4 we construct the windowed GLR test and demonstrate its first-order asymptotic optimality. In Section 5, we present the dynamic Shiryaev-Roberts algorithm, the QCD algorithm with recursive change point estimation, and the mixture CUSUM algorithm and develop the lower bounds on their MTFA. In Section 6, we review and instantiate Fuh's test proposed in Fuh (2003) for our problem. In Section 7 we numerically compare the different detection schemes presented in this work. Finally, in Section 8, we conclude our article.

2. Problem model

2.1. Stochastic model

Consider a network of L nodes denoted by $\mathcal{L} = \{1, \ldots, L\}$. Define by $\boldsymbol{X}_k = [X_{1,k}, \ldots, X_{L,k}]^\top$ the vector of observations obtained by the nodes at time k, where $X_{\ell,k}$ denotes the measurement provided by node ℓ at time k. At some *deterministic* but *unknown* time ν, an anomaly appears in the network and affects one of the nodes. In particular, at each time instant k where $k \geq \nu$, the index of the affected node is denoted by $S_k \in \mathcal{L}$, which is not directly observable. For notational convenience, if there is no anomaly—that is, $k < \nu$—we let $S_k \triangleq 0$. We note that in this article we focus on the scenario where there is one and only one affected node at each time after the anomaly appears in the network. The results in this article can be easily generalized to the case with multiple nodes being affected at the same time.

It is assumed that before the anomaly appears $(k < \nu)$, the samples generated by node ℓ are i.i.d. generated by a probability density function (p.d.f.) $f_{\ell,0}$ for all $\ell \in \mathcal{L}$, and that the samples are independent across different nodes. Then, the joint distribution of $\boldsymbol{X}_1, ..., \boldsymbol{X}_k$ for $k < \nu$ is given by

$$f_0(\boldsymbol{X}_1, ..., \boldsymbol{X}_k) = \prod_{j=1}^{k} \prod_{i=1}^{L} f_{i,0}(X_{i,j}), \text{ for } k \leq \nu. \tag{2.1}$$

If $k \geq \nu$ and at time k the affected node is ℓ—that is, $S_k = \ell$—then $X_{\ell,k}$ is generated according to a post-change distribution $f_{\ell,1}$, and the samples of the other nodes $X_{i,k}$'s still follow the prechange distribution $f_{i,0}$, for $i \neq \ell$. We further assume that conditioned on ν and S_k, the samples across different nodes are independent. Specifically, conditioning on $S_k = \ell$ and $k \geq \nu, \boldsymbol{X}_k$ is generated according to the following joint probability distribution:

$$f_\ell(\boldsymbol{X}_k) \triangleq \left(\prod_{i \neq \ell} f_{i,0}(X_{i,k}) \right) f_{\ell,1}(X_{\ell,k}). \tag{2.2}$$

We denote by $\mathbb{P}_\nu(\cdot)$ $(\mathbb{E}_\nu[\cdot])$ the probability measure (expectation) when the anomaly occurs at time ν. To be more specific, we denote by $\mathbb{P}_\infty(\cdot)$ $(\mathbb{E}_\infty[\cdot])$ the probability measure (expectation) when $\nu = \infty$; that is, when there is no anomaly. We further let $\mathcal{B}(\mathbb{R}^L)$ denote the Borel σ-algebra with respect to $\mathbb{R}^L$, and μ is a σ-finite measure on $\mathbb{R}^L$.

In this article, we study the case where the anomaly is dynamic; that is, S_k changes with time k for $k \geqslant \nu$. We model the change of S_k as a DTMC. More specifically, for any $k \geq \nu$,

$$\mathbb{P}_\nu(S_{k+1} | S_1, ..., S_k, \boldsymbol{X}_1, ..., \boldsymbol{X}_k) = \mathbb{P}_\nu(S_{k+1} | S_k) \triangleq \lambda_{S_k, S_{k+1}}, \tag{2.3}$$

where $\lambda_{i,j} \in [0, 1]$ denotes the probability that the anomaly moves from node i to node j for any $i, j \in \mathcal{L}$. Furthermore, for any k, conditioned on S_k, $\boldsymbol{X}_k$ is independent from anything else. To be more explicit, for any $B \in \mathcal{B}(\mathbb{R}^L)$, we have that

$$\mathbb{P}_\nu(\boldsymbol{X}_k \in B | \boldsymbol{X}_1, \boldsymbol{X}_2, ..., S_1, S_2, ...) = \mathbb{P}_\nu(\boldsymbol{X}_k \in B | S_k) = \int_B f_{S_k}(\boldsymbol{X}_k) d\mu. \tag{2.4}$$

Before the anomaly appears in the network, the observations from the nodes are i.i.d. according to the pre-change distribution f_0 in (2.1). After the anomaly emerges in the network, the underlying stochastic process of this problem can be viewed as an HMM, where $\{S_k\}_{k=\nu}^{\infty}$ is a finite state Markov chain, which is not directly observable. The transition probability matrix is given by $[\lambda_{i,j}]_{i,j \in \mathcal{L}}$. Then the sequence of random vectors $\{\boldsymbol{X}_k\}_{k=\nu}^{\infty}$ is adjoint to this Markov chain according to (2.3) and (2.4). Therefore, after the anomaly appears in the network, there is a change in the underlying stochastic process from an i.i.d. model to an HMM.

2.2. Performance criteria

The goal is to design stopping times that can detect the anomaly at time ν as quickly as possible while ensuring that the frequency of false alarm events is below an acceptable level. A stopping time τ with respect to the observed sequence $\{\boldsymbol{X}_k\}_{k=1}^{\infty}$ is an integer-valued random variable, such that for each $k \geq 1, \{\tau \leq k\} \in \sigma(\boldsymbol{X}_1, ..., \boldsymbol{X}_k)$, where $\sigma(\boldsymbol{X}_1, ..., \boldsymbol{X}_k)$ denotes the σ-algebra generated by $\boldsymbol{X}_1, ..., \boldsymbol{X}_k$. In other words, the decision to stop at time k is determined only by $\boldsymbol{X}_1, ..., \boldsymbol{X}_k$.

In this article, we focus on the minimax setting, where the change point ν is assumed to be deterministic and unknown. In order to measure the frequency of false alarm events, we define the mean time to false alarm (MTFA) as $\mathbb{E}_\infty[\tau]$. We further define the worst-case average detection delay of a stopping time τ under Lorden's criterion, introduced in Lorden (1971), by

$$\text{WADD}(\tau) = \sup_{\nu \geq 1} \text{ess sup} \mathbb{E}_\nu \left[(\tau - \nu + 1)^+ | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right], \tag{2.5}$$

where $(x)^+ = \max\{x, 0\}$ and under Pollak's criterion (see Pollak, 1985) by

$$\text{CADD}(\tau) = \sup_{\nu \geq 1} \mathbb{E}_\nu [\tau - \nu | \tau \geq \nu]. \tag{2.6}$$

The Worst-Case Average Detection Delay (WADD) metric is a more pessimistic metric than the Conditional Average Detection Delay (CADD) metric (for more details, see, e.g., Veeravalli and Banerjee, 2013); in particular,

$$\text{WADD}(\tau) \geq \text{CADD}(\tau). \tag{2.7}$$

In this article, we aim to design a stopping rule τ to minimize WADD and CADD subject to a constraint on the MTFA

$$\mathbb{E}_\infty[\tau] \geq \gamma,$$

where $\gamma > 0$ is a predetermined constant. In particular, our goal is to design stopping rules that solve the following constrained stochastic optimization problems:

$$\inf_{\tau : \mathbb{E}_\infty[\tau] \geq \gamma} \text{WADD}(\tau), \tag{2.8}$$

$$\inf_{\tau : \mathbb{E}_\infty[\tau] \geq \gamma} \text{CADD}(\tau). \tag{2.9}$$

2.3. Assumptions on the HMM

We assume that the DTMC defined in (2.3) has a stationary distribution denoted by a vector $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_L]^\top$ and that the DTMC is also initialized with $\boldsymbol{\alpha}$; that is, that for all $\ell \in \mathcal{L}, \mathbb{P}_\nu(S_\nu = \ell) = \alpha_\ell$. For $t_1 \leq \nu \leq t_2$, we then denote by

$$g_\nu(\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{t_2}) \triangleq f_0(\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{\nu-1}) \times \sum_{i_\nu, ..., i_{t_2} \in \mathcal{L}} \left\{ \alpha_{i_\nu} f_{i_\nu}(\boldsymbol{X}_\nu) \times \prod_{j=\nu+1}^{t_2} [\lambda_{i_{j-1}, i_j} f_{i_j}(\boldsymbol{X}_j)] \right\} \tag{2.10}$$

the joint probability distribution of $\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{t_2}$ conditioned on a change point ν. We also define by

$$\Lambda_\nu(\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{t_2}) \triangleq \frac{g_\nu(\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{t_2})}{f_0(\boldsymbol{X}_{t_1}, ..., \boldsymbol{X}_{t_2})}$$

the likelihood ratio of $\boldsymbol{X}_{t_1}, \ldots, \boldsymbol{X}_{t_2}$ between the hypothesis that the anomaly appears at time ν and the hypothesis that the anomaly never appears.

For the asymptotic analysis in this article, we make the following assumptions on the DTMC. In particular we assume the following:

(C.1) Under $\mathbb{P}_1(\cdot)$, the DTMC $\{S_k\}_{k=1}^{\infty}$ is ergodic (positive recurrent, irreducible, and aperiodic). Furthermore, if we define the random matrices

$$\mathsf{M} = \begin{bmatrix} f_1(\boldsymbol{X}_1) & \ldots & 0 \\ \vdots & \ddots & \\ 0 & & f_L(\boldsymbol{X}_1) \end{bmatrix}$$

and

$$\mathsf{N} = \begin{bmatrix} \lambda_{1,1} f_1(\boldsymbol{X}_2) & \ldots & \lambda_{1,L} f_L(\boldsymbol{X}_2) \\ \vdots & \ddots & \\ \lambda_{L,1} f_1(\boldsymbol{X}_2) & & \lambda_{L,L} f_L(\boldsymbol{X}_2) \end{bmatrix},$$

then M and N are almost surely invertible under $\mathbb{P}_1(\cdot)$ and $\mathbb{P}_\infty(\cdot)$.

(C.2) There exists $r > 0$ such that $\int_{\mathbb{R}} |x|^{r+1} f_{\ell,0}(x) d\mu < \infty$, and $\int_{\mathbb{R}} |x|^{r+1} f_{\ell,1}(x) d\mu < \infty$, for all $\ell \in \mathcal{L}$.

The above assumptions cover many interesting examples of HMMs, as noted in Fuh and Tartakovsky (2019).

We further define the following effective Kullback-Leibler (KL) number:

$$I = \lim_{n\to\infty} \frac{1}{n} \frac{g_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)}{f_0(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)} = \mathbb{E}_1\left[\log \frac{g_1(\boldsymbol{X}_1, \boldsymbol{X}_2)}{f_0(\boldsymbol{X}_1) f_0(\boldsymbol{X}_2)}\right], \tag{2.11}$$

where the underlying probability measure is $\mathbb{P}_1(\cdot)$. Such a limit is assumed to exist almost surely with $0 < I < \infty$, which is the case if the pre- and post-change data generating distributions are distinct for each node.

3. Universal lower bound on the WADD and CADD

In this section, we develop the universal lower bound on the CADD (and thus on the WADD) for any stopping rule τ that satisfies the false alarm constraint: $\mathbb{E}_\infty[\tau] \geq \gamma$.

Theorem 3.1. *Consider the statistical model defined in Section 2. If conditions C.1 and C.2 are satisfied, then as $\gamma \to \infty$,*

$$\inf_{\tau:\mathbb{E}_\infty[\tau]\geq\gamma} \mathrm{WADD}(\tau) \geq \inf_{\tau:\mathbb{E}_\infty[\tau]\geq\gamma} \mathrm{CADD}(\tau) \geq \frac{\log \gamma}{I}(1 + o(1)). \tag{3.1}$$

Proof. Let $\epsilon > 0$. Define $K_\gamma \triangleq \frac{\log \gamma}{I}$. By Markov's inequality, it follows that

$$\mathbb{E}_\nu[\tau - \nu | \tau \geq \nu] \geq \mathbb{P}_\nu(\tau - \nu \geq K_\gamma(1-\epsilon) | \tau \geq \nu) K_\gamma(1-\epsilon). \tag{3.2}$$

Then to prove the theorem, it suffices to show that for any τ satisfying $\mathbb{E}_\infty[\tau] \geq \gamma$, there exists some $\nu \geq 1$ such that $\mathbb{P}_\nu(\nu \leq \tau < \nu + K_\gamma(1-\epsilon) | \tau \geq \nu) = o(1)$, as $\gamma \to \infty$; that is, that

$$\lim_{\gamma\to\infty} \sup_{\tau:\mathbb{E}_\infty[\tau]\geq\gamma} \inf_{\nu\geq 1} \mathbb{P}_\nu(\nu \leq \tau < \nu + K_\gamma(1-\epsilon)|\tau \geq \nu) = 0. \tag{3.3}$$

Define $a \triangleq (1-\epsilon^2)\log\gamma$. Then for any ν, we have that

$$\begin{aligned} &\mathbb{P}_\nu(\nu \leq \tau < \nu + K_\gamma(1-\epsilon)|\tau \geq \nu) \\ &= \mathbb{P}_\nu\Big(\nu \leq \tau < \nu + K_\gamma(1-\epsilon), \Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_\tau) > e^a|\tau \geq \nu\Big) \\ &+ \mathbb{P}_\nu\Big(\nu \leq \tau < \nu + K_\gamma(1-\epsilon), \Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_\tau) \leq e^a|\tau \geq \nu\Big). \end{aligned} \tag{3.4}$$

The first term in (3.4) can be upper bounded as follows:

$$\begin{aligned} &\mathbb{P}_\nu\Big(\nu \leq \tau < \nu + K_\gamma(1-\epsilon), \Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_\tau) > e^a|\tau \geq \nu\Big) \\ &\overset{(a)}{\leq} \mathbb{P}_\nu\Big(\max_{\nu\leq j<K_\gamma(1-\epsilon)+\nu} \log\Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_j) > a|\tau \geq \nu\Big) \\ &\overset{(b)}{=} \mathbb{P}_\nu\Big(\max_{\nu\leq j<K_\gamma(1-\epsilon)+\nu} \log\Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_j) > a\Big) \\ &\overset{(c)}{=} \mathbb{P}_1\left(\frac{\max_{1\leq j<K_\gamma(1-\epsilon)+1} \log\Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)}{K_\gamma(1-\epsilon)} > I(1+\epsilon)\right), \end{aligned} \tag{3.5}$$

where (a) is due to the fact that

$$\{\nu \leq \tau < \nu + K_\gamma(1-\epsilon), \log\Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_\tau) > a\} \subseteq \Big\{\max_{\nu\leq j<K_\gamma(1-\epsilon)+\nu} \log\Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_j) > a\Big\}; \tag{3.6}$$

(b) is due to the facts that $\{\tau \geq \nu\} \in \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\nu-1})$ and the pre- and post-change observations are independent; and (c) follows by the definition of Λ_ν and the independence between the pre- and post-change observations.

By lemma A.1 in Fellouris and Tartakovsky (2017), if

$$\frac{\log\Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)}{k} \underset{\text{a.s.}}{\overset{k\to\infty}{\longrightarrow}} I \tag{3.7}$$

under $\mathbb{P}_1(\cdot)$, then it follows that

$$\lim_{\gamma\to\infty} \mathbb{P}_1\left(\frac{\max_{1\leq j<K_\gamma(1-\epsilon)+1} \log\Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)}{K_\gamma(1-\epsilon)} > I(1+\epsilon)\right) = 0. \tag{3.8}$$

By lemma 1.(i) in Fuh and Tartakovsky (2019), it follows that if C.1 and C.2 are satisfied, then (3.7) holds and consequently (3.8) holds.

We then analyze the second term in (3.4). By a change of measure argument similar to the one in Lai (1998), we have that there exists $\nu \geq 1$ such that

$$\begin{aligned} &\mathbb{P}_\nu\Big(\nu \leq \tau < \nu + K_\gamma(1-\epsilon), \Lambda_\nu(\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_\tau) \leq e^a|\tau \geq \nu\Big) \\ &\overset{(a)}{=} \mathbb{P}_1\Big(1 \leq \tau < 1 + K_\gamma(1-\epsilon), \Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau) \leq e^a\Big) \\ &\overset{(b)}{=} \mathbb{E}_\infty[\mathbb{1}_{\{1\leq\tau<1+K_\gamma(1-\epsilon), \Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau)\leq e^a\}}\Lambda_1(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\tau)] \\ &\leq e^a\mathbb{P}_\infty(\nu \leq \tau < \nu + K_\gamma(1-\epsilon)|\tau \geq \nu) \overset{(c)}{\leq} \frac{K_\gamma(1-\epsilon)e^a}{\gamma} \overset{\gamma\to\infty}{\longrightarrow} 0, \end{aligned} \tag{3.9}$$

where (a) follows similarly as steps (b) and (c) in (3.5); (b) follows by a change of measure argument; and (c) follows from the fact (see the proof of theorem 1 in Lai, 1998) that for any positive integer $m < \gamma$, if $\mathbb{E}_\infty[\tau] \geq \gamma$, then there exists some $\nu \geq 1$ such that

$$\mathbb{P}_\infty(\tau \geq \nu) > 0, \text{ and } \mathbb{P}_\infty(\tau < \nu + m | \tau \geq \nu) \leq m/\gamma. \tag{3.10}$$

Combining (3.8), (3.9), the fact that the upper bound in (3.9) is independent of τ, and the fact that for any stopping time τ, $\text{WADD}(\tau) \geq \text{CADD}(\tau)$, the theorem is established. □

The proof of the asymptotic universal lower bound is based on a change-of-measure argument similar to the one employed in Lai (1998) and the law of large numbers for log-likelihood ratio statistics of HMMs proposed in Fuh and Tartakovsky (2019). In contrast, the asymptotic lower bound analysis in Fuh (2003) follows Lorden's technique in Lorden (1971), which is based on interpreting the proposed test as a sequence of sequential probability ratio tests.

4. The windowed GLR Test

In this section, we first construct the windowed GLR test and then demonstrate its asymptotic optimality.

4.1. Algorithm Construction

The quickest moving anomaly detection problem in this article can be posed as a dynamic composite hypothesis testing problem, where at each time k we distinguish between the following two hypotheses:

$$H_1^k: \text{ the anomaly appears at time } \nu \leq k, \tag{4.1}$$

$$H_0^k: \text{ the anomaly appears at time } \nu > k. \tag{4.2}$$

Note that under the alternative hypothesis the change point ν is unknown. We then take a GLR approach to construct the detection statistic (see, e.g., Veeravalli and Banerjee, 2013 for the interpretation of classic QCD tests through the GLR approach). Specifically, the likelihood under these two hypotheses can be expressed respectively as follows:

$$H_1^k: \prod_{i=1}^{\nu-1} f_0(\boldsymbol{X}_i) \prod_{i=\nu}^{k} g_\nu(\boldsymbol{X}_i | \boldsymbol{X}_\nu, ..., \boldsymbol{X}_{i-1}), \tag{4.3}$$

$$H_0^k: \prod_{i=1}^{k} f_0(\boldsymbol{X}_i), \tag{4.4}$$

where $g_\nu(\boldsymbol{X}_i | \boldsymbol{X}_\nu, ..., \boldsymbol{X}_{i-1})$ denotes the post-change conditional distribution of $\boldsymbol{X}_i$ given the past observations (see (2.10)). Then, the GLR test statistic between the two hypotheses can be written as

$$W'_k = \max_{1 \le \nu \le k} \sum_{i=\nu}^{k} \log \frac{g_\nu(\boldsymbol{X}_i | \boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)}, \tag{4.5}$$

and the corresponding stopping rule is given by

$$\tau'_W = \inf\{k \ge 1 : W'_k > b\}. \tag{4.6}$$

Although the conditional pdf $g_\nu(\boldsymbol{X}_i | \boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_{i-1})$ in (4.5) can be calculated recursively (as shown below), to compute W'_k, the number of quantities that need to be stored scales with time k, which is not feasible for a real-time algorithm. Thus, to design an implementable GLR test, we consider a windowed version of W'_k. Denote the windowed version of the GLR statistic in (4.5) by

$$W_k = \max_{k-m \le j \le k} \sum_{i=j}^{k} \log \frac{g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)} \tag{4.7}$$

and the corresponding stopping time by

$$\tau_W = \inf\{k \ge 1 : W_k > b\}. \tag{4.8}$$

As will be observed later, the window length m needs to scale with the threshold b (and as a result γ), and also depends on the KL number I.

Note that for a fixed j, $g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1})$ can be calculated recursively. In particular, by using the Bayes rule, it can be easily shown that

$$g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}) = \sum_{\ell=1}^{L} f_\ell(\boldsymbol{X}_i) \mathbb{P}_j(S_i = \ell | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}), \tag{4.9}$$

$$\mathbb{P}_j(S_i = \ell | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}) = \sum_{\ell'=1}^{L} \mathbb{P}_j(S_{i-1} = \ell' | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}) \lambda_{\ell', \ell}, \tag{4.10}$$

$$\mathbb{P}_j(S_{i-1} = l | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}) = \frac{\mathbb{P}_j(S_{i-1} = l | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-2}) f_l(\boldsymbol{X}_{i-1})}{\sum_{\ell'=1}^{L} \mathbb{P}_j(S_{i-1} = \ell' | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-2}) f_{\ell'}(\boldsymbol{X}_{i-1})}, \tag{4.11}$$

where the recursion is initialized with the stationary probability of the DTMC:

$$\mathbb{P}_j(S_j = \ell | \boldsymbol{X}_{j-1}) \triangleq \alpha_\ell. \tag{4.12}$$

4.2. Asymptotic optimality

In this subsection, we establish the first-order asymptotic optimality of the windowed GLR test in (4.7) and (4.8) under both Lorden's and Pollak's criteria.

We start our analysis by presenting a lower bound on the MTFA.

Proposition 4.1. *For the stopping rule defined in (4.7) and (4.8), the MTFA can be lower bounded as follows:*

$$\mathbb{E}_\infty[\tau_W] \ge e^b. \tag{4.13}$$

Proof. Note that $W'_k \geq W_k$; thus, $\tau_W \geq \tau'_W$. Therefore,

$$\mathbb{E}_\infty[\tau_W] \geq \mathbb{E}_\infty\left[\tau'_W\right]. \tag{4.14}$$

Let

$$V_k \triangleq \sum_{j=1}^{k} \prod_{i=j}^{k} \frac{g_j(\boldsymbol{X}_j|\boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)} = \sum_{j=1}^{k} \Lambda_j(\boldsymbol{X}_j, \ldots, \boldsymbol{X}_k), \tag{4.15}$$

and

$$\tau_V = \inf\{k \geq 1 : V_k > e^b\}. \tag{4.16}$$

Note that

$$\begin{aligned} V_k &= \sum_{j=1}^{k} \frac{g_j(\boldsymbol{X}_k|\boldsymbol{X}_j, \ldots, \boldsymbol{X}_{k-1})}{f_0(\boldsymbol{X}_k)} \Lambda_j(\boldsymbol{X}_j, \ldots, \boldsymbol{X}_{k-1}) \\ &= \sum_{j=1}^{k-1} \frac{g_j(\boldsymbol{X}_k|\boldsymbol{X}_j, \ldots, \boldsymbol{X}_{k-1})}{f_0(\boldsymbol{X}_k)} \Lambda_j(\boldsymbol{X}_j, \ldots, \boldsymbol{X}_{k-1}) + \Lambda_j(\boldsymbol{X}_k). \end{aligned} \tag{4.17}$$

Then, from (4.15) and (4.17), we have that $\mathbb{E}_\infty[V_k|X_{k-1}, \ldots, X_1] = 1 + V_{k-1}$, and $\mathbb{E}_\infty[V_k] = k$. This implies that $\{V_k - k\}_{k=1}^\infty$ is a zero-mean martingale under $\mathbb{P}_\infty$. Thus, by the optional sampling theorem (see, e.g., Poor and Hadjiliadis, 2009) and the fact that $V_k \geq e^{W'_k}$, we have that

$$\mathbb{E}_\infty[\tau_W] \geq \mathbb{E}_\infty\left[\tau'_W\right] \geq \mathbb{E}_\infty[\tau_V] = \mathbb{E}_\infty[V_{\tau_V}] \geq e^b. \tag{4.18}$$

Next, we establish an asymptotic upper bound on the WADD and CADD of the windowed GLR test in (4.7) and (4.8).

Theorem 4.1. *Consider the stopping rule defined in (4.7) and (4.8). Consider the window length $m = m(b)$ such that*

$$\liminf_{b\to\infty} \frac{m(b)}{b} > \frac{1}{I}. \tag{4.19}$$

Then, under conditions C.1 and C.2, we have that as $b \to \infty$,

$$\text{CADD}(\tau_W) \leq \text{WADD}(\tau_W) \leq \frac{b}{I}(1 + o(1)). \tag{4.20}$$

Proof. Let $\epsilon > 0, \delta > 0$ and $n_b \triangleq \frac{b(1+\epsilon)}{I}$. It can be shown that for any $\nu \geq 1$,

$$\begin{aligned} &\text{ess sup}\mathbb{E}_\nu\left[\frac{(\tau_W - \nu + 1)^+}{n_b}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\nu-1}\right] \leq \text{ess sup}\sum_{l=0}^{\infty} \mathbb{P}_\nu(\tau_W - \nu + 1 > ln_b|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\nu-1}) \\ &\leq \sum_{l=0}^{\infty} \text{ess sup}\mathbb{P}_\nu(\tau_W > ln_b + \nu - 1|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\nu-1}) \\ &\leq 1 + \sum_{l=1}^{\infty} \text{ess sup}\mathbb{P}_\nu(\tau_W > ln_b + \nu - 1|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\nu-1}). \end{aligned} \tag{4.21}$$

For any $l \geq 1$, it follows that

$$\begin{aligned}
&\mathbb{P}_\nu(\tau_W > ln_b + \nu - 1 | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1}) \\
&= \mathbb{P}_\nu\left(\max_{1 \leq k \leq ln_b + \nu - 1} W_k < b | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right) \\
&= \mathbb{P}_\nu\left(\max_{1 \leq k \leq ln_b + \nu - 1} \max_{k-m \leq j \leq k} \sum_{i=j}^{k} \log \frac{g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)} < b | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right) \\
&\leq \mathbb{P}_\nu\left(\bigcap_{\xi \in \{1, ..., l\}} \left\{ \max_{\xi n_b + \nu - 1 - m \leq j \leq \xi n_b + \nu - 1} \sum_{i=j}^{\xi n_b + \nu - 1} \log \frac{g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)} < b \right\} | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right).
\end{aligned} \tag{4.22}$$

Without loss of generality, we choose m such that $m \geq n_b$ for large b. This further implies that $\xi n_b + \nu - m \leq (\xi - 1)n_b + \nu$ for large b. As a result, for large b, (4.22) can be further upper bounded as follows:

$$\mathbb{P}_\nu(\tau_W > ln_b + \nu - 1 | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1}) \leq \mathbb{P}_\nu\left(\bigcap_{\xi \in \{1, ..., l\}} A_\xi | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right), \tag{4.23}$$

where for simplicity of notation, we denote the event by

$$A_\xi \triangleq \left\{ \sum_{i=(\xi-1)n_b+\nu}^{\xi n_b + \nu - 1} \log \frac{g_{(\xi-1)n_b+\nu}(\boldsymbol{X}_i | \boldsymbol{X}_{(\xi-1)n_b+\nu}, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)} < b \right\}, \tag{4.24}$$

for all $\xi \geq 1$. It is clear that $A_\xi \in \sigma(\boldsymbol{X}_{(\xi-1)n_b+\nu}, ..., \boldsymbol{X}_{\xi n_b + \nu - 1})$. Then, it follows that

$$\mathbb{P}_\nu\left(\bigcap_{\xi \in \{1, ..., l\}} A_\xi \middle| \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right) = \prod_{\xi=1}^{l} \mathbb{P}_\nu\big(A_\xi | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1}, A_1, ..., A_{\xi-1}\big). \tag{4.25}$$

This further implies that

$$\text{ess sup}\mathbb{P}_\nu\left(\bigcap_{\xi \in \{1, ..., l\}} \mathrm{A}_\xi \middle| \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right) \leq \prod_{\xi=1}^{1} \text{ess sup}\mathbb{P}_\nu\left(\mathrm{A}_\xi | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1}, \mathrm{A}_1, ..., \mathrm{A}_{\xi-1} \right). \tag{4.26}$$

If the following holds that for any $\xi \geq 1$,

$$\text{ess sup}\mathbb{P}_\nu(\mathrm{A}_\xi | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1}, \mathrm{A}_1, ..., \mathrm{A}_{\xi-1}) \leq \delta, \tag{4.27}$$

where δ is independent of ν, and can be arbitrarily small for large b, then

$$\text{ess sup}\mathbb{P}_\nu\left(\bigcap_{\xi \in \{1, ..., l\}} \mathrm{A}_\xi | \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right) \leq \delta^l \tag{4.28}$$

and, further,

$$\sup_\nu \text{ess sup}\mathbb{E}_\nu\left[\frac{(\tau_W - \nu + 1)^+}{n_b} \middle| \boldsymbol{X}_1, ..., \boldsymbol{X}_{\nu-1} \right] \leq 1 + \sum_{\ell=1}^{\infty} \delta^l = \frac{1}{1 - \delta}. \tag{4.29}$$

This implies that

$$\sup_{\nu} \operatorname{ess\,sup} \mathbb{E}_{\nu}\left[\frac{(\tau_W-\nu+1)^+}{n_b}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1}\right] \leq \frac{b(1+\epsilon')}{I}, \tag{4.30}$$

where $\epsilon' = (1+\epsilon)/(1-\delta)$. Because ϵ is arbitrary and δ can be arbitrarily small for large b, the proof is complete if we can show that (4.27) is true.

In the following, we prove that (4.27) is true. We first note that by our notation,

$$\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1}) = \sum_{i=(\xi-1)n+\nu}^{\xi n+\nu-1} \log \frac{g_{(\xi-1)n+\nu}(\boldsymbol{X}_i|\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)}. \tag{4.31}$$

By the Markov property of the problem model as in (2.3) and (2.4), it follows that

$$\begin{aligned}
&\mathbb{P}_{\nu}\left(\frac{1}{n}\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1}) < \frac{I}{1+\epsilon}\middle|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1},A_1,\ldots,A_{\xi-1}\right)\\
&= \sum_{s\in\mathcal{S}} \mathbb{P}_{\nu}\Bigg(\frac{1}{n}\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1}) < \frac{I}{1+\epsilon},\\
&\qquad S_{(\xi-1)n+\nu} = s|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1},A_1,\ldots,A_{\xi-1}\Bigg)\\
&= \sum_{s\in\mathcal{S}} \mathbb{P}_{\nu}\left(\frac{1}{n}\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1}) < \frac{I}{1+\epsilon}\middle|S_{(\xi-1)n+\nu} = s\right)\\
&\qquad \times \mathbb{P}_{\nu}(S_{(\xi-1)n+\nu} = s|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1},A_1,\ldots,A_{\xi-1}).
\end{aligned} \tag{4.32}$$

From lemma A.1. in Fuh and Tartakovsky (2019), it follows that for any $s\in\mathcal{S}$ and any $\xi\geq 1$,

$$\lim_{n\to\infty} \mathbb{P}_{\nu}\left(\frac{1}{n}\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1}) < \frac{I}{1+\epsilon}\middle|S_{(\xi-1)n+\nu} = s\right) = 0. \tag{4.33}$$

It then follows that for any $\xi \geq 1$,

$$\begin{aligned}
&\lim_{n\to\infty} \operatorname{ess\,sup} \mathbb{P}_{\nu}\Bigg(\sum_{i=(\xi-1)n+\nu}^{\xi n+\nu-1} \frac{1}{n}\log \frac{g_{(\xi-1)n+\nu}(\boldsymbol{X}_i|\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)}\\
&\quad < \frac{I}{1+\epsilon}\Bigg|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1},A_1,\ldots,A_{\xi-1}\Bigg)\\
&= \lim_{n\to\infty} \operatorname{ess\,sup} \mathbb{P}_{\nu}\Bigg(\frac{1}{n}\log \Lambda_{(\xi-1)n+\nu}(\boldsymbol{X}_{(\xi-1)n+\nu},\ldots,\boldsymbol{X}_{\xi n+\nu-1})\\
&\quad < \frac{I}{1+\epsilon}\Bigg|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{\nu-1},A_1,\ldots,A_{\xi-1}\Bigg) = 0,
\end{aligned} \tag{4.34}$$

which further implies that (4.27) is true. This concludes the proof. □

The following theorem demonstrates the asymptotic optimality of the windowed GLR test, which follows directly from Proposition 4.1 and Theorems 3.1 and 4.1.

Theorem 4.2. *Consider the stopping rule defined in (4.7) and (4.8) with $b = \log\gamma$ and m chosen to satisfy*

$$\liminf_{b\to\infty}\frac{m(b)}{b} > \frac{1}{I}. \tag{4.35}$$

Then under conditions C.1 and C.2, the windowed GLR test is asymptotically optimal under both Lorden's and Pollak's criteria; that is, as $\gamma \to \infty$,

$$\mathrm{WADD}(\tau_W) \sim \mathrm{CADD}(\tau_W) \sim \frac{\log\gamma}{I}. \tag{4.36}$$

Proof. The result follows directly from Proposition 4.1 and Theorems 3.1 and 4.1. □

5. Alternative detection schemes

In this section, we develop several alternative algorithms for the problem of moving anomaly detection in networks and derive lower bounds on their MTFAs. We first design a dynamic Shiryaev-Roberts (D-S-R) algorithm by modeling the change point as a geometric random variable with parameter ρ and then letting $\rho \to 0$. The advantage of the D-S-R algorithm is that it can be updated recursively. We then develop a QCD algorithm with recursive change point estimation. This test recursively estimates the unknown change point and then constructs a CUSUM-type algorithm using the estimated change point. Finally, we design a mixture CUSUM algorithm, which is applicable for the case where the Markov transition probabilities are unknown.

5.1. Dynamic Shiryaev-Roberts algorithm

We first assume that the change point is a geometric random variable with parameter ρ. We denote the change point by Γ. Specifically,

$$\mathbb{P}(\Gamma = m) = \rho(1-\rho)^{m-1}, \; m \in \mathbb{N}. \tag{5.1}$$

In the following, we will show how we design a recursive test under such a Bayesian framework. We will further let $\rho \to 0$ so that the designed algorithm does not depend on ρ and can be applied to the minimax setting described in Section 2, where the change point is deterministic and unknown.

Under the Bayesian assumption of the change point, we introduce one additional state 0 to denote the state where there is no anomaly in the network. Then the transition from the pre-change mode to the post-change mode can be represented by the transition from state 0 to any state $\ell \in \mathcal{L}$. Specifically, for all $\ell \in \{1, \ldots, L\}$, we denote by $\lambda_{0,\ell}$ the probability that the anomaly first emerges at node ℓ; that is,

$$\mathbb{P}(S_k = \ell | S_{k-1} = 0) = \lambda_{0,\ell}. \tag{5.2}$$

It is clear that $\rho = \sum_{\ell=1}^{L}\lambda_{0,\ell}$. We further note that $\lambda_{\ell,0} = 0$, for any $\ell \in \mathcal{L}$, and $\ell_{0,0} = 1-\rho$. For any $\ell \in \{0, 1, \ldots, L\}$, and $k \geq 1$, define by

$$p_{\ell,k} \triangleq \mathbb{P}(S_k = \ell | \mathbf{X}_1, \ldots, \mathbf{X}_k) \tag{5.3}$$

the posterior probability that the network is at state ℓ at time k. A natural way to construct a test is to compare with a threshold the posterior probability that the network is in the pre-change state.

In particular, $p_{\ell,k}$ can be updated recursively. For any $\ell \in \{0\} \cup \mathcal{L}$, by the Bayes rule we have that

$$\begin{aligned} p_{\ell,k} &= \frac{\mathbb{P}(S_k = \ell | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}, \boldsymbol{X}_k) f(\boldsymbol{X}_k | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})}{f(\boldsymbol{X}_k | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})} \\ &= \frac{\mathbb{P}(S_k = \ell | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k | S_k = \ell, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})}{\sum_{i=0}^{L} f(\boldsymbol{X}_k, S_k = i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})} \\ &= \frac{\mathbb{P}(S_k = \ell | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k | S_k = \ell, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})}{\sum_{i=0}^{L} \mathbb{P}(S_k = i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k | S_k = i, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})} \\ &= \frac{A_{\ell,k}}{\sum_{i=0}^{L} A_{i,k}}, \end{aligned} \tag{5.4}$$

where $f(\cdot|\cdot)$ denotes the conditional probability density function of $\boldsymbol{X}_k$ and

$$\begin{aligned} A_{i,k} &\triangleq \mathbb{P}(S_k = i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k | S_k = i, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) \\ &= \mathbb{P}(S_k = i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f_i(\boldsymbol{X}_k). \end{aligned} \tag{5.5}$$

We then compute $A_{i,k}$ as follows:

$$\begin{aligned} A_{i,k} &= \mathbb{P}(S_k = i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k | S_k = i) \\ &= \sum_{j=0}^{L} \mathbb{P}(S_k = i, S_{k-1} = j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f_i(\boldsymbol{X}_k) \\ &= \sum_{j=0}^{L} \mathbb{P}(S_{k-1} = j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) \mathbb{P}(S_k = i | S_{k-1} = j, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f_i(\boldsymbol{X}_k) \\ &= \sum_{j=0}^{L} \mathbb{P}(S_{k-1} = j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) \mathbb{P}(S_k = i | S_{k-1} = j) f_i(\boldsymbol{X}_k) \\ &= \left[\sum_{j=0}^{L} p_{j,k-1} \lambda_{j,i} \right] f_i(\boldsymbol{X}_k). \end{aligned} \tag{5.6}$$

Combining (5.4) and (5.6) implies that $p_{\ell,k}$ can be updated recursively.

We note that for $1 \leq i \leq L$,

$$\begin{aligned} A_{i,k} &= \sum_{j=0}^{L} \mathbb{P}(S_{k-1} = j | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) \mathbb{P}(S_k = i | S_{k-1} = j) f_i(\boldsymbol{X}_k) \\ &= \left[\sum_{j=0}^{L} p_{j,k-1} \lambda_{j,i} \right] f_i(\boldsymbol{X}_k) = \left[p_{0,k-1} \lambda_{0,i} + \sum_{j=1}^{L} p_{j,k-1} \lambda_{j,i} \right] f_i(\boldsymbol{X}_k). \end{aligned} \tag{5.7}$$

Furthermore, for $i = 0$ we have that

$$A_{0,k} = \mathbb{P}(S_k = 0|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}) f(\boldsymbol{X}_k|S_k = 0) = p_{0,k-1}\lambda_{0,0} f_0(\boldsymbol{X}_k). \tag{5.8}$$

The recursion is initialized with $p_{0,0} = 1$ and $p_{\ell,0} = 1$ for $\ell \in \mathcal{L}$.

We further define the following invertible mapping:

$$q_{\ell,k} = \frac{p_{\ell,k}}{\rho p_{0,k}} \Longleftrightarrow p_{\ell,k} = \frac{q_{\ell,k}}{\sum_{j=0}^{L} q_{j,k}}. \tag{5.9}$$

It then follows that

$$p_{0,k} = \frac{1}{1 + \left[\rho \sum_{j=1}^{L} q_{j,k}\right]}, \tag{5.10}$$

where $q_{\ell,k}$ can be computed recursively by

$$q_{\ell,k} = \frac{\left[\frac{\lambda_{0,\ell}}{\rho} + \sum_{j=1}^{L} q_{j,k-1}\lambda_{j,\ell}\right] f_\ell(\boldsymbol{X}_k)}{\lambda_{0,0} f_0(\boldsymbol{X}_k)}, \tag{5.11}$$

with the following priors: $q_{0,k} = 1/\rho$ and $q_{\ell,0} = 0$, $\ell \in \{1, \ldots, L\}$. From (5.10), it follows that comparing $p_{0,k}$ to a threshold B is equivalent to comparing $\sum_{j=1}^{L} q_{j,k}$ to a threshold $(1/B - 1)/\rho$.

To obtain a test that does not depend on ρ and can be applied to the non-Bayesian setting, we take the limit $\rho \to 0$. In particular, we assume that as $\rho \to 0$,

$$\frac{\lambda_{0,\ell}}{\rho} \to \alpha_\ell \tag{5.12}$$

for all $\ell \in \mathcal{L}$. Practically, this means that the change point is treated as an unknown but deterministic variable and that after the change occurs, the initial location of the anomaly is distributed according to $\boldsymbol{\alpha}$. As a result, if we define

$$r_{\ell,k} = \lim_{\rho \to 0} q_{\ell,k},$$

for $\ell \in \{1, \ldots, L\}$, then the recursion of $r_{\ell,k}$ is

$$r_{\ell,k} = \left[\alpha_\ell + \sum_{j=1}^{L} r_{j,k-1}\lambda_{j,\ell}\right] \frac{f_{\ell,1}(X_{\ell,k})}{f_{\ell,0}(X_{\ell,k})} \tag{5.13}$$

with $r_{\ell,0} \triangleq 0$. Define the test statistic

$$R_k = \sum_{\ell=1}^{L} r_{\ell,k}. \tag{5.14}$$

The corresponding stopping rule is then given by

$$\tau_R = \inf\{k \geq 1 : \log R_k \geq b\}. \tag{5.15}$$

This test involves calculating a test statistic for each node in the network. At each time k, the test statistic for one node is calculated by first weighing the test statistics of all of the nodes at the previous time instant according to the corresponding transition probabilities and then multiplying the likelihood ratio of the sample taken by that node. Thus, knowledge of the transition probabilities is needed in order to implement this test.

We note that the D-S-R algorithm is developed by letting $\rho \rightarrow 0$. Such a change point can be intuitively interpreted as a "uniformly" distributed random variable on the entire timescale. Therefore, this algorithm may not perform as well as the windowed GLR test under both Lorden's and Pollak's criteria, because both criteria are defined for the worst-case scenario over all possible change points. An experimental study will be provided in Section 7.

Next, we derive a lower bound on the MTFA for the D-S-R algorithm.

Lemma 5.1. *For the stopping rule defined in (5.14) and (5.15), its MTFA is lower bounded as follows:*

$$\mathbb{E}_\infty[\tau_R] \geq e^b. \tag{5.16}$$

Proof. Note that

$$\begin{aligned}\mathbb{E}_\infty[R_k|\boldsymbol{X}_{k-1},\dots,\boldsymbol{X}_1] &= \mathbb{E}_\infty\left[\sum_{j=1}^{L} r_{j,k}\middle|\boldsymbol{X}_1,\dots,\boldsymbol{X}_{k-1}\right]\\ &= \mathbb{E}_\infty\left[\sum_{j=1}^{L}\left(\left\{\alpha_\ell + \sum_{q=1}^{L} r_{q,k-1}\lambda_{j,q}\right\}\frac{f_{j,1}(X_{j,k})}{f_{j,0}(X_{j,k})}\right)|\boldsymbol{X}_1,\dots,\boldsymbol{X}_{k-1}\right]\\ &= 1 + \sum_{j=1}^{L}\sum_{q=1}^{L}\lambda_{j,q} r_{j,k-1}\\ &= 1 + \sum_{j=1}^{L} r_{j,k-1}\\ &= 1 + R_{k-1},\end{aligned} \tag{5.17}$$

which implies that $\{R_k - k\}_{k=1}^{\infty}$ is a martingale under $\mathbb{P}_\infty(\cdot)$. It can also be shown that $\mathbb{E}_\infty[R_k - k] = 0$. As a result, by the optimal stopping theorem (see, e.g., Poor and Hadjiliadis, 2009), it follows that $\mathbb{E}_\infty[W_{\tau_R} - \tau_R] = 0$. This further implies that

$$\mathbb{E}_\infty[\tau_R] = \mathbb{E}_\infty[R_{\tau_R}] \geq e^b. \tag{5.18}$$

□

5.2. QCD Algorithm with Recursive Change Point Estimation

In the windowed GLR test, the change point is implicitly estimated by the maximum likelihood approach over a finite window. The estimation does not have a recursive form and thus is not as computationally efficient, which is why a windowed approach is used. An interesting question is whether we can design a test that can recursively and inherently estimate the change point and then construct a CUSUM-type algorithm using the estimated change point.

QCD algorithms based on recursive change point estimation were proposed in Lau et al. (2019) to solve the semiparametric QCD problem and in Lorden and Pollak (2008) to solve the composite QCD problem (for prior work in composite QCD, see Lai, 1998 and Lorden, 1971). The main idea is motivated by the CUSUM algorithm, for which, before the changepoint, the test statistic takes values around zero and therefore an estimate of the change point is the last time that the test statistic became zero. Following a similar idea, we design a QCD algorithm with recursive change point estimation. In particular, define the following test statistic:

$$U_k = \max_{\zeta_{k-1} \le j \le k+1} \sum_{i=j}^{k} \log \frac{g_{\zeta_{k-1}}(\boldsymbol{X}_i | \boldsymbol{X}_{\zeta_{k-1}}, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)}, \tag{5.19}$$

where ζ_k denotes the estimate of the change point at time k and $g_j(\boldsymbol{X}_i | \boldsymbol{X}_j, \ldots, \boldsymbol{X}_{i-1}) \triangleq 0$ for $j > i$. The estimate of the change point is defined by

$$\zeta_k = \operatorname*{argmax}_{\zeta_{k-1} \le j \le k+1} \sum_{i=j}^{k} \log \frac{g_{\zeta_{k-1}}(\boldsymbol{X}_i | \boldsymbol{X}_{\zeta_{k-1}}, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)}. \tag{5.20}$$

Following steps similar to those in Lau et al. (2019), it can be shown that the detection statistics in (5.19) and (5.20) can be updated recursively as follows:

$$U_{k+1} = \left(U_k + \log \frac{g_{\zeta_k}(\boldsymbol{X}_{k+1} | \boldsymbol{X}_{\zeta_k}, \ldots, \boldsymbol{X}_k)}{f_0(\boldsymbol{X}_{k+1})} \right)^+, \tag{5.21}$$

and

$$\zeta_{k+1} = \begin{cases} \zeta_k, & U_k > 0 \text{ or } \zeta_k = k+1, \\ k+2, & \text{else}, \end{cases} \tag{5.22}$$

where $U_0 \triangleq 0$ and $\zeta_0 \triangleq 1$. The corresponding stopping rule is

$$\tau_U = \inf\{k \ge 1 : U_k \ge b\}. \tag{5.23}$$

The advantage of such a test is that it is an approximation to the GLR test, which can be implemented recursively. We now present a lower bound for the MTFA for the algorithm defined in (5.19)–(5.23).

Lemma 5.2. *For the stopping rule defined in (5.19)–(5.23), the MTFA can be lower bounded as follows:*

$$\mathbb{E}_\infty[\tau_U] \ge e^b. \tag{5.24}$$

Proof. Let

$$\hat{\tau} = \inf\left\{ k \ge 1 : \hat{S}_k \triangleq \sum_{i=1}^{k} \log \frac{g_1(\boldsymbol{X}_i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)} \ge b \right\}. \tag{5.25}$$

By expressing the algorithm in (5.19)–(5.23) as a sequence of i.i.d. circles of (5.25), it can be easily shown that by using Wald's identity (see, e.g., Veeravalli and Banerjee, 2013)

$$\mathbb{E}_\infty[\tau_U] \geq \frac{\mathbb{E}_\infty[\hat{\tau}]}{\mathbb{P}_\infty(\hat{S}_{\hat{\tau}} \geq b)} \geq \frac{1}{\mathbb{P}_\infty(\hat{S}_{\hat{\tau}} \geq b)}. \tag{5.26}$$

Consider the event

$$E_t = \left\{ \prod_{i=1}^{t} \frac{g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)} \geq e^b, \hat{\tau} = t \right\}.$$

It then follows that

$$\begin{aligned}
&\mathbb{P}_\infty(\hat{S}_{\hat{\tau}} \geq b) \\
&= \sum_{t=1}^{\infty} \mathbb{P}_\infty(E_t) = \sum_{t=1}^{\infty} \mathbb{E}_\infty[\mathbb{1}_{E_t}] \\
&= \sum_{t=1}^{\infty} \mathbb{E}_\infty\left[\prod_{i=1}^{t} \frac{g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)} \prod_{i=1}^{t} \frac{f_0(\boldsymbol{X}_j)}{g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1})} \mathbb{1}_{E_t} \right] \\
&\leq e^{-b} \sum_{t=1}^{\infty} \mathbb{E}_\infty\left[\prod_{i=1}^{t} \frac{g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_j)} \mathbb{1}_{E_t} \right] \\
&\leq e^{-b} \sum_{t=1}^{\infty} \mathbb{P}_1(E_t) \\
&\leq e^{-b}.
\end{aligned} \tag{5.27}$$

The result then follows by combining (5.26) and (5.27). □

Due to the use of the recursive change point estimate ζ_k, the analysis of the detection delay for this algorithm is challenging, and we leave this as an open problem for future research.

5.3. Mixture CUSUM algorithm

In practice, it might be hard to acquire complete knowledge of the transition probabilities of the DTMC in (2.3). However, it might be possible to have a good estimate of the stationary distribution of the DTMC; for example, based on symmetries in the network, we may be able to approximate the stationary distribution by a uniform distribution. In this case, we approximate the postchange data generating distribution by a mixture of $f_{\ell,1}$, where the weights are the stationary distribution $\boldsymbol{\alpha}$, and construct a CUSUM algorithm that tests the change from the pre-change distribution to the mixture distribution.

In particular, the mixture CUSUM test statistic is defined as follows:

$$C_k = \max_{1 \leq j \leq k+1} \sum_{i=j}^{k} \log\left(\sum_{\ell=1}^{L} \alpha_\ell \frac{f_{\ell,1}(X_{\ell,i})}{f_{\ell,0}(X_{\ell,i})} \right). \tag{5.28}$$

Note that this statistic can be updated recursively:

$$C_{k+1} = \left(C_k + \log \left(\sum_{\ell=1}^{L} \alpha_\ell \frac{f_{\ell,1}(X_{\ell,k+1})}{f_{\ell,0}(X_{\ell,k+1})} \right) \right)^+ \tag{5.29}$$

with $C_0 \triangleq 0$. The mixture CUSUM stopping rule is

$$\tau_C = \inf\{k \geq 1 : C_k \geq b\}. \tag{5.30}$$

Because this test is essentially a CUSUM algorithm that tests a change from the pre-change distribution to a mixture post-change distribution, its MTFA can be lower bounded similar to the CUSUM algorithm.

Lemma 5.3. *For the mixture CUSUM algorithm defined in (5.28)–(5.30), the MTFA can be lower bounded as follows:*

$$\mathbb{E}_\infty[\tau_C] \geq e^b. \tag{5.31}$$

Proof. The result follows directly from the lower bound on the MTFA for the CUSUM algorithm (see Lai, 1998; Lorden, 1971; Pollak, 1985). □

Because the mixture CUSUM algorithm only employs the stationary distribution of the DTMC, we might expect a loss in performance compared to the other algorithms that make use of the entire transition matrix. However, as will be seen in Section 7, the mixture CUSUM performs competitively with the asymptotically optimal algorithms.

6. Fuh's recursive approximation test

In this section, we review Fuh's recursive approximation algorithm, and instantiate it for our moving anomaly detection problem.

As discussed in Section 4, the GLR-based test does not admit a recursion. To address this problem, Fuh (2003) approximates the conditional p.d.f. $g_\nu(\boldsymbol{X}_i|\boldsymbol{X}_\nu, \ldots, \boldsymbol{X}_{i-1})$ in (4.5) using $g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1})$. Such an approximation inherently uses the likelihood when the change point is at time 1 to approximate the likelihood when the change point is at ν. In this way, the log-likelihood ratio does not depend on the change point ν and thus the test statistic can be updated recursively. Specifically, the detection statistic of Fuh's recursive approximation test is

$$F_k = \max_{1 \leq j \leq k+1} \sum_{i=j}^{k} \log \frac{g_1(\boldsymbol{X}_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1})}{f_0(\boldsymbol{X}_i)}. \tag{6.1}$$

Then, F_k can be written recursively as follows:

$$F_{k+1} = \left(F_k + \log \frac{g_1(\boldsymbol{X}_{k+1}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)}{f_0(\boldsymbol{X}_{k+1})} \right)^+, \tag{6.2}$$

where $F_0 \triangleq 0$. The corresponding stopping rule is defined as

$$\tau_F = \inf\{k \geq 1 : F_k > b\}. \tag{6.3}$$

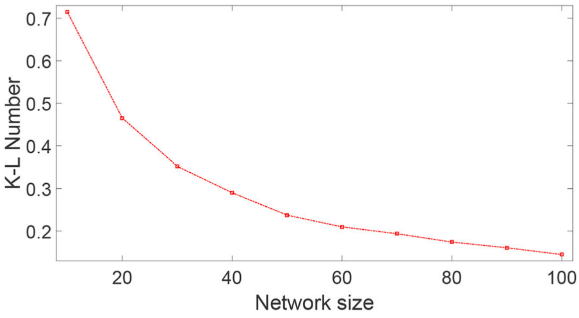

Figure 2. *I* versus *L*.

In Fuh (2003), Fuh used the stationarity properties of Markov chains to prove the first-order asymptotic optimality of τ_F. For completeness, we include his result in the next theorem.

Theorem 6.1. (Fuh, 2003) *Consider the stopping rule defined in (6.1)–(6.3) with* $b = \log\gamma$. *Then we have that* $\mathbb{E}_\infty[\tau_F] \geq \gamma$ *and that*

$$\mathrm{WADD}(\tau_F) \sim \mathrm{CADD}(\tau_F) \sim \frac{\log\gamma}{I}, \tag{6.4}$$

as $\gamma \to \infty$.

7. Numerical results

In this section, we conduct a numerical study for the moving anomaly detection problem. We set $f_{\ell,0} = \mathcal{N}(0,1)$ and $f_{\ell,0} = \mathcal{N}(2,1)$ for all $\ell \in \mathcal{L}$. We consider different values of network size L and compare all of the algorithms discussed in this article.

For the windowed GLR test, the QCD algorithm with recursive change point estimation, and Fuh's recursive approximation test, the worst-case detection delay is not necessarily attained at $\nu = 1$ for the WADD or CADD (also see Fuh and Mei, 2015). As a result, it is difficult to analytically or numerically calculate the worst-case detection delay for these algorithms. For the D-S-R and mixture CUSUM tests, the WADD and CADD are attained at $\nu = 1$. For the purpose of illustration, we simulate the average detection delay $\mathbb{E}_\nu[\tau - \nu | \tau \geq \nu]$ for different values of the change point ν, which serves as an approximation for the WADD and CADD.

In Figure. 2, we evaluate the value of I as a function of the network size L. The KL number I was calculated by the Monte Carlo method according to (2.11). Note that I decreases with network size. This implies that for a large network, the windowed GLR test requires a large window size. In Figure. 3 we plot the evolution of statistics for $\nu = 100$. It can be seen that the statistics for all of the algorithms grow after the change point. In Figure. 4 we plot the average detection delay vs. MTFA for the algorithms discussed in this article for $\nu = 1$, $L = 10$, and $m = 30$. Among all of the tests, the windowed GLR test, Fuh's recursive approximation algorithm, and the mixture CUSUM test perform the best. In the remainder of this section, we mainly compare these three algorithms.

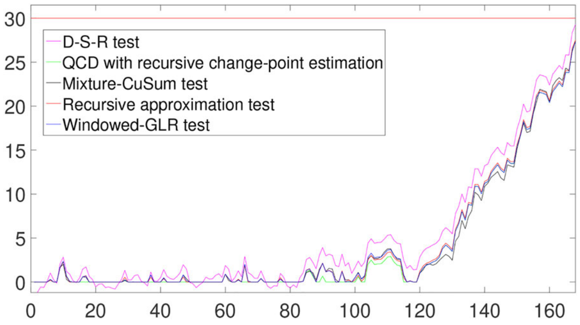

Figure 3. Evolution of the test statistics for $L = 100$ and $\nu = 120$.

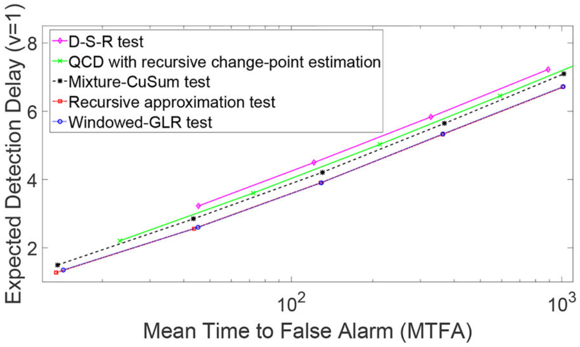

Figure 4. $\mathbb{E}_1[\tau - 1|\tau \geq 1]$ versus MTFA for $L = 10$.

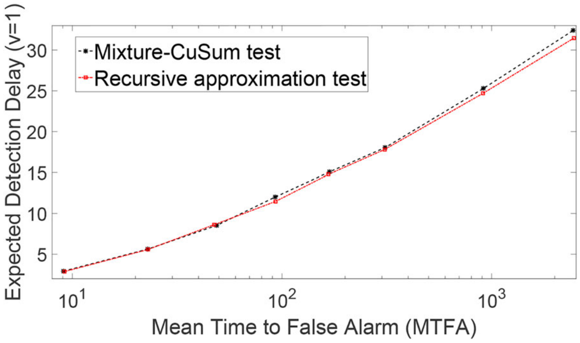

Figure 5. $\mathbb{E}_1[\tau - 1|\tau \geq 1]$ versus MTFA for $L = 100$.

In Figure. 5 we first compare Fuh's test with the mixture CUSUM test for $L = 100$ and $\nu = 1$. We note that although the mixture CUSUM algorithm only employs the stationary distribution of the DTMC and does not use the transition probabilities, it provides very good performance compared to Fuh's recursive approximation test, which is provably first-order asymptotically optimal. Furthermore, Fuh's test can be computationally expensive for a large L, because it requires $O(L^2)$ computations per time step, whereas the computational complexity for the mixture CUSUM algorithm is only $O(L)$. Thus, for large networks, the mixture CUSUM test might be a better choice if the computational resource is limited. In Figure. 6, we repeat the comparison for $L = 10$,

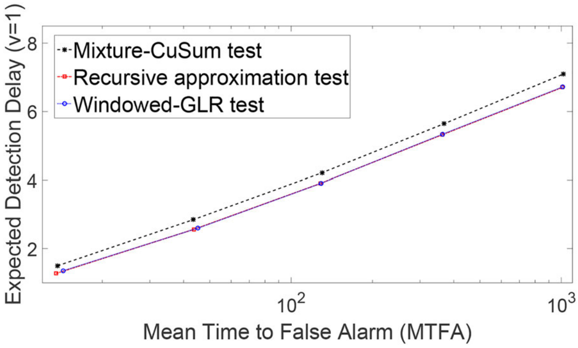

Figure 6. $\mathbb{E}_1[\tau - 1|\tau \geq 1]$ versus MTFA for $L = 10$.

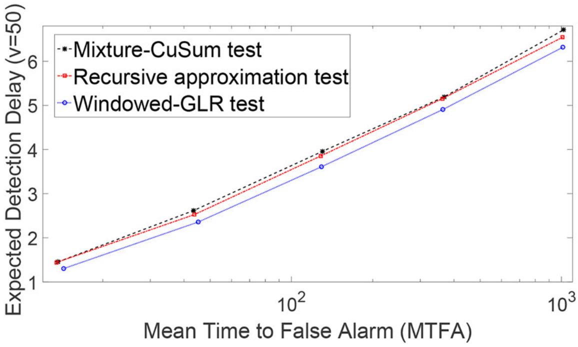

Figure 7. $\mathbb{E}_{50}[\tau - 50|\tau \geq 50]$ versus MTFA for $L = 10$.

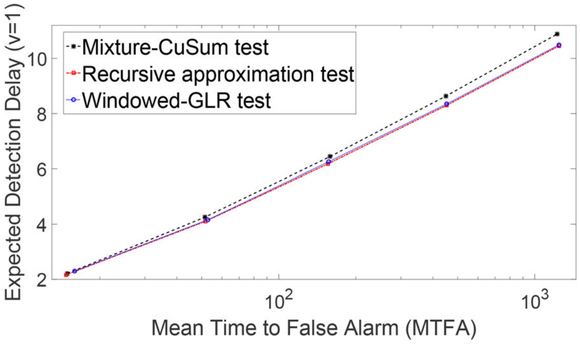

Figure 8. $\mathbb{E}_1[\tau - 1|\tau \geq 1]$ versus MTFA for $L = 20$.

$m = 30$, and $\nu = 1$ by adding the windowed GLR test, and similar observations are obtained. Note that in this case Fuh's recursive test offers performance identical to that of the windowed GLR, because the former inherently assumes that the change occurs at $\nu = 1$.

In Figure. 7, we further compare Fuh's test and the mixture CUSUM test with the windowed GLR test for $L = 10$, $m = 30$, and $\nu = 50$. Note that although for the case of $\nu = 1$ the windowed GLR test has performance similar to that of Fuh's algorithm, the windowed GLR test performs better for $\nu \neq 1$. This phenomenon is expected because

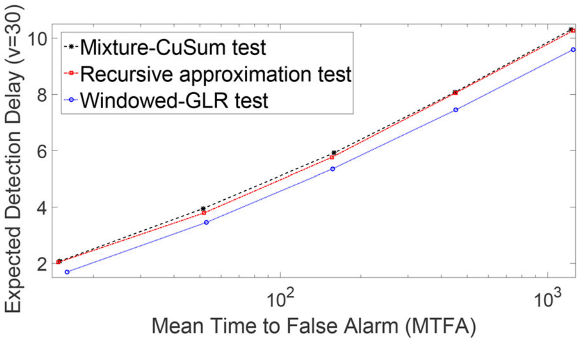

Figure 9. $\mathbb{E}_{30}[\tau - 30|\tau \geq 30]$ versus MTFA for $L = 20$.

Fuh's test uses the likelihood when $\nu = 1$ as an approximation. Finally, in Figures. 8 and 9 we compare the three tests for the case of $L = 20$ with $m = 50$, $\nu = 1$, and $\nu = 30$.

8. Conclusions

In this article, we studied the problem of moving anomaly detection in networks. The trajectory of the moving anomaly after it emerges in the network is modeled as a DTMC, which results in the observation model being an HMM. We constructed the windowed GLR test for the detection problem and established its first-order asymptotic optimality. We also constructed three alternative tests, including the D-S-R test, the QCD test with recursive change point estimation, and the mixture CUSUM test. For each of the three alternative tests, we derived lower bounds on the MTFA, which can be used for false alarm control in practice.

We have conducted comprehensive numerical studies for the proposed algorithms in this article. Our windowed GLR test provides the best performance in terms of the trade-off between the MTFA and the WADD (CADD) among all of the tests considered. However, it may suffer from high computational complexity, especially for large networks. Fuh's approximation test does not perform as well when the change point is a time different than 1, which may limit its use in practice. Our mixture CUSUM test has a computational complexity of $\mathcal{O}(L)$, which is the most efficient among all of the tests. Moreover, it does not require knowledge of the transition probabilities, which might be hard to estimate in practice.

Future work includes developing low-complexity solutions for detecting multiple anomalies, as well as developing methods for implementing the detection algorithms in a distributed manner across the network.

Funding

The work of G. Rovatsos and V. V. Veeravalli was supported by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 and by the National Science Foundation (NSF) under grant CCF 16-18658. The work of S. Zou was supported in part by the National Science Foundation (NSF) under grant CCF-1948165.

References

Chen, B. and Willett, P. (1997). Quickest Detection of Hidden Markov Models, in *Proceedings of IEEE Conference on Decision and Control*, San Diego, CA, December.

Fellouris, G. and Sokolov, G. (2016). Second-order Asymptotic Optimality in Multisensor Sequential Change Detection, *IEEE Transactions on Information Theory* 62: 3662–3675. doi:10.1109/TIT.2016.2549042

Fellouris, G. and Tartakovsky, A. G. (2017). Multichannel Sequential Detection—Part I: Non-I.I.D. Data, *IEEE Transactions on Information Theory* 63: 4551–4571. doi:10.1109/TIT.2017.2689785

Fienberg, S. E. and Shmueli, G. (2005). Statistical Issues and Challenges Associated with Rapid Detection of Bio-Terrorist Attacks, *Statistics in Medicine* 24: 513–529. doi:10.1002/sim.2032

Frisn, M. (2009). Optimal Sequential Surveillance for Finance, Public Health, and Other Areas, *Sequential Analysis* 28: 310–337.

Fuh, C. D. (2003). SPRT and CUSUM in Hidden Markov Models, *Annals of Statistics* 31: 942–977. doi:10.1214/aos/1056562468

Fuh, C. D. (2004). Asymptotic Operating Characteristics of an Optimal Change Point Detection in Hidden Markov Models, *Annals of Statistics* 32: 2305–2339. doi:10.1214/009053604000000580

Fuh, C. D. and Mei, Y. (2015). Quickest Change Detection and Kullback-Leibler Divergence for Two-State Hidden Markov Models, *IEEE Transactions on Signal Processing* 63: 4866–4878. doi:10.1109/TSP.2015.2447506

Fuh, C. D. and Tartakovsky, A. G. (2019). Asymptotic Bayesian Theory of Quickest Change Detection for Hidden Markov Models, *IEEE Transactions on Information Theory* 65: 511–529. doi:10.1109/TIT.2018.2843379

Hadjiliadis, O., Zhang, H., and Poor, H. V. (2009). One Shot Schemes for Decentralized Quickest Change Detection, *IEEE Transactions on Information Theory* 55: 3346–3359. doi:10.1109/TIT.2009.2021311

Lai, L., Fan, Y., and Poor, H. V. (2008). Quickest Detection in Cognitive Radio: A Sequential Change Detection Framework, in *Proceedings of Global Telecommunications Conference (GLOBECOM)*, New Orleans, LA, November.

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

Lau, T. S., Tay, W. P., and Veeravalli, V. V. (2019). A Binning Approach to Quickest Change Detection with Unknown Post-Change Distribution, *IEEE Transactions on Signal Processing* 67: 609–621. doi:10.1109/TSP.2018.2881666

Lorden, G. (1971). Procedures for Reacting to a Change in Distribution, *Annals of Mathematical Statistics* 42: 1897–1908. doi:10.1214/aoms/1177693055

Lorden, G. and Pollak, M. (2008). Sequential Change-Point Detection Procedures That Are Nearly Optimal and Computationally Simple, *Sequential Analysis* 27: 476–512. doi:10.1080/07474940802446244

Ludkovski, M. (2012). Bayesian Quickest Detection in Sensor Arrays, *Sequential Analysis* 31: 481–504. doi:10.1080/07474946.2012.719437

Mechitov, K., Kim, W., Agha, G., and Nagayama, T. (2004). High-Frequency Distributed Sensing for Structure Monitoring, in *Proceedings of International Workshop on Networked Sensing Systems (INSS)*, Tokyo, Japan, June.

Mei, Y. (2010). Efficient Scalable Schemes for Monitoring a Large Number of Data Streams, *Biometrika* 97: 419–433. doi:10.1093/biomet/asq010

Moustakides, G. V. (2019). Detecting Changes in Hidden Markov Models, in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, Paris, France, July.

Moustakides, G. V. and Veeravalli, V. V. (2016). Sequentially Detecting Transitory Changes, in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, July.

Pollak, M. (1985). Optimal Detection of a Change in Distribution, *Annals of Statistics* 13: 206–227. doi:10.1214/aos/1176346587

Poor, H. V. and Hadjiliadis, O. (2009). *Quickest Detection*, Cambridge: Cambridge University Press.

Raghavan, V. and Veeravalli, V. V. (2010). Quickest Change Detection of a Markov Process across a Sensor Array, *IEEE Transactions on Information Theory* 56: 1961–1981. doi:10.1109/TIT.2010.2040869

Rice, J., Mechitov, K., Sim, S., Nagayama, T., Jang, S., Kim, R., Spencer, B., Agha, G., and Fujino, Y. (2010). Flexible Smart Sensor Framework for Autonomous Structural Health Monitoring, *Smart Structures and Systems* 6: 423–438. doi:10.12989/sss.2010.6.5_6.423

Rovatsos, G., Jiang, X., Domínguez-García, A. D., and Veeravalli, V. V. (2016). Comparison of Statistical Algorithms for Power System Line Outage Detection, in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March.

Rovatsos, G., Jiang, J., Domínguez-García, A. D., and Veeravalli, V. V. (2017a). Statistical Power System Line Outage Detection under Transient Dynamics, *IEEE Transactions on Signal Processing* 65: 2787–2797. doi:10.1109/TSP.2017.2673802

Rovatsos, G., Zou, S., and Veeravalli, V. V. (2017b). Quickest Change Detection under Transient Dynamics, in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, March.

Rovatsos, G., Zou, S., and Veeravalli, V. V. (2019). Quickest Detection of a Moving Target in a Sensor Network, in *Proceeding of IEEE International Symposium on Information Theory (ISIT)*, Paris, France, July.

Shewhart, W. A. (1925). The Application of Statistics as an Aid in Maintaining Quality of a Manufactured Product, *Journal of American Statistical Association* 20: 546–548. doi:10.1080/01621459.1925.10502930

Tartakovsky, A. G., Nikiforov, I. V., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Change-Point Detection*, Boca Raton, FL: CRC Press.

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006). A Novel Approach to Detection of Intrusions in Computer Networks via Adaptive Sequential and Batch-Sequential Change-Point Detection Methods, *IEEE Transactions on Signal Processing* 54: 3372–3382.

Tartakovsky, A. G. and Veeravalli, V. V. (2004). Change-Point Detection in Multichannel and Distributed Systems, in *Applied Sequential Methodologies: Real-World Examples with Data Analysis*, S. Datta, S. Chattopadhyay, and N. Mukhopadhyay, eds., pp 339–370, New York: Dekker.

Veeravalli, V. V. and Banerjee, T. (2013). *Quickest Change Detection*, Amsterdam: Elsevier.

Xie, Y. and Siegmund, D. (2013). Sequential Multi-Sensor Change-Point Detection, *Annals of Statistics* 41: 670–692. doi:10.1214/13-AOS1094

Zou, S., Fellouris, G., and Veeravalli, V. V. (2019). Quickest Change Detection Under Transient Dynamics: Theory and Asymptotic Analysis, *IEEE Transactions on Information Theory* 65: 1397–1412. doi:10.1109/TIT.2018.2877972

Zou, S. and Veeravalli, V. V. (2018). Quickest Detection of Dynamic Events in Sensor Networks, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Alberta, Canada, April.