Quickest Change Detection in Autoregressive Models

Zhongchang Sun, *Student Member, IEEE*, and Shaofeng Zou, *Member, IEEE*

Abstract— The problem of quickest change detection (QCD) in autoregressive (AR) models is investigated. A system is being monitored with sequentially observed samples. At some unknown time, a disturbance signal occurs and changes the distribution of the observations. The disturbance signal follows an AR model, which is dependent over time. Before the change, observations only consist of measurement noise, and are independent and identically distributed (i.i.d.). After the change, observations consist of the disturbance signal and the measurement noise, are dependent over time, which essentially follow a continuous-state hidden Markov model (HMM). The goal is to design a stopping time to detect the disturbance signal as quickly as possible subject to false alarm constraints. Existing approaches for general non-i.i.d. settings and discrete-state HMMs cannot be applied due to their high computational complexity and memory consumption, and they usually assume some asymptotic stability condition. In this paper, the asymptotic stability condition is firstly theoretically proved for the AR model by a novel design of forward variable and auxiliary Markov chain. A computationally efficient Ergodic CuSum algorithm that can be updated *recursively* is then constructed and is further shown to be asymptotically optimal. The data-driven setting where the disturbance signal parameters are unknown is further investigated, and an online and computationally efficient gradient ascent CuSum algorithm is designed. The algorithm is constructed by iteratively updating the estimate of the unknown parameters based on the maximum likelihood principle and the gradient ascent approach. The lower bound on its average running length to false alarm is also derived for practical false alarm control. Simulation results are provided to demonstrate the performance of the proposed algorithms.

Index Terms— Hidden Markov model, forward variable, sequential change detection, asymptotic optimality, non-i.i.d..

I. INTRODUCTION

THE problem of quickest change detection (QCD) has been widely studied in the literature [1], [2], [3], [4], where the goal is to detect an abrupt change in the data-generating distribution as quickly as possible, subject to

false alarm constraints. For the basic setting where the observations are independent over time, algorithms, e.g., cumulative sum (CuSum) algorithm and Shiryaev-Roberts algorithm have been proposed and have been shown to be optimal [5], [6], [7], [8], [9]. However, for a wide range of practical applications, observations may not be independent over time. For example, in power systems, the faulty signal is dependent in time [10] and such dependency is usually exploited by an autoregressive (AR) model [11], [12], [13], [14], [15].

In this paper, we investigate the problem of QCD in AR models. Specifically, before the change, the observed signal only consists of Gaussian measurement noise. After the change, a non-i.i.d. disturbance signal occurs in the system, and the observation consists of the disturbance signal and the measurement noise. The disturbance signal follows an AR model, and thus the observation equivalently follows a hidden Markov model (HMM) [16]. The goal is to detect the occurrence of the disturbance signal as quickly as possible subject to false alarm constraints.

A. Related Works

The problem in this paper is closely related to QCD in Markov chains [17], [18], [19], [20], [21], [22], [23], [24], [25], where the pre- and/or post-change samples follow the Markov model. In [17], optimal change detection algorithms were proposed for finite-state Markov chains under the Bayesian setting and the minimax setting. The continuous state Markov chain was studied in [24], where the CuSum algorithm was proved to be asymptotically optimal and its performance was characterized. In [19], the change detection in sensor networks under the Bayesian setting was studied where the change propagates across sensors and its propagation was modeled as a Markov process. It was shown that an exactly optimal algorithm for QCD of Markov chains under the Bayesian setting is a threshold test on the posterior of no change has happened in [23]. In [21], two CuSum type schemes were proposed for the QCD of finite-state Markov chain with unknown post-change transition probabilities, and bounds on the average detection delay (ADD) and the average running length (ARL) were further given for the proposed schemes. A maximum mean discrepancy (MMD) based method for change detection of Markov kernels with unknown post-change kernel was proposed in [22] and [25] for general state Markov chain. However, the algorithms and analyses for Markov chains cannot be applied to our problem directly. This is because with measurement noise, the

post-change samples in this paper essentially follow an HMM, and the hidden state is not directly observable.

The problem of QCD in discrete-state HMMs has been studied in [26], [27], [28], [29], [30], and [31]. In [26], Page's test was extended to the QCD of discrete-state HMMs and the ARL under both hypotheses were approximated. In [28], the detection of the change point at which the discrete-state HMM enters the absorbing state and the identification of the absorbing set was investigated and optimal decision rule was provided. The QCD problem in two-state HMM was studied in [29], where two computationally efficient schemes were developed. In [30], the Shiryaev algorithm was proved to be asymptotically optimal under some regularity conditions for the QCD in discrete-state HMMs under the Bayesian setting. The exact optimal algorithm for QCD in discrete-state HMM under the Bayesian setting was established in [31]. These works mainly focus on the discrete-state HMM. The AR model in our paper is an HMM model with a continuous state space. To the best of the author's knowledge, research on QCD in continuous HMMs is rather limited. The performance characterization of Shiryaev algorithm for the discrete-state HMM relies on the limiting theory for products of random matrices [32] and the computational complexity scales with the size of the state space. For the continuous-state Markov chain, the theory of random matrices is not applicable anymore since the number of state is uncountably infinite.

The AR models are commonly used in power system to characterize the disturbance signal and inter-area oscillations [11], [12], [13], [14], [15]. In [12], [33], [34], [35], [36], [37], and [38], the change detection in the AR model was studied where the observation is an AR time series. Our problem is fundamentally different from [12], [33], [34], [35], [36], [37], and [38] since the disturbance signal which has AR structure is unobservable in our problem. The observation is a noised version of the AR time series which makes our problem more challenging. The quickest detection of cyber-attacks in discrete-time linear dynamic system, which is an AR model, was studied in [14]. The Kalman filter [39] was first performed to estimate the state and the generalized CuSum algorithm was further designed using the state estimation. Though the Kalman filter is effective in state estimation, the performance of the detection algorithm is difficult to characterize due to the estimation error. In [40], a robust algorithm was proposed for distributed change detection in AR models with noised observations. However, no theoretical guarantee was provided. In our paper, we design an Ergodic CuSum algorithm directly and show that it is asymptotically optimal under the Lorden's criterion [41].

In our paper, we consider both the model-based setting where parameters of the disturbance signal are known and the data-driven setting where parameters of the disturbance signal are unknown. For the data-driven setting, our problem can be viewed as a composite QCD problem [24], [42], [43], [44], [45], [46], [47], [48]. Existing works mainly assume that the samples are i.i.d., while in the AR models, the observations are dependent in time. In [15], the QCD in AR model was studied where the properties of the post-change signal are unknown. Based on the assumption that the parameters of the post-change signal were small, the generalized likelihood ratio test (GLRT) was proposed. Our work is different from the works in [15] since we do not assume that the parameter of the post-change signal is small. Moreover, simulation results demonstrate that our proposed algorithm outperforms the GLRT in [15].

B. Contributions and Major Challenges

For the model-based setting, we design an Ergodic CuSum algorithm which can be updated recursively and thus is computationally efficient, and we theoretically prove that it is asymptotically optimal. While the general theory for non-i.i.d. setting were developed in e.g., [24], [49], [50], and [51], it relies on some asymptotic stability condition that the normalized log likelihood ratio between the post- and pre-change distributions converges to some finite and positive number. However, there are very few studies that try to verify this condition, e.g., discrete HMMs in [30] and Markov model in [49]. For the AR model, whether such stability condition holds remains unexplored. In this paper, we show that the normalized log likelihood ratio converges to some $\mathcal{K} > 0$ almost surely under the post-change distribution, i.e., satisfies the stability condition.

The difficulty of analyzing the convergence of the log likelihood ratio lies in that the log likelihood function for the AR model has a non-additive form due to the non-i.i.d. nature of the observation. For the discrete-state HMM in [30], the likelihood ratio function is represented by the L_1-norm of products of Markov random matrices and thus the log likelihood ratio has an additive form. However, the Markov random matrices techniques can not be applied to our problem since our AR model has uncountably infinite hidden states. To overcome this difficulty, we represent the likelihood function as the integration of the product of a sequence of functions based on the hidden Markov structure of the observation. We further show that the likelihood function can be represented as the integration of a scaled Gaussian density function. We define this scaled Gaussian density function as the forward variable and show that it can be updated recursively. Since the scaled Gaussian density function has parameterized representation, we develop a novel approach to write the conditional likelihood ratio using the parameter of the forward variable and the observation. The log likelihood ratio can then be written as the sum of the conditional log likelihood ratio and thus has an additive form. To characterize the convergence of the log likelihood ratio, we design an auxiliary Markov chain using the parameter of the forward variable and the observation and represent the likelihood function as a function of the auxiliary Markov chain. We further show that the stationary distribution of this auxiliary Markov chain exists and then show that $\mathcal{K}$ exists by applying the ergodic theorem [52], [53] to the auxiliary Markov chain. With the convergence of the log likelihood ratio, we derive the universal lower bound on the worst-case average detection delay (WADD).

For the CuSum algorithm designed for the general non-i.i.d. setting, e.g., [24], it is computationally expensive for

our AR model since the likelihood ratio for our AR model depends on the change point and we need to compute the likelihood ratio for every possible change point. Motivated by the dependence of the WADD lower bound on $\mathcal{K}$, we aim to find a computationally efficient test statistic that has a positive drift of $\mathcal{K}$ under the post-change distribution and has a negative drift under the pre-change distribution. We propose a novel Ergodic CuSum algorithm by exploiting the ergodicity of the underlying HMM and employing the likelihood ratio of the observations with change point at time 1. Since the likelihood function can be written as a function of the observation and the forward variable, our Ergodic CuSum algorithm can then be updated recursively, and thus is computationally efficient. From the ergodic theorem [52], [53] of Markov chain, the long term statistical performance of our test statistic is independent of the state when the change occurs. Based on this fact, the test statistic of our Ergodic CuSum algorithm converges to $\mathcal{K}$ in the post-change regime despite of the fact that time 1 is not the true change point. Under the pre-change distribution, we show that our test statistic has a negative drift. Moreover, since our test statistic is a likelihood ratio function of the current observation, the ARL lower bound can be derived following the proof of the ARL lower bound for general non-i.i.d. case. Therefore, we show that when satisfying the constraint on the average running length (ARL), the WADD upper bound of our Ergodic CuSum algorithm matches with the universal lower bound and thus the asymptotic optimality of our Ergodic CuSum algorithm follows.

For the data-driven setting, the generalized likelihood ratio test (GLRT) which replaces the unknown post-change parameter with its maximum likelihood estimate (MLE) [24] is not computationally efficient. We propose an online gradient ascent CuSum algorithm (OGA-CuSum) based on the online convex optimization (OCO) algorithm [48], [54]. We iteratively update the estimate of the unknown parameters based on the maximum likelihood principle using gradient ascent. Specifically, at each time, we compute the gradient of the log likelihood ratio function with respect to the unknown parameters based on the current observation and update the estimate in the direction of the gradient. We then replace the unknown parameter in the Ergodic CuSum with its estimate to construct the OGA-CuSum. The OGA-CuSum can be updated recursively using only the most recent sample and thus is memory and computationally efficient. We derive a lower bound on its ARL so that a threshold can be chosen analytically to control the false alarm. We also provide simulation results to demonstrate the performance of our algorithms.

C. Paper Organization

The remainder of this paper is organized as follows. In Section II, we present the problem formulation. In Section III, we present the design of a forward variable and establish the universal lower bound on WADD. In Section IV, we develop a computationally efficient Ergodic CuSum algorithm and show that it is asymptotically optimal under Lorden's criterion. In Section V, we design an online gradient ascent CuSum algorithm for the case where the

post-change parameters are unknown. Numerical results are provided to demonstrate the performance of our algorithms in VI. In Section VII, we present some concluding remarks.

II. PROBLEM FORMULATION

Consider a system being monitored by a sequentially observed signal $\boldsymbol{y}_t \in \mathbb{R}^K, t = 1, 2, \cdots$. At some unknown time t_0, a disturbance signal occurs in the system and changes the distribution of the observed signal. Specifically, before the change point t_0, the observed signal consists of the measurement noise $\boldsymbol{\nu}_t$ only, which is Gaussian distributed and independent over time:

$$\boldsymbol{y}_t = \boldsymbol{\nu}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \ t < t_0, \tag{1}$$

where $\boldsymbol{I}$ is the $K \times K$ identity matrix. The post-change signal consists of the disturbance signal $\boldsymbol{x}_t$ and the measurement noise:

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\nu}_t, \ t \geq t_0, \tag{2}$$

Note that the identity covariance matrix of $\boldsymbol{\nu}_t$ can be satisfied by applying a linear transformation to whiten the noise in the observation $\boldsymbol{y}_t$. The disturbance signal follows an AR model [11], [12], [13], [14], [15]:

$$\boldsymbol{x}_t = \sum_{i=1}^{q} \boldsymbol{A}_i \boldsymbol{x}_{t-i} + \boldsymbol{\omega}_t, \tag{3}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{K \times K}$ is the matrix coefficient and is invertible, $\boldsymbol{\omega}_t \in \mathbb{R}^K$ is the innovation noise vector and $\boldsymbol{\omega}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_\omega)$ and q denotes the order of the AR model. The disturbance signal $\boldsymbol{x}_t$ is not directly observable. The goal is to detect the change at time t_0 as soon as possible subject to false alarm constraints.

Denote by $\mathbb{P}_\infty$ the probability measure when there is no change, and denote by p_∞ the corresponding probability density. For any $t_0 > 0$, denote by $\mathbb{P}_{t_0}$ the probability measure when the change happens at t_0, and denote by p_{t_0} the corresponding probability density. It is clear that

$$p_\infty(\boldsymbol{y}_t|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t-1}) = p_\infty(\boldsymbol{y}_t),$$
$$p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t-1}) = p_\infty(\boldsymbol{y}_t), \ \text{if } t < t_0,$$
$$p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t-1}) = p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}), \ \text{if } t \geq t_0. \tag{4}$$

The AR model governs the evolution of $\boldsymbol{x}_t$. The combination of the state evolution model (3) and the observation model (2) is equivalent to an HMM. Let $f_1(\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_0+1}, \cdots, \boldsymbol{x}_{t_0+q-1})$ be the joint probability density of the initial state $(\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_0+1}, \cdots, \boldsymbol{x}_{t_0+q-1})$. Let $f(\boldsymbol{x}_t|\boldsymbol{x}_{t-q}, \cdots, \boldsymbol{x}_{t-1})$ denote the transitional probability density of the HMM and $g(\boldsymbol{y}_t|\boldsymbol{x}_t)$ be the conditional probability density of the observation $\boldsymbol{y}_t$ given the hidden state $\boldsymbol{x}_t$. The probability density p_{t_0} then satisfies that for $t \geq t_0 + q$

$$p_{t_0}(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)$$
$$= p_\infty(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t_0-1}) \cdot \int f_1(\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_0+1}, \cdots, \boldsymbol{x}_{t_0+q-1})$$
$$\cdot g(\boldsymbol{y}_{t_0}|\boldsymbol{x}_{t_0}) \cdots g(\boldsymbol{y}_{t_0+q-1}|\boldsymbol{x}_{t_0+q-1})$$

$$\cdot f(\boldsymbol{x}_{t_0+q}|\boldsymbol{x}_{t_0}, \cdots, \boldsymbol{x}_{t_0+q-1}) \cdots g(\boldsymbol{y}_t|\boldsymbol{x}_t)d\boldsymbol{x}_{t_0}\boldsymbol{x}_{t_0+1}\cdots\boldsymbol{x}_t. \tag{5}$$

In this paper, we consider a deterministic but unknown change point t_0. The goal is to detect the change as quickly as possible subject to false alarm constraints based on the sequentially observed samples. Let $\mathcal{F}_t$ be the σ-algebra generated by the first t samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_t$. A stopping time τ is a random variable with the property that for each t, the event $\{\tau = t\} \in \mathcal{F}_t$. We define the worse-case average detection delay (WADD) under Lorden's criterion [41] and the average running length (ARL) for any stopping time τ as follows:

$$\text{WADD}(\tau) \triangleq \sup_{t_0 \geq 1} \text{esssup} \mathbb{E}_{t_0} \left[(\tau - t_0)^+ | \boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t_0-1} \right],$$

$$\text{ARL}(\tau) \triangleq \mathbb{E}_\infty[\tau], \tag{6}$$

where $\mathbb{E}_{t_0}$ ($\mathbb{E}_\infty$) denotes the expectation under the probability measure $\mathbb{P}_{t_0}$ ($\mathbb{P}_\infty$). The goal is to design a stopping time that minimizes the WADD subject to the constraint on the ARL:

$$\inf_{\tau:\text{ARL}(\tau)\geq\gamma} \text{WADD}(\tau), \tag{7}$$

where $\gamma > 0$ is a pre-specified threshold.

III. Universal Lower Bound on WADD

For convenience, we first present results for the first-order AR model to illustrate our approach. We will then show the generalization to any q-th order AR model. We note that any q-th order AR model can be equivalently converted to a first-order AR model (see details in Appendix A).

Specifically, the first-order AR model is defined as:

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\omega}_t, \tag{8}$$

where $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ is the matrix coefficient and is invertible. The initial disturbance signal $\boldsymbol{x}_{t_0}$ is assumed to be Gaussian with probability density $f_1(\boldsymbol{x}_{t_0})$ and is independent of the pre-change observations. We assume that the operator norm of $\boldsymbol{A}$ is strictly less than 1. This assumption guarantees the stability of this system [55].

In the following, we derive a universal lower bound on the WADD for any τ satisfying the false alarm constraint: $\mathbb{E}_\infty[\tau] \geq \gamma$. To derive the universal lower bound, we first prove the following stability condition [24], [49], [50], [51]: convergence of the log likelihood ratio

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}. \tag{9}$$

Note that for the AR model, the log likelihood ratio in (9) does not admit an additive form, and thus its convergence analysis is challenging. To overcome this difficulty, we first introduce a forward variable α_t, which is a scaled Gaussian density function and can be updated recursively. We show that the conditional density $p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})$ can be represented using $\boldsymbol{y}_t$ and the parameters of the forward variable α_t. The log likelihood ratio in (9) can then be written in an additive form. To further prove the convergence of the log likelihood ratio, we construct an auxiliary Markov chain using $\boldsymbol{y}_t$ and the parameter of the forward variable α_t. We show

that the stationary distribution of this auxiliary Markov chain exists and then show the convergence of the log likelihood ratio in (9) by applying the ergodic theorem [52], [53]. The universal lower bound on the WADD can then be derived.

A. Forward Variable

We first introduce a forward variable, which plays an important role in deriving the universal lower bound on the WADD. Moreover, our proposed algorithm in Section IV can be recursively updated using the forward variable.

Note that for the i.i.d. setting, the log likelihood ratio of $\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t$ can be written as the sum of the log likelihood ratio of each individual sample. For the discrete state HMM in [30], the likelihood ratio function is represented by the L_1-norm of products of Markov random matrices and thus the log likelihood ratio has an additive form. However, the Markov random matrices techniques can not be applied to our problem since our AR model has uncountably infinite hidden states.

Observe that $p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t) = \int f_1(\boldsymbol{x}_{t_0})g(\boldsymbol{y}_{t_0}|\boldsymbol{x}_{t_0})f(\boldsymbol{x}_{t_0+1}|\boldsymbol{x}_{t_0}) \cdots g(\boldsymbol{y}_t|\boldsymbol{x}_t)d\boldsymbol{x}_{t_0}\boldsymbol{x}_{t_0+1}\cdots\boldsymbol{x}_t$ is the integration of the products of a sequence of functions. This motivates us to represent the likelihood ratio function by replacing the Markov random matrices and the L_1-norm of matrices in [30] with functions and integration of functions, respectively. Moreover, we leverage the Gaussian property of the innovation noise and the measurement noise in our AR model to further show that these functions can be represented using the parameters of the Gaussian density function. Therefore, the likelihood ratio function admits an additive form and can be updated efficiently.

We define the following forward variable to compute the likelihood function after the change point ($t \geq t_0$),

$$\alpha_t(\boldsymbol{x}_t) = p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t, \boldsymbol{x}_t). \tag{10}$$

Note that $\alpha_t(\boldsymbol{x}_t)$ also depends on $\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t$. For notational convenience, we treated $\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t$ as known parameters here and write α_t as only a function of $\boldsymbol{x}_t$. We then have that

$$p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t) = \int \alpha_t(\boldsymbol{x}_t)d\boldsymbol{x}_t. \tag{11}$$

It can be easily verified that the following recursion holds for the forward variable:

$$\alpha_{t+1}(\boldsymbol{x}_{t+1}) = \int \alpha_t(\boldsymbol{x}_t)f(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)g(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})d\boldsymbol{x}_t.$$

We first provide a formal definition for the Gaussian function.

Definition 1: A Gaussian function $f'(\cdot) : \mathbb{R}^K \to \mathbb{R}$ is a function of the form

$$f'(\boldsymbol{x}) = a \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{12}$$

where a is a constant, $\boldsymbol{\mu} \in \mathbb{R}^K$ is an arbitrary vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$ is a positive definite matrix.

In the following lemma, we show that $\alpha_t(\boldsymbol{x}_t)$ is a Gaussian function of $\boldsymbol{x}_t$.

Lemma 1: $\alpha_t(\boldsymbol{x}_t)$ is a Gaussian function of $\boldsymbol{x}_t$.

Proof: We prove this by induction. If the distribution of $\boldsymbol{x}_{t_0}$ is Gaussian, $\alpha_1(\boldsymbol{x}_{t_0}) = f_1(\boldsymbol{x}_{t_0})g(\boldsymbol{y}_{t_0}|\boldsymbol{x}_{t_0})$ is a Gaussian distribution times a coefficient, which is a Gaussian function. This can be proved by computing the product of two Gaussian density functions. Assume that for some $c_{t-1}, \boldsymbol{\Sigma}_{t-1}$ and $\boldsymbol{\mu}_{t-1}$,

$$\alpha_{t-1}(\boldsymbol{x}_{t-1}) = \frac{c_{t-1}}{\sqrt{(2\pi)^K \det(\boldsymbol{\Sigma}_{t-1})}}$$
$$\cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_{t-1})^\top \boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_{t-1})\right), \quad (13)$$

which is a Gaussian function of $\boldsymbol{x}_{t-1}$. We will show that $\alpha_t(\boldsymbol{x}_t)$ is also a Gaussian function of $\boldsymbol{x}_t$. From the product and the convolution of Gaussian density functions, we have that

$$\alpha_t(\boldsymbol{x}_t) = \int \alpha_t(\boldsymbol{x}_{t-1}) f(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) g(\boldsymbol{y}_t|\boldsymbol{x}_t) d\boldsymbol{x}_{t-1}$$
$$= \frac{c_t}{\sqrt{(2\pi)^K \det(\boldsymbol{\Sigma}_t)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_t)\right), \quad (14)$$

where

$$\boldsymbol{\Sigma}_t = (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1},$$
$$\boldsymbol{\mu}_t = (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\mu}_{t-1}$$
$$+ (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{y}_t, \quad (15)$$

and

$$\frac{c_t}{c_{t-1}} = \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}}$$
$$\cdot \frac{1}{\sqrt{\det((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}$$
$$\cdot \exp\left(-\frac{1}{2}\left((\boldsymbol{A}\boldsymbol{\mu}_{t-1})^\top (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1})\right.\right.$$
$$+ \boldsymbol{y}_t^\top \boldsymbol{y}_t - ((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}) + \boldsymbol{y}_t)^\top$$
$$\cdot \left((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I}\right)^{-1}$$
$$\cdot \left.\left.((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}) + \boldsymbol{y}_t)\right)\right). \quad (16)$$

Moreover, from the update rule of $\boldsymbol{\Sigma}_t$, it can be seen that $\boldsymbol{\Sigma}_t$ is positive definite. Therefore, $\alpha_t(\boldsymbol{x}_t)$ is a scaled Gaussian distribution of $\boldsymbol{x}_t$ with coefficient c_t. □

With the forward variable, the conditional density $p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})$ can be written as follows:

$$p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}) = \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}$$
$$= \frac{\int \alpha_t(\boldsymbol{x}_t) d\boldsymbol{x}_t}{\int \alpha_{t-1}(\boldsymbol{x}_{t-1}) d\boldsymbol{x}_{t-1}} = \frac{c_t}{c_{t-1}}, \quad (17)$$

where the last equality is due to the fact that $\alpha_t(\boldsymbol{x}_t)$ is a scaled Gaussian density function of $\boldsymbol{x}_t$ with coefficient c_t. Note that before the change, the observations are i.i.d.. Therefore, it suffices to consider the additive form for the post-change log likelihood function. We have the following additive form for the log likelihood function $\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)$:

$$\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)$$

$$= \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})} + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-2})}$$
$$+ \cdots + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1})}{p_{t_0}(\boldsymbol{y}_{t_0})} + \log p_{t_0}(\boldsymbol{y}_{t_0})$$
$$= \sum_{i=t_0}^{t} \frac{c_i}{c_{i-1}}, \quad (18)$$

where $\frac{c_{t_0}}{c_{t_0-1}} = \log p_{t_0}(\boldsymbol{y}_{t_0})$. As shown in (16), $\frac{c_t}{c_{t-1}}$ is a function of $\boldsymbol{y}_t, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}$. Therefore, the log likelihood function can be written in an additive form using the parameters of the forward variable.

It can be seen that at each time step, we only need to compute $c_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$, and the forward variable α_t can be fully recovered. Moreover, $c_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$ can be updated recursively, and thus the likelihood function $p_{t_0}(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)$ can be computed efficiently using the forward variable α_t.

B. Universal Lower Bound on WADD

The general theory for QCD with non-i.i.d. samples relies on the assumption that the normalized log likelihood ratio between the post- and pre-change distributions converges to some finite and positive number [24], [49], [50], [51]. For the AR model, whether such assumption holds remains unexplored. Here, we first show that as $t \to \infty$, the limit of $\frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$ exists and is positive, denoted by $\mathcal{K}$, and further provide the universal lower bound on WADD in the following theorem. The explicit expression of $\mathcal{K}$ will be provided later after we introduce necessary notations.

Theorem 1: We have that

$$\lim_{t\to\infty} \frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})} = \mathcal{K} \quad (19)$$

almost surely under $\mathbb{P}_{t_0}$ with $\mathcal{K} > 0$. Moreover, as $\gamma \to \infty$,

$$\inf_{\tau:\text{ARL}(\tau)\geq\gamma} \text{WADD}(\tau) \geq \frac{\log\gamma}{\mathcal{K}}(1 + o(1)). \quad (20)$$

Since $\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1}, \cdots, \boldsymbol{y}_t$ are dependent, the Law of Large Number used in the i.i.d. setting is not applicable, even if the log likelihood function has an additive form in (18). We develop a novel approach to show the convergence of $\lim_{t\to\infty} \frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$ under $\mathbb{P}_{t_0}$.

Observe that $\frac{c_t}{c_{t-1}}$ is a function of $\boldsymbol{y}_t, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}$ and $\boldsymbol{y}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$ are dependent over time. This motivates us to formulate a new Markov chain using $\boldsymbol{y}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$ and apply the ergodic theorem [52], [53] to characterize the convergence of $\lim_{t\to\infty} \frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$ under $\mathbb{P}_{t_0}$.

From the update rule of $\boldsymbol{\Sigma}_t$ in (15), it can be seen that $\boldsymbol{\Sigma}_t$ does not depend on $\boldsymbol{y}_t$ and thus is deterministic. We then show that $\boldsymbol{\Sigma}_t$ converges as $t \to \infty$ and let $\boldsymbol{\Sigma}^* \triangleq \lim_{t\to\infty} \boldsymbol{\Sigma}_t$. Let $\boldsymbol{\mu}_t^*$ be the parameter of the forward variable when $\boldsymbol{\Sigma}_{t_0} = \boldsymbol{\Sigma}^*$. We formulate an auxiliary Markov chain using $\boldsymbol{y}_t, \boldsymbol{\mu}_t^*$ as an intermediate step to prove the convergence of the log likelihood ratio. Specifically, denote by $p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)$ the likelihood function when $\boldsymbol{\Sigma}_{t_0} = \boldsymbol{\Sigma}^*$. With the auxiliary Markov chain $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$, we show that $\lim_{t\to\infty} \frac{1}{t} p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})$ converges under $\mathbb{P}_{t_0}$. The

convergence of $\lim_{t\to\infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}$ under $\mathbb{P}_{t_0}$ is then proved by showing that

$$\lim_{t\to\infty} \frac{1}{t}\big(\log p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})$$
$$- \log p_{t_0}^*(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})\big) = 0 \qquad (21)$$

under $\mathbb{P}_{t_0}$ almost surely.

We first show that $\boldsymbol{\Sigma}_t$ converges in the following Lemma.

Lemma 2: As $t \to \infty$, $\boldsymbol{\Sigma}_t$ converges to $\boldsymbol{\Sigma}^*$.

Proof: The proof can be found in Appendix B. □

We then construct the auxiliary Markov chain to facilitate our proof. For the auxiliary Markov chain, we let $\boldsymbol{\Sigma}_{t_0} = \boldsymbol{\Sigma}^*$. Then $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}^*, \forall t \geq t_0$. It follows that $\boldsymbol{\mu}_t^*$ can be updated recursively as follows:

$$\boldsymbol{\mu}_t^* = (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*$$
$$+ (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{y}_t$$
$$= (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\mu}_{t-1}^* + \boldsymbol{\Sigma}^*\boldsymbol{y}_t. \qquad (22)$$

In the following, we denote the stationary distribution of a Markov chain in the post-change phase by π. For examples, $\pi(\boldsymbol{x}_t)$ is the stationary distribution of $\{\boldsymbol{x}_t\}_{t=t_0}^\infty$, $\pi(\boldsymbol{x}_t,\boldsymbol{y}_t)$ is the stationary distribution of $\{\boldsymbol{x}_t,\boldsymbol{y}_t\}_{t=t_0}^\infty$. We first show that $\{\boldsymbol{x}_t,\boldsymbol{y}_t\}_{t=t_0}^\infty$ has a unique stationary distribution in the following lemma, which will be used to formulate and characterize the property of the auxiliary Markov chain.

Lemma 3: The HMM $\{\boldsymbol{x}_t,\boldsymbol{y}_t\}_{t=t_0}^\infty$ is π-irreducible. Moreover, $\pi(\boldsymbol{x}_t)g(\boldsymbol{y}_t|\boldsymbol{x}_t)$ is the unique stationary distribution of $\{\boldsymbol{x}_t,\boldsymbol{y}_t\}_{t=t_0}^\infty$, where $\pi(\boldsymbol{x}_t)$ is a Gaussian distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma} = \sum_{i=0}^\infty (\boldsymbol{A}^\top)^i \boldsymbol{R}_\omega \boldsymbol{A}^i$.

Proof: The detailed proof can be found in Appendix C. □

In the following lemma, we show that $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain and the stationary distribution of $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ exists and is unique.

Lemma 4: $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain. Moreover, $\int \pi(\boldsymbol{\mu}_{t-1}^*) p_{t_0}(\boldsymbol{y}_t,\boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*)d\boldsymbol{\mu}_{t-1}^*$ is the unique stationary distribution of $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$, where $\pi(\boldsymbol{\mu}_{t-1}^*)$ is the stationary distribution of $\boldsymbol{\mu}_{t-1}^*$ and is guaranteed to exist.

Proof: The proof can be found in Appendix D. □

With the auxiliary Markov chain $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ and its stationary distribution in Lemma 4, we are ready to show the convergence of $\lim_{t\to\infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}$ and prove Theorem 1.

Proof Sketch of Theorem 1: We first consider the auxiliary Markov chain $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ and show that

$$\lim_{t\to\infty} \frac{1}{t} \log p_{t_0}^*(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1}) = \mathbb{E}_\pi[h(\boldsymbol{\mu},\boldsymbol{y})] \qquad (23)$$

almost surely under $\mathbb{P}_{t_0}$ by applying the ergodic theorem [52, Theorem 4], [53, Theorem 17.0.1] to the Markov chain $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$, where $h(\boldsymbol{\mu},\boldsymbol{y})$ is a quadratic function of $\boldsymbol{\mu}, \boldsymbol{y}$ and $(\boldsymbol{y},\boldsymbol{\mu})$ follows the stationary distribution of the auxiliary Markov chain $\{\boldsymbol{y}_t,\boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$. The explicit expression of $h(\boldsymbol{\mu},\boldsymbol{y})$ is as follows

$$h(\boldsymbol{\mu},\boldsymbol{y}) = \log\Big(\frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}}$$

$$\cdot \frac{1}{\sqrt{\det((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}\Big)$$
$$- \frac{1}{2}\Big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu} - \boldsymbol{\Sigma}^*\boldsymbol{y}))^\top$$
$$\cdot (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu} - \boldsymbol{\Sigma}^*\boldsymbol{y}))$$
$$+ \boldsymbol{y}^\top\boldsymbol{y} - ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})$$
$$\cdot (\boldsymbol{\mu} - \boldsymbol{\Sigma}^*\boldsymbol{y})) + \boldsymbol{y})^\top((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})^{-1}$$
$$\cdot ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu} - \boldsymbol{\Sigma}^*\boldsymbol{y}))$$
$$+ \boldsymbol{y})\Big). \qquad (24)$$

Moreover, if we let $\boldsymbol{y} = \boldsymbol{y}_t$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_t^*$, we have that $h(\boldsymbol{\mu}_t^*,\boldsymbol{y}_t) = p_{t_0}^*(\boldsymbol{y}_t|\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t-1})$.

We then show that $\lim_{t\to\infty} \frac{1}{t} \log p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1}) = \mathbb{E}_\pi[h(\boldsymbol{\mu},\boldsymbol{y})]$ under $\mathbb{P}_{t_0}$ almost surely by showing that

$$\lim_{t\to\infty} \frac{1}{t}\big(\log p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})$$
$$- \log p_{t_0}^*(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})\big) = 0 \qquad (25)$$

under $\mathbb{P}_{t_0}$ almost surely.

Since the observations are independent under $\mathbb{P}_\infty$, we have that

$$\lim_{t\to\infty} \frac{1}{t} \log p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1}) = \lim_{t\to\infty} \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \log p_\infty(\boldsymbol{y}_i).$$

We then have that under $\mathbb{P}_{t_0}$,

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \log p_\infty(\boldsymbol{y}_i) = \mathbb{E}_\pi[\log p_\infty(\boldsymbol{y})]$$

almost surely from the ergodic theorem of Markov chain [52, Theorem 4], [53, Theorem 17.0.1].

Let

$$\mathcal{K} = \mathbb{E}_\pi[h(\boldsymbol{\mu},\boldsymbol{y})] - \mathbb{E}_\pi[\log p_\infty(\boldsymbol{y})]. \qquad (26)$$

It then follows that for any initial state $\boldsymbol{y}_{t_0},\boldsymbol{\mu}_{t_0}$,

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})} = \mathcal{K}$$

under $\mathbb{P}_{t_0}$ almost surely. Therefore, for any $\eta > 0$,

$$\lim_{t\to\infty} \sup_{t_0\geq 1} \operatorname{esssup}\mathbb{P}_{t_0}\Bigg\{ \max_{k\leq t} \sum_{i=t_0}^{t_0+k-1} \log \frac{p_{t_0}(\boldsymbol{y}_i|\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i)}$$
$$\geq \mathcal{K}(1+\eta)t\Big|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}\Bigg\} = 0. \qquad (27)$$

Then (20) follows from [24, Theorem 1]. The full proof can be found in Appendix E. □

Remark 1: Results in [24] assumes $\lim_{t\to\infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_{t_0+t-1})}$ exists under $\mathbb{P}_{t_0}$ while in our results, we prove its existence and characterize its value.

C. q-th Order AR Models

In this section, we show that our results for first-order AR models can be generalized to any q-th order AR models.

In the following theorem, we show that for any q-th order AR model, as in Section III-B, as $t \to \infty$, the limit of $\frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$ exists and is positive, denoted by $\widetilde{\mathcal{K}}$. The expression of $\widetilde{\mathcal{K}}$ can be derived similarly as in (26). The universal lower bound on WADD then follows from [24, Theorem 1].

Theorem 2: For a q-th order AR model, we have that

$$\lim_{t \to \infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})} = \widetilde{\mathcal{K}} \qquad (28)$$

almost surely under $\mathbb{P}_{t_0}$ where $\widetilde{\mathcal{K}} > 0$. Moreover, as $\gamma \to \infty$,

$$\inf_{\tau:\text{ARL}(\tau) \geq \gamma} \text{WADD}(\tau) \geq \frac{\log \gamma}{\widetilde{\mathcal{K}}}(1 + o(1)). \qquad (29)$$

The proof of Theorem 2 is similar to the proof of Theorem 1, and the idea is to convert the q-th order AR model to a first-order AR model, and then apply the proof of Theorem 1. Below, we provide a proof sketch.

Proof Sketch: Note that for a q-th order AR model, it can be converted to a first-order AR model (See Appendix A). Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to x. For notational convenience, let $\boldsymbol{y}_{t_1, t_2}$ denote the sequence $\boldsymbol{y}_{t_1}, \boldsymbol{y}_{t_1+1}, \cdots, \boldsymbol{y}_{t_2}$. We have that

$$\lim_{t \to \infty} \frac{1}{t} \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$$
$$= \lim_{t \to \infty} \frac{1}{t} \left(\log \frac{p_{t_0}(\boldsymbol{y}_{t_0, t_0+q-1})}{p_\infty(\boldsymbol{y}_{t_0, t_0+q-1})} \right.$$
$$+ \log \frac{p_{t_0}(\boldsymbol{y}_{t_0+q, t_0+2q-1}|\boldsymbol{y}_{t_0, t_0+q-1})}{p_\infty(\boldsymbol{y}_{t_0+q, t_0+2q-1}|\boldsymbol{y}_{t_0, t_0+q-1})} + \cdots$$
$$+ \log \frac{p_{t_0}(\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1}|\boldsymbol{y}_{t_0, t_0+q(\lfloor \frac{t}{q} \rfloor-1)-1})}{p_\infty(\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1}|\boldsymbol{y}_{t_0, t_0+q(\lfloor \frac{t}{q} \rfloor-1)-1})}$$
$$\left. + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0+q\lfloor \frac{t}{q} \rfloor, t_0+t-1}|\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1})}{p_\infty(\boldsymbol{y}_{t_0+q\lfloor \frac{t}{q} \rfloor, t_0+t-1}|\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1})} \right)$$
$$= \lim_{t \to \infty} \frac{1}{t} \left(\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+q-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+q-1})} \right.$$
$$+ \log \frac{p_{t_0}(\boldsymbol{y}_{t_0+q, t_0+2q-1}|\boldsymbol{y}_{t_0, t_0+q-1})}{p_\infty(\boldsymbol{y}_{t_0+q, t_0+2q-1}|\boldsymbol{y}_{t_0, t_0+q-1})} + \cdots$$
$$\left. + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1}|\boldsymbol{y}_{t_0, t_0+q(\lfloor \frac{t}{q} \rfloor-1)-1})}{p_\infty(\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1}|\boldsymbol{y}_{t_0, t_0+q(\lfloor \frac{t}{q} \rfloor-1)-1})} \right)$$
$$= \lim_{t \to \infty} \frac{1}{t} \log \frac{p_{t_0}\big((\boldsymbol{y}_{t_0, t_0+q-1}), \cdots, (\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1})\big)}{p_\infty\big((\boldsymbol{y}_{t_0, t_0+q-1}), \cdots, (\boldsymbol{y}_{t_0+q(\lfloor \frac{t}{q} \rfloor-1), t_0+q\lfloor \frac{t}{q} \rfloor-1})\big)}$$
$$= \lim_{t' \to \infty} \frac{1}{qt'} \log \frac{p_{t_0}\big((\boldsymbol{y}_{t_0, t_0+q-1}), \cdots, (\boldsymbol{y}_{t_0+q(t'-1), t_0+qt'-1})\big)}{p_\infty\big((\boldsymbol{y}_{t_0, t_0+q-1}), \cdots, (\boldsymbol{y}_{t_0+q(t'-1), t_0+qt'-1})\big)}$$
$$= \widetilde{\mathcal{K}} \qquad (30)$$

almost surely under $\mathbb{P}_{t_0}$, where the second equality is due to the fact that $(\boldsymbol{y}_{t_0+q\lfloor \frac{t}{q} \rfloor}, \cdots, \boldsymbol{y}_{t_0+t-1})$ only contains finite additional observations (less than q) and the last equality is from (19) and the fact that $\{(\boldsymbol{y}_{t_0+qi}, \cdots, \boldsymbol{y}_{t_0+q(i+1)-1})\}_{i=0}^{\infty}$ follows a first-order AR model. For the quickest change

detection in q-th order AR models, we then have that as $\gamma \to \infty$,

$$\inf_{\tau:\text{ARL}(\tau) \geq \gamma} \text{WADD}(\tau) \geq \frac{\log \gamma}{\widetilde{\mathcal{K}}}(1 + o(1)). \qquad (31)$$

$\square$

IV. ASYMPTOTICALLY OPTIMAL STOPPING TIME

In this section, we first present the algorithm and its optimality results for the first-order AR model. We then show the generalization to the q-th order AR model.

A. First-Order AR Model

The CuSum algorithm based on the generalized likelihood ratio (GLR) approach has been widely used for QCD problems. For the AR model, the GLR statistic is defined as follows:

$$W_t = \max_{1 \leq k \leq t} \sum_{i=k}^{t} \log \frac{p_k(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1})}. \qquad (32)$$

The CuSum algorithm [24] can then be designed:

$$\tau_c = \inf \Big\{ t : W_t \geq c \Big\}. \qquad (33)$$

In [24], a special non-i.i.d. case was studied where the post-change distribution does not depend on the change point t_0, i.e., $p_k(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1})$ in (32) does not depend on k. Under the assumption that $\frac{1}{t}\sum_{i=t_0}^{t_0+t-1} \log \frac{p_{t_0}(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1})}$ converges in probability under $\mathbb{P}_{t_0}$ to some positive constant $\mathcal{K}$, the asymptotic optimality of the CuSum algorithm was proved. The non-i.i.d. case under the Bayesian setting was studied in [56], where the post-change distribution does not depend on the change point t_0 as in [24]. The asymptotic optimality of Shiryaev procedure was established under the same stability assumption.

If the post-change distribution does not depend on the change-point, then the CuSum algorithm can be updated recursively and is computationally efficient. However, for our AR model, under $\mathbb{P}_{t_0}$, $p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})$ depends on t_0 for $t \geq t_0$. At each time t, we need to update $p_k(\boldsymbol{y}_t|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t-1})$ for every $1 \leq k \leq t$, the complexity of which scales with t, which is not practically feasible.

In this section, we propose a computationally efficient Ergodic CuSum algorithm and further show that it is asymptotically optimal.

Motivated by the fact that the WADD is lower bounded by $\frac{\log \gamma}{\mathcal{K}}(1 + o(1))$, we aim to find a computationally efficient test statistic that has a positive drift of $\mathcal{K}$ under the post-change distribution and has a negative drift under the pre-change distribution. Define the likelihood ratio of the first t observations when the change point $t_0 = 1$ by

$$L_t = \frac{p_1(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)}, \qquad (34)$$

and let $L_0 = 1$. For any $t_0 \geq 1$, we have that

$$\lim_{t \to \infty} \frac{1}{t} \log \frac{p_1(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)}$$

$$\begin{aligned}
&= \lim_{t\to\infty} \frac{1}{t}\bigg(\log\frac{p_1(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}\\
&\quad + \log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}\bigg)\\
&= \lim_{t\to\infty} \frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}.
\end{aligned}\tag{35}$$

Note that different sample trajectories $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}$ lead to different values of $\boldsymbol{\mu}_{t_0}$. However, the ergodic theorem of Markov chain [52, Theorem 4], [53, Theorem 17.0.1] implies that the convergence of $\frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}$ does not depend on $\boldsymbol{\mu}_{t_0}$ and thus does not depend on the sample trajectory $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}$. Therefore, we have that for any $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}$, under $\mathbb{P}_{t_0}$ almost surely

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}\\
&= \lim_{t\to\infty} \frac{1}{t}\log\frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t)}\\
&= \mathcal{K},
\end{aligned}\tag{36}$$

where the first equality is due to the fact that the convergence of the ergodic Markov chain doesn't depend on the initial state [52, Theorem 4], [53, Theorem 17.0.1]. Therefore, under the post-change distribution, L_t has a positive drift $\mathcal{K}$ on average. Let $D(p\|q)$ denote the KL-Divergence between two distributions. Under the pre-change distribution, we have that $\mathbb{E}_\infty\big[\log\frac{p_1(\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1})}{p_\infty(\boldsymbol{y}_t)}\big] = -D\big(p_\infty(\boldsymbol{y}_t)\|p_1(\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1})\big) \le 0$ where the equality holds when $p_1(\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1}) = p_\infty(\boldsymbol{y}_t)$ almost surely. Therefore, under the pre-change distribution, L_t has a negative drift as long as $p_1(\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1}) \ne p_\infty(\boldsymbol{y}_t)$.

Motivated by these facts, we define the Ergodic CuSum statistic

$$\begin{aligned}
S_t &= \max_{0\le i\le t}(\log L_t - \log L_i)\\
&= \max\big(0, S_{t-1} + \log L_t - \log L_{t-1}\big).
\end{aligned}\tag{37}$$

The Ergodic CuSum algorithm is then defined as

$$\tau_c^* = \inf\{t: S_t \ge c\}.\tag{38}$$

At each time t, we only need to compute L_t. It can be easily verified using (18) that L_t admits an additive form. Moreover, L_t can be recursively updated using the forward variable as shown in Lemma 1. Specifically, let $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ be the parameters of the initial distribution of $\boldsymbol{x}_0$. We have that $\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$ can be updated recursively according to (15) for each t. We then have that $L_t = \sum_{i=1}^t \frac{c_i}{c_{i-1}} = L_{t-1} + \frac{c_t}{c_{t-1}}$, where $\frac{c_t}{c_{t-1}}$ can be computed using $\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t$ as shown in (16).

We note that S_t is not the actual generalized likelihood ratio, and is different from W_t in τ_c. However, at each time t, $\log\frac{p_1(\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1})}{p_\infty(\boldsymbol{y}_t)}$ can still be viewed as a log likelihood ratio of the current sample $\boldsymbol{y}_t$. Therefore, the ARL lower bound can be derived following the proof of the ARL lower bound in [24] for general non-i.i.d. case. Since the convergence of $\lim_{t\to\infty} \frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}$ under $\mathbb{P}_{t_0}$ does not depend on the initial state $\boldsymbol{\mu}_{t_0}, \boldsymbol{y}_{t_0}$, it can be shown that τ_c^* is asymptotically optimal for (7).

In the following theorem, we show 1) the ARL lower bound of τ_c^* and 2) the WADD upper bound of τ_c^*.

Theorem 3: 1) Let $c = \log\gamma$ in (38), then $\mathbb{E}_\infty[\tau_c^*] \ge \gamma$; and 2) as $\gamma\to\infty$, $\text{WADD}(\tau_c^*) \le \frac{\log\gamma}{\mathcal{K}}(1+o(1))$.

Proof: The proof can be found in Appendix F. □

Based on Theorem 1 and Theorem 3, we establish the asymptotic optimality of τ_c^* in the following theorem.

Theorem 4: τ_c^* is asymptotically optimal.

Proof: By Theorem 1 and Theorem 3, we establish the asymptotic optimality of τ_c^*. □

The Ergodic CuSum algorithm in (38) is computationally efficient and asymptotically optimal for detecting the change in the AR model in Section II.

B. q-th Order AR Model

We first convert the post-change q-th order AR model equivalently into a first-order AR model (see Appendix A). We then partition the sequence of the observations into a sequence of non-overlapping blocks with size q. Specially, define $\tilde{y}_t = (y_{(t-1)q+1}, \ldots, y_{tq})$, for $t = 1, 2, \ldots$. We then apply our Ergodic Cusum algorithm in (38) on the sequence of $\{\tilde{y}_t\}_{t=1}^\infty$. We show that our Ergodic Cusum algorithm is asymptotically optimal for the problem of QCD in q-th order AR models.

For the ARL lower bound and WADD upper bound of τ_c^*, we have the following theorem.

Theorem 5: For the QCD problem in q-th order AR models, consider τ_c^* applied on $\{\tilde{y}_t\}_{t=1}^\infty$. 1) Let $c = \log\gamma$ in (38), then $\mathbb{E}_\infty[\tau_c^*] \ge \gamma$. 2) As $\gamma\to\infty$, $\text{WADD}(\tau_c^*) \le \frac{\log\gamma}{\widetilde{\mathcal{K}}}(1+o(1))$.

The proof of Theorem 5 is similar to the proof of Theorem 3. Here, we only provide a proof sketch.

Proof Sketch: When the post-change disturbance signal follows a q-th order AR model, we have that for any $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}$, under $\mathbb{P}_{t_0}$ almost surely

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t}\log\frac{p_1(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}\\
&= \lim_{t\to\infty} \frac{1}{t}\bigg(\log\frac{p_1(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})} + \cdots\\
&\quad + \log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}\bigg)\\
&= \lim_{t\to\infty} \frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}\\
&= \lim_{t\to\infty} \frac{1}{t}\log\frac{p_{t_0}(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t)}\\
&= \lim_{t\to\infty} \frac{1}{t}\log\frac{p_{t_0}\big((\boldsymbol{y}_{t_0,t_0+q-1}),\cdots,(\boldsymbol{y}_{t_0+q(\lfloor\frac{t}{q}\rfloor-1),t_0+q\lfloor\frac{t}{q}\rfloor-1})\big)}{p_\infty\big((\boldsymbol{y}_{t_0,t_0+q-1}),\cdots,(\boldsymbol{y}_{t_0+q(\lfloor\frac{t}{q}\rfloor-1),t_0+q\lfloor\frac{t}{q}\rfloor-1})\big)}\\
&= \widetilde{\mathcal{K}}.
\end{aligned}\tag{39}$$

where the third equality is due to the fact that the convergence of $\frac{1}{t}\log\frac{p_1(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}{p_\infty(\boldsymbol{y}_{t_0},\cdots,\boldsymbol{y}_t|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1})}$ does not depend on the sample trajectory $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}$ and the last equality is from (30) and the fact that $\{(\boldsymbol{y}_{t_0+qi},\cdots,\boldsymbol{y}_{t_0+q(i+1)-1})\}_{i=0}^\infty$ follows a first-order AR model. Following the same steps as in Theorem 3, we have that when $c = \log\gamma$, $\mathbb{E}_\infty[\tau_c^*] \ge \gamma$ and $\text{WADD}(\tau_c^*) \le \frac{\log\gamma}{\widetilde{\mathcal{K}}}(1+o(1))$ as $\gamma\to\infty$. □

It can be seen that the WADD upper bound of τ_c^* matches with the universal lower bound in Theorem 2 for the q-th order AR models. We then have the following result of asymptotic optimality.

Theorem 6: For the problem of QCD in q-th order AR models, τ_c^* (applied on its first-order equivalence) is asymptotically optimal.

Proof: By Theorem 2 and Theorem 5, we establish the asymptotic optimality of τ_c^*. □

Remark 2: In practice, when the order q of the AR model is unknown, we can first estimate q using the method in [57], and then implement our Ergodic CuSum algorithm based on the estimation.

V. DATA-DRIVEN SETTING: ONLINE GRADIENT ASCENT CUSUM

In this section, we consider the practical data-driven setting where the post-change parameters are unknown. This is motivated by practical applications such as detecting cyber-attacks in dynamic systems, where there is no prior knowledge about the disturbance signal $\boldsymbol{x}_t$. As shown in Sections III and IV, the QCD problem in a q-th order AR model can be equivalently solved using a first-order AR model. Therefore, in this section, we focus on the first-order AR model. Specifically, the post-change parameters $\boldsymbol{A}, \boldsymbol{R}_\omega$ are unknown.

The generalized likelihood ratio test (GLRT) which replaces the unknown post-change parameter with its maximum likelihood estimate (MLE) is widely used when there are unknown parameters. Specifically, at each time t, hypothesizing on the change point being k, we use samples $\boldsymbol{y}_k, \cdots, \boldsymbol{y}_t$ to compute the MLE of the unknown post-change parameters, i.e., $\hat{\theta}_{k,t} = \arg\max_\theta p_{k,\theta}(\boldsymbol{y}_k, \cdots, \boldsymbol{y}_t)$, where $p_{k,\theta}(\boldsymbol{y}_k, \cdots, \boldsymbol{y}_t)$ denotes the likelihood of $\boldsymbol{y}_k, \cdots, \boldsymbol{y}_t$ when the change point is k and post-change parameter is θ. The MLE of the hypothesized change point is then computed. Specifically, let $S_{\hat{\theta}_{k,t}}$ be the test statistic with post-change parameters $\hat{\theta}_{k,t}$, and take the maximum of $S_{\hat{\theta}_{k,t}}$ over all k, i.e, $S_t = \max_k S_{\hat{\theta}_{k,t}}$. This approach needs to store all the historical samples $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t$ and recompute $\hat{\theta}_{k,t}$ when there is a new sample, which is not efficient for memory and computation. Though in practice, a sliding-window approach can be used, but it is still hard in the non-i.i.d. setting [24]. The Markov chain Monte Carlo method [58] can also be used to estimate the unknown parameters. However, the Monte Carlo method consists of particle filtering and importance sampling, which requires to sample multiple trajectories and is not efficient in the online setting.

In this section, we apply an online gradient ascent algorithm [48], [54] to estimate the unknown parameters $\boldsymbol{A}, \boldsymbol{R}_\omega$ and plug them back to the Ergodic CuSum algorithm to design our online gradient ascent CuSum algorithm (OGA-CuSum). The OGA-CuSum can be updated using only the most recent sample and in a recursive way, and thus is much more memory and computationally efficient.

Recall the definition of $h(\boldsymbol{\mu}, \boldsymbol{y})$ in (24). Define $h_t(\boldsymbol{\mu}, \boldsymbol{y})$ by replacing $\boldsymbol{\Sigma}^*$ with $\boldsymbol{\Sigma}_t$ in $h(\boldsymbol{\mu}, \boldsymbol{y})$. From Appendix E, letting $\boldsymbol{y} = \boldsymbol{y}_t$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_t$, we have that $h_t(\boldsymbol{\mu}_t, \boldsymbol{y}_t) =$

$p_{t_0}(\boldsymbol{y}_t | \boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})$. Given the initial $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$, it then follows that

$$\log L_t = \sum_{i=1}^t h_i(\boldsymbol{\mu}_i, \boldsymbol{y}_i) - \log p_\infty(\boldsymbol{y}_i). \qquad (40)$$

Since $\boldsymbol{A}, \boldsymbol{R}_\omega$ are unknown, $h_t(\boldsymbol{\mu}_t, \boldsymbol{y}_t)$ cannot be computed. We then propose a one-step update rule to efficiently estimate $\boldsymbol{A}, \boldsymbol{R}_\omega$ and further estimate $h_t(\boldsymbol{\mu}_t, \boldsymbol{y}_t)$ at each time t.

Denote by $\widehat{\boldsymbol{A}}_t, \widehat{\boldsymbol{R}}_{\omega,t}$ the estimate of $\boldsymbol{A}, \boldsymbol{R}_\omega$ at time t. Denote by $\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{t+1}$ the parameters of the forward variable. From (15), we have that $\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{t+1}$ can be updated recursively using $\widehat{\boldsymbol{A}}_t, \widehat{\boldsymbol{R}}_{\omega,t}$, i.e.,

$$\widehat{\boldsymbol{\Sigma}}_{t+1} = (\widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Sigma}}_t \widehat{\boldsymbol{A}}_t^\top + \widehat{\boldsymbol{R}}_{\omega,t})(\widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Sigma}}_t \widehat{\boldsymbol{A}}_t^\top + \widehat{\boldsymbol{R}}_{\omega,t} + \boldsymbol{I})^{-1},$$
$$\widehat{\boldsymbol{\mu}}_{t+1} = (\widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Sigma}}_t \widehat{\boldsymbol{A}}_t^\top + \widehat{\boldsymbol{R}}_{\omega,t} + \boldsymbol{I})^{-1} \widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\mu}}_t$$
$$+ (\widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Sigma}}_t \widehat{\boldsymbol{A}}_t^\top + \widehat{\boldsymbol{R}}_{\omega,t})(\widehat{\boldsymbol{A}}_t \widehat{\boldsymbol{\Sigma}}_t \widehat{\boldsymbol{A}}_t^\top + \widehat{\boldsymbol{R}}_{\omega,t} + \boldsymbol{I})^{-1} \boldsymbol{y}_{t+1}.$$
$$(41)$$

Define $\widehat{h}_{t+1}(\widehat{\boldsymbol{\mu}}_{t+1}, \boldsymbol{y}_{t+1})$ by replacing $\boldsymbol{A}, \boldsymbol{R}_\omega, \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}$ in $h_{t+1}(\boldsymbol{\mu}_{t+1}, \boldsymbol{y}_{t+1})$ with $\widehat{\boldsymbol{A}}_t, \widehat{\boldsymbol{R}}_{\omega,t}, \widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{t+1}$. We iteratively update the estimate of the parameters when there is a new observation based on the maximum likelihood principle using gradient ascent. We note that $h_t(\boldsymbol{\mu}_t, \boldsymbol{y}_t) - \log p_\infty(\boldsymbol{y}_t)$ is the log-likelihood ratio of the observation $\boldsymbol{y}_t$ at time t, i.e., $h_t(\boldsymbol{\mu}_t, \boldsymbol{y}_t) - \log p_\infty(\boldsymbol{y}_t) = \frac{p_1(\boldsymbol{y}_t | \boldsymbol{y}_1, \cdots, \boldsymbol{y}_{t-1})}{p_\infty(\boldsymbol{y}_t)}$. Therefore, based on the maximum likelihood principle, the online gradient ascent estimator is updated as follows

$$\widehat{\boldsymbol{A}}_t = \widehat{\boldsymbol{A}}_{t-1} + \beta \nabla_{\boldsymbol{A}} \widehat{h}_t(\widehat{\boldsymbol{\mu}}_t, \boldsymbol{y}_t),$$
$$\widehat{\boldsymbol{R}}_{\omega,t} = \mathrm{Proj}(\widehat{\boldsymbol{R}}_{\omega,t-1} + \beta \nabla_{\boldsymbol{R}_\omega} \widehat{h}_t(\widehat{\boldsymbol{\mu}}_t, \boldsymbol{y}_t)), \qquad (42)$$

where β is a pre-specified step-size and $\mathrm{Proj}(\boldsymbol{X})$ is an operator that projects a matrix $\boldsymbol{X}$ to the set of positive definite matrices to guarantee that $\widehat{\boldsymbol{R}}_{\omega,t}$ is positive definite. Define the eigenvalue decomposition $\boldsymbol{X} = \sum_{i=1}^n \lambda_i v_i v_i^\top$, where λ_i is the eigenvalue of $\boldsymbol{X}$ and v_i is its corresponding eigenvector. We define $\mathrm{Proj}(\boldsymbol{X}) = \sum_{i=1}^n \max\{\lambda_i, \epsilon\} v_i v_i^\top$ [59], where $\epsilon > 0$ is chosen to guarantee that the eigenvalue of $\mathrm{Proj}(\boldsymbol{X})$ is at least ϵ. With $\widehat{\boldsymbol{A}}_t, \widehat{\boldsymbol{R}}_{\omega,t}$, we can update $\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{t+1}$ according to (41). We further plug them back to (37) and (40) to compute $\widehat{S}_{t+1}$, which serves as the estimate of S_{t+1}. When $\widehat{S}_{t+1} < 0$, we claim that the change hasn't occurred. Therefore, we ignore the previous samples and reset our parameters $\widehat{\boldsymbol{A}}_{t+1}, \widehat{\boldsymbol{R}}_{\omega,t+1}$. The OGA-CuSum algorithm is then defined as follows

$$\tau_{\mathrm{OGA}} = \inf\{t : \widehat{S}_t \geq c\}. \qquad (43)$$

We summarize the algorithm in Algorithm 1.

In general, it is hard to obtain theoretical optimality performance guarantees for data-driven approaches. Nevertheless, we provide a lower bound on the ARL for our OGA-CuSum algorithm, so that a threshold can be chosen analytically to control the false alarm in practice. We also provide simulation results to demonstrate the good performance of our algorithm.

In the following theorem, we present a lower bound on the ARL for our OGA-CuSum algorithm.

Theorem 7: Let $c = \log \gamma$ in Algorithm 1, then $\mathbb{E}_\infty[\tau_{\mathrm{OGA}}] \geq \gamma$.

Proof: The proof can be found in Appendix G. □

Algorithm 1 Online gradient ascent CuSum algorithm

Require: Initial parameters $c, \beta, \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\Sigma}}_0, \widehat{\boldsymbol{A}}_0, \widehat{\boldsymbol{R}}_{\omega,0}, \ t \leftarrow 0,$
$\widehat{L}_0 \leftarrow 1, \widehat{S}_0 \leftarrow 0$
  while $\widehat{S}_t < c$ **do**
    Take a new observation $\boldsymbol{y}_{t+1}$
    Update $\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{t+1}$ according to (41)
    $\log \widehat{L}_{t+1} \leftarrow \log \widehat{L}_t + \widehat{h}_{t+1}(\widehat{\boldsymbol{\mu}}_{t+1}, \boldsymbol{y}_{t+1}) - \log p_\infty(\boldsymbol{y}_{t+1})$
    $\widehat{S}_{t+1} \leftarrow \widehat{S}_t + \log \widehat{L}_{t+1} - \log \widehat{L}_t$
    if $\widehat{S}_{t+1} < 0$ **then**
      $\widehat{\boldsymbol{A}}_{t+1} \leftarrow \widehat{\boldsymbol{A}}_0$
      $\widehat{\boldsymbol{R}}_{\omega,t+1} \leftarrow \widehat{\boldsymbol{R}}_{\omega,0}$
      $\widehat{S}_{t+1} \leftarrow 0$
      $\widehat{L}_{t+1} \leftarrow 1$
    else
      $\widehat{\boldsymbol{A}}_{t+1} = \widehat{\boldsymbol{A}}_t + \beta \nabla_{\boldsymbol{A}} \widehat{h}_{t+1}(\widehat{\boldsymbol{\mu}}_{t+1}, \boldsymbol{y}_{t+1})$
      $\widehat{\boldsymbol{R}}_{\omega,t+1} = \text{Proj}\big(\widehat{\boldsymbol{R}}_{\omega,t} + \beta \nabla_{\boldsymbol{R}_\omega} \widehat{h}_{t+1}(\widehat{\boldsymbol{\mu}}_{t+1}, \boldsymbol{y}_{t+1})\big)$
    end if
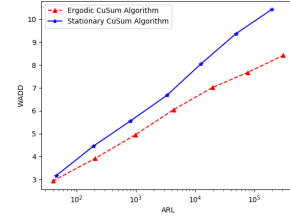 $t \leftarrow t+1$
  end while

Fig. 1. Comparison of the two algorithms: case 1.

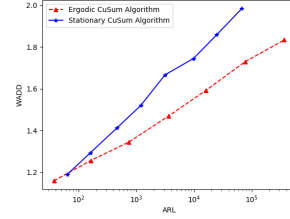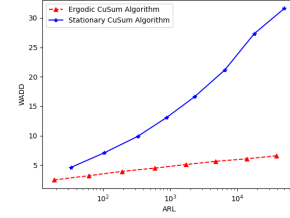

Fig. 2. Comparison of the two algorithms: case 2.

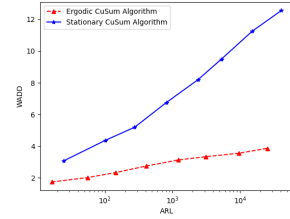Fig. 3. Comparison of the two algorithms: case 3.

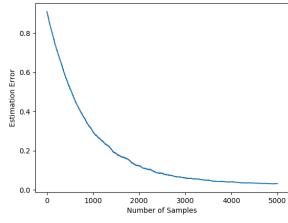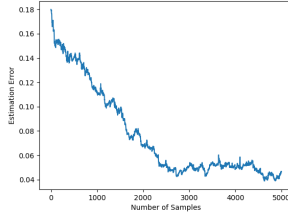Fig. 4. Comparison of the two algorithms: case 4.

VI. SIMULATION RESULTS

In this section, we provide some numerical results to demonstrate the performance of our Ergodic CuSum algorithm and OGA-CuSum algorithm.

A. Model-Based Setting

In this section, we consider the setting where the parameters of the disturbance signal are known. We compare the Ergodic CuSum algorithm τ_c^* with a stationary CuSum algorithm designed using the stationary distribution of $\boldsymbol{y}$. The stationary CuSum algorithm uses the CuSum algorithm that detects a change from p_∞ to the stationary distribution π:

$$\tau_s = \inf\left\{ t : \max_{1 \leq k \leq t} \sum_{i=k}^{t} \log \frac{\pi(\boldsymbol{y}_i)}{p_\infty(\boldsymbol{y}_i)} \geq c \right\}. \quad (44)$$

We consider four different cases. For the first case, we set $\boldsymbol{A} = [[0.7, 0.4], [0.2, 0.6]]$ and $\boldsymbol{R}_\omega = [[1, 0.5], [0.5, 1]]$. For the second case, we let $\boldsymbol{A}$ be a 10×10 matrix and $\boldsymbol{R}_\omega = \boldsymbol{I}$. For the third case, we consider a 2-th order AR model. We set $\boldsymbol{A}_1 = [[0.4, 0.3], [0.2, 0.1]], \boldsymbol{A}_2 = [[0.3, 0.2], [0.1, 0.2]]$ and $\boldsymbol{R}_\omega = \boldsymbol{I}$. For the fourth case, we consider a 5-th order AR model. We set $\boldsymbol{A}_1 = [[0.2, 0.1], [0.1, 0.2]], \boldsymbol{A}_2 = [[0.2, 0.1], [0.1, 0.2]], \boldsymbol{A}_3 = [[0.1, 0], [0, 0.1]], \boldsymbol{A}_4 = [[0.1, 0], [0, 0.1]], \boldsymbol{A}_5 = [[0.1, 0], [0, 0.1]]$ and $\boldsymbol{R}_\omega = \boldsymbol{I}$. Note that our Ergodic CuSum algorithm can be easily generalized to the q-th order AR model by reformulating the q-th order AR model to a first-order AR model. Since our test statistic S_t has initial value 0 and remains non-negative, the delay is the largest when the change happens at $t_0 = 0$. To evaluate the WADD, we let the change point $t_0 = 0$. In Fig. 1, Fig. 2, Fig. 3 and Fig. 4, we plot the WADD as a function of ARL. It can be seen from Fig. 1, Fig. 2, Fig. 3 and Fig. 4 that with the same constraint on the ARL, our Ergodic CuSum algorithm has a lower detection delay. Therefore, our Ergodic CuSum algorithm outperforms the stationary CuSum algorithm under both the low-dimensional setting and the high-dimensional setting. This further implies that for detecting changes in the AR model, simply applying the stationary distribution and ignoring the dependence among samples may not perform well. Moreover, the relationship between WADD and log of the WARL is linear, which validates our theoretical analysis.

B. Data-Driven Setting

In this section, we focus on the data-driven setting. We first plot the estimation errors of $\widehat{\boldsymbol{A}}_t, \widehat{\boldsymbol{R}}_{\omega,t}$ as functions of number of samples under the pre- and post-change distribution, where the estimation errors are defined as the Frobenius norm of $\widehat{\boldsymbol{A}}_t - \boldsymbol{A}$ and $\widehat{\boldsymbol{R}}_{\omega,t} - \boldsymbol{R}_\omega$. Under the post-change distribution, we set $\boldsymbol{A} = [[0.7, 0.4], [0.2, 0.6]]$ and $\boldsymbol{R}_\omega = [[1, 0.5], [0.5, 1]]$. We note that the pre-change distribution is equivalent to the AR model with $\boldsymbol{A} = \boldsymbol{0}$ and $\boldsymbol{R}_\omega = \boldsymbol{0}$. It can be seen from Fig. 5, Fig. 6, Fig. 7 and Fig. 8 that the estimated

Fig. 5. Convergence of $\widehat{\boldsymbol{A}}_t$ under the pre-change distribution.

Fig. 6. Convergence of $\widehat{\boldsymbol{A}}_t$ under the post-change distribution.

Fig. 7. Convergence of $\widehat{\boldsymbol{R}}_{\omega,t}$ under the pre-change distribution.

Fig. 8. Convergence of $\widehat{\boldsymbol{R}}_{\omega,t}$ under the post-change distribution.

Fig. 9. Comparison of the three algorithms: case 1.

Fig. 10. Comparison of the three algorithms: case 4.

Fig. 11. Comparison of the three algorithms: case 3.

Fig. 12. Comparison of the three algorithms: case 4.

parameters $\widehat{\boldsymbol{A}}_t$, $\widehat{\boldsymbol{R}}_{\omega,t}$ converge to the true parameters under the pre- and post-change distribution, which demonstrates that our OGA-CuSum scheme provides reliable estimates for the true parameters of the disturbance signal.

We then compare our OGA-CuSum algorithm with an existing approach in [15]. In [15], a GLRT based data-driven algorithm was proposed to detect disturbance signals in the AR model. We also plot the curve of the Ergodic CuSum algorithm under the model-based setting, which serves as a lower bound for the data-driven setting. We use the same parameters as in Section VI-A and plot the WADD as a function of ARL. From Fig. 9, Fig. 10, Fig. 11 and Fig. 12, it can be seen that our OGA-CuSum algorithm outperforms the GLRT based algorithm in [15]. Moreover, the performance of the OGA-CuSum algorithm is close to the performance of the Ergodic CuSum algorithm. Therefore, our OGA-CuSum algorithm has a small performance loss compared with

the asymptotically optimal algorithm under the model-based setting.

VII. CONCLUSION

In this paper, we studied the QCD problem in AR models under the model-based setting and the data-driven setting. For the model-based setting, we proposed a novel Ergodic CuSum algorithm. By introducing the forward variable for general state HMMs, our Ergodic CuSum algorithm can be implemented efficiently. We further showed that our algorithm is asymptotically optimal under the Lorden's criterion [41] based on the ergodic theorem [52], [53]. For the data-driven setting, we proposed an OGA-CuSum algorithm, which replaces the unknown parameters in the Ergodic CuSum algorithm with their estimates based on the online convex optimization. We provided simulation results to demonstrate the performance of our algorithm. Our approaches provide useful insights for general detection problems for AR models

and general state HMMs. In the future, it is of great interest to investigate the setting where the observation noise and the innovation noise are beyond the Gaussian model, for which the Monte Carlo method [58], [60] based CuSum algorithm can be designed.

APPENDIX A
CONVERT q-TH ORDER AR MODEL TO FIRST-ORDER AR MODEL

Consider the case with $q = 2$. Let $\tilde{\boldsymbol{x}}_t^{(2)} = [\boldsymbol{x}_{2t-1}^\top, \boldsymbol{x}_{2t}^\top]^\top$. From (3), we have that

$$\tilde{\boldsymbol{x}}_t^{(2)} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}_{t-1}^{(2)} + \tilde{\boldsymbol{\omega}}_t^{(2)}, \tag{45}$$

where $\tilde{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{A}_2 & \boldsymbol{A}_1 \\ \boldsymbol{A}_1\boldsymbol{A}_2 & \boldsymbol{A}_1\boldsymbol{A}_1 + \boldsymbol{A}_2 \end{pmatrix}$ and $\tilde{\boldsymbol{\omega}}_t^{(2)} = \begin{pmatrix} \boldsymbol{\omega}_{2t-1} \\ \boldsymbol{A}_1\boldsymbol{\omega}_{2t-1} + \boldsymbol{\omega}_{2t} \end{pmatrix}$ is Gaussian distributed. Therefore, $\tilde{\boldsymbol{x}}_t^{(2)}$ follows a first-order AR model. Let $\tilde{\boldsymbol{y}}_t^{(2)} = [\boldsymbol{y}_{2t-1}^\top, \boldsymbol{y}_{2t}^\top]^\top$. We have that $\tilde{\boldsymbol{y}}_t^{(2)} = \tilde{\boldsymbol{x}}_t^{(2)} + \tilde{\boldsymbol{\nu}}_t^{(2)}$, where $\tilde{\boldsymbol{\nu}}_t^{(2)} = \begin{pmatrix} \boldsymbol{\nu}_{2t-1} \\ \boldsymbol{\nu}_{2t} \end{pmatrix}$.

Following the same steps, for a q-th order AR model, let $\tilde{\boldsymbol{x}}_t^{(q)} = [\boldsymbol{x}_{q(t-1)+1}^\top, \boldsymbol{x}_{q(t-1)+2}^\top, \cdots, \boldsymbol{x}_{qt}^\top]^\top$ and $\tilde{\boldsymbol{y}}_t^{(q)} = [\boldsymbol{y}_{q(t-1)+1}^\top, \boldsymbol{y}_{q(t-1)+2}^\top, \cdots, \boldsymbol{y}_{qt}^\top]^\top$. We have that

$$\begin{aligned} \tilde{\boldsymbol{x}}_t^{(q)} &= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}_{(t-1)}^{(q)} + \tilde{\boldsymbol{\omega}}_t^{(q)}, \\ \tilde{\boldsymbol{y}}_t^{(q)} &= \tilde{\boldsymbol{x}}_t^{(q)} + \tilde{\boldsymbol{\nu}}_t^{(q)}, \end{aligned} \tag{46}$$

where $\tilde{\boldsymbol{\omega}}_t^{(q)}$ and $\tilde{\boldsymbol{\nu}}_t^{(q)}$ are the innovative noise and measurement noise respectively for the q-th order AR model. We denote the covariance of $\tilde{\boldsymbol{\omega}}_t^{(q)}$ by $\tilde{\boldsymbol{R}}_{\boldsymbol{\omega}}$. Here, we omit the expression of $\tilde{\boldsymbol{A}}$ since it is cumbersome but it can be computed directly. It can be easily verified that the covariance matrix of $\tilde{\boldsymbol{\nu}}_t^{(q)}$ is $\boldsymbol{I}$. Therefore, $\{\tilde{\boldsymbol{x}}_t^{(q)}, \tilde{\boldsymbol{y}}_t^{(q)}\}_{t=1}^\infty$ is a first-order AR model.

APPENDIX B
PROOF OF LEMMA 2

Proof: Let $\|\cdot\|$ denote the operator norm of a matrix and the Euclidean norm of a vector. To show that $\boldsymbol{\Sigma}_t$ converges, it suffices to show that for any $\epsilon > 0$, there exists an integer T such that for any $s, t > T$, $\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_t\| < \epsilon$. Without loss of generality, we assume that $s > t$.

We first note that

$$\begin{aligned} \boldsymbol{\Sigma}_t &= \left((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1}, \\ \boldsymbol{\Sigma}_s &= \left((\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1}. \end{aligned} \tag{47}$$

We then have that

$$\begin{aligned} &\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_t \\ &= \boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_s^{-1})\boldsymbol{\Sigma}_s \\ &= \left((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1}\left((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\right. \\ &\quad \left. - (\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\right)\left((\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1} \\ &= \left((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1}(\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} \\ &\quad \left((\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}}) - (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})\right) \\ &\quad (\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\left((\boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} + \boldsymbol{I}\right)^{-1} \end{aligned}$$

$$\begin{aligned} &= (\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\boldsymbol{A}(\boldsymbol{\Sigma}_{s-1} - \boldsymbol{\Sigma}_{t-1})\boldsymbol{A}^\top \\ &\quad (\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}. \end{aligned} \tag{48}$$

From the update rule of $\boldsymbol{\Sigma}_t$ in (47), $\boldsymbol{\Sigma}_t$ is positive definite for every $t \geq t_0$. Therefore, $\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}}$ is positive definite. It then allows the eigendecomposition and can be factorized as $\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q}$ is the square $K \times K$ matrix whose ith column is the eigenvector q_i of $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix whose ith diagonal elements are the corresponding eigenvalues λ_i. We then have that

$$\begin{aligned} (\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1} &= \left(\boldsymbol{Q}(\boldsymbol{\Lambda} + \boldsymbol{I})\boldsymbol{Q}^{-1}\right)^{-1} \\ &= \boldsymbol{Q}(\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{Q}^{-1}. \end{aligned} \tag{49}$$

Since $\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}}$ is positive definite for any $t > 0$, we have that $\lambda_i > 0$ for $i = 1, \cdots, K$. Therefore, all the diagonal elements of $\boldsymbol{\Lambda} + \boldsymbol{I}$ are strictly larger than 1. Therefore, the diagonal elements of $(\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}$ are strictly less than 1. From the definition of operator norm, we have that there exists a $0 < \delta < 1$ such that $\|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\| < (1 - \delta)$ for any $t > 0$.

It then follows that

$$\begin{aligned} &\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_t\| \\ &= \|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\boldsymbol{A}(\boldsymbol{\Sigma}_{s-1} - \boldsymbol{\Sigma}_{t-1})\boldsymbol{A}^\top \\ &\quad (\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\| \\ &\leq \|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\|\|\boldsymbol{A}\|\|\boldsymbol{\Sigma}_{s-1} - \boldsymbol{\Sigma}_{t-1}\|\|\boldsymbol{A}^\top\| \\ &\quad \|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\| \\ &< (1 - \delta)^2\|\boldsymbol{\Sigma}_{s-1} - \boldsymbol{\Sigma}_{t-1}\|, \end{aligned} \tag{50}$$

where the first inequality is from the submultiplicative of operator norm and the second inequality is due to the facts that $\|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\| < 1 - \delta$, $\|(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{\Sigma}_{s-1}\boldsymbol{A}^\top + \boldsymbol{R}_{\boldsymbol{\omega}})^{-1}\| < 1 - \delta$ and the assumption $\|\boldsymbol{A}\| < 1$. By applying (50) recursively, we have that

$$\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_t\| < (1 - \delta)^{2(t-t_0)}\|\boldsymbol{\Sigma}_{s-t_0+t} - \boldsymbol{\Sigma}_{t_0}\|. \tag{51}$$

Moreover, from the update rule of $\boldsymbol{\Sigma}_t$, it can be easily verified that $\|\boldsymbol{\Sigma}_{s-t_0+t} - \boldsymbol{\Sigma}_1\|$ is bounded. Therefore, for any $\epsilon > 0$, there exists an integer T such that for any $s, t > T$, $\|\boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_t\| < \epsilon$. This completes the proof. $\square$

APPENDIX C
PROOF OF LEMMA 3

Proof: Let $\pi(\boldsymbol{x}_t)$ be a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We will show that $\pi(\boldsymbol{x}_t)$ is a stationary distribution of $\{\boldsymbol{x}_t\}_{t=t_0}^\infty$ if $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \sum_{i=0}^\infty (\boldsymbol{A}^\top)^i \boldsymbol{R}_{\boldsymbol{\omega}} \boldsymbol{A}^i$. It can be easily proved that $\boldsymbol{\Sigma}$ exists since $\|\boldsymbol{A}\| < 1$, where $\|\cdot\|$ denotes the operator norm of a matrix.

From the definition of stationary distribution, $\pi(\boldsymbol{x}_t)$ should satisfy that

$$\int \pi(\boldsymbol{x}_{t-1})f(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1} = \pi(\boldsymbol{x}_t). \tag{52}$$

We have that

$$\int \pi(\boldsymbol{x}_{t-1})f(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1}$$

$$= \int \pi(\boldsymbol{x}_{t-1}) \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{R}_\omega)}}$$
$$\exp\left(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{x}_{t-1})^\top \boldsymbol{R}_\omega^{-1}(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{x}_{t-1})\right) d\boldsymbol{x}_{t-1}$$
$$= \int \pi(\boldsymbol{x}_{t-1}) \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{R}_\omega)}} \exp\Big($$
$$-\frac{1}{2}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{x}_{t-1})^\top \boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) d\boldsymbol{x}_{t-1}$$
$$= \int \pi(\boldsymbol{x}_{t-1}) \frac{\sqrt{\det((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1})}}{\sqrt{\det(\boldsymbol{R}_\omega)}\sqrt{(2\pi)^K \det((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A}))^{-1}}} \exp$$
$$\Big(-\frac{1}{2}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{x}_{t-1})^\top \boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) d\boldsymbol{x}_{t-1}$$
$$\overset{(a)}{=} \frac{\sqrt{\det((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1})}}{\sqrt{\det(\boldsymbol{R}_\omega)}\sqrt{(2\pi)^K \det((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma})}} \exp\Big($$
$$-\frac{1}{2}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{\mu})^\top \big((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma}\big)^{-1}(\boldsymbol{A}^{-1}\boldsymbol{x}_t - \boldsymbol{\mu})\Big)$$
$$= \frac{\sqrt{\det((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1}) \det\big(\boldsymbol{A}((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma})\boldsymbol{A}^\top\big)}}{\sqrt{\det(\boldsymbol{R}_\omega) \det\big((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma}\big)}}$$
$$\frac{1}{\sqrt{(2\pi)^K \det\big(\boldsymbol{A}((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma})\boldsymbol{A}^\top\big)}} \exp\Big(-\frac{1}{2}$$
$$(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{\mu})^\top \big(\boldsymbol{A}((\boldsymbol{A}^\top \boldsymbol{R}_\omega^{-1}\boldsymbol{A})^{-1} + \boldsymbol{\Sigma})\boldsymbol{A}^\top\big)^{-1}(\boldsymbol{x}_t - \boldsymbol{A}\boldsymbol{\mu})\Big),$$
$$\tag{53}$$

where (a) is from the convolution of two Gaussian functions. Let $\boldsymbol{\Sigma} = \sum_{i=0}^\infty (\boldsymbol{A}^\top)^i \boldsymbol{R}_\omega \boldsymbol{A}^i$. We have that $\boldsymbol{\Sigma} = \boldsymbol{R}_\omega + \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$. If we choose $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = \sum_{i=0}^\infty (\boldsymbol{A}^\top)^i \boldsymbol{R}_\omega \boldsymbol{A}^i$, it can then be easily verified that (52) holds. Therefore, $\pi(\boldsymbol{x}_t)$ is a stationary distribution of $\{\boldsymbol{x}_t\}_{t=t_0}^\infty$. It then follows that $\{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=t_0}^\infty$ has a stationary distribution $\pi(\boldsymbol{x}_t)g(\boldsymbol{y}_t|\boldsymbol{x}_t)$.

Let $P^1(\{\boldsymbol{x}, \boldsymbol{y}\}, E)$ denote the probability of reaching a measurable set E from state $\{\boldsymbol{x}, \boldsymbol{y}\}$ in one step. We have that for any $\{\boldsymbol{x}, \boldsymbol{y}\}$ and $E \in \mathbb{R}^{2K}$ such that $\pi(E) > 0$,

$$P^1(\{\boldsymbol{x}, \boldsymbol{y}\}, E) = \int_E f(\boldsymbol{x}'|\boldsymbol{x})g(\boldsymbol{y}'|\boldsymbol{x}')d\boldsymbol{x}'\boldsymbol{y}' > 0, \tag{54}$$

where the inequality is due to the fact that Gaussian density functions are positive. From the definition of irreducible [53] Markov chain, we have that $\{\boldsymbol{x}_t, \boldsymbol{y}_t\}_{t=t_0}^\infty$ is π-irreducible. $\square$

APPENDIX D
PROOF OF LEMMA 4

Proof: To prove Lemma 4, we will first show that $\{\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain. We will then show that the stationary distribution for $\boldsymbol{\mu}_t^*$ exists by finding the limiting distribution of $\boldsymbol{\mu}_t^*$. We then construct a new Markov chain using $\boldsymbol{y}_t, \boldsymbol{\mu}_t^*$ and show that its stationary distribution exists and is unique.

Step 1. We first show that $\{\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain. From the update rule of $\boldsymbol{\mu}_t^*$, we have that for $t \geq t_0$,

$$p_{t_0}(\boldsymbol{\mu}_t^*|\boldsymbol{x}_{t_0}, \cdots, \boldsymbol{x}_t, \boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t, \boldsymbol{\mu}_{t_0}^*, \cdots, \boldsymbol{\mu}_{t-1}^*)$$
$$= p_{t_0}(\boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*, \boldsymbol{y}_t). \tag{55}$$

We then have that

$$p_{t_0}(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{x}_{t_0}, \cdots, \boldsymbol{x}_{t-1}, \boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*, \cdots, \boldsymbol{\mu}_{t-1}^*)$$
$$= p_{t_0}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p_{t_0}(\boldsymbol{y}_t|\boldsymbol{x}_t)p_{t_0}(\boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*, \boldsymbol{y}_t)$$
$$= p_{t_0}(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t-1}^*). \tag{56}$$

Therefore, $\{\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain.

Step 2. We then show that the stationary distribution for $\boldsymbol{\mu}_t^*$ exists by finding the limiting distribution of $\boldsymbol{\mu}_t^*$. Let the initial state be $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$. From (22), we have that

$$\boldsymbol{\mu}_t^* = \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\big)^{t-t_0}\boldsymbol{\mu}_{t_0}^*$$
$$+ \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\big)^{t-t_0-1}\boldsymbol{\Sigma}^*\boldsymbol{y}_{t_0+1} + \cdots$$
$$+ (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{y}_{t-1} + \boldsymbol{\Sigma}^*\boldsymbol{y}_t. \tag{57}$$

Note that $\boldsymbol{\mu}_t^*$ is the sum of Gaussian random variables. If $\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t$ are jointly Gaussian distributed, then $\boldsymbol{\mu}_t^*$ is a Gaussian random variable. Since the Gaussian density function is continuous in its mean and covariance, it then suffices to show that the limiting mean and limiting covariance matrix of $\boldsymbol{\mu}_t^*$ exist and are independent of the initial state.

Given the initial state $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$, we have that

$$p_{t_0}(\boldsymbol{y}_{t_0+1}|\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*)$$
$$= \int p_{t_0}(\boldsymbol{x}_{t_0+1}|\boldsymbol{x}_{t_0})p_{t_0}(\boldsymbol{y}_{t_0+1}|\boldsymbol{x}_{t_0+1})d\boldsymbol{x}_{t_0+1}$$
$$= \int \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{R}_\omega)}}$$
$$\exp\left(-\frac{1}{2}(\boldsymbol{x}_{t_0+1} - \boldsymbol{A}\boldsymbol{x}_{t_0})^\top \boldsymbol{R}_\omega^{-1}(\boldsymbol{x}_{t_0+1} - \boldsymbol{A}\boldsymbol{x}_{t_0})\right)\frac{1}{\sqrt{(2\pi)^K}}$$
$$\exp(-\frac{1}{2}(\boldsymbol{y}_{t_0+1} - \boldsymbol{x}_{t_0+1})^\top(\boldsymbol{y}_{t_0+1} - \boldsymbol{x}_{t_0+1}))d\boldsymbol{x}_{t_0+1}$$
$$= \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{R}_\omega + \boldsymbol{I})}} \exp\Big(-\frac{1}{2}(\boldsymbol{y}_{t_0+1} - \boldsymbol{A}\boldsymbol{x}_{t_0})^\top$$
$$(\boldsymbol{I} + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{y}_{t_0+1} - \boldsymbol{A}\boldsymbol{x}_{t_0})\Big). \tag{58}$$

Therefore, conditioning on $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$, $\boldsymbol{y}_{t_0+1}$ is Gaussian distributed with mean $\boldsymbol{A}\boldsymbol{x}_{t_0}$ and covariance matrix $\boldsymbol{I} + \boldsymbol{R}_\omega$. Similarly, we can show that for any $t > t_0$, conditioning on $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$, $\boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{A}^{(t-t_0)}\boldsymbol{x}_{t_0}, \boldsymbol{I} + \boldsymbol{R}_\omega + \boldsymbol{A}^\top \boldsymbol{R}_\omega \boldsymbol{A} + \cdots + (\boldsymbol{A}^\top)^{(t-t_0-1)}\boldsymbol{R}_\omega \boldsymbol{A}^{t-t_0-1})$. Moreover, since $\boldsymbol{x}_t$ is Gaussian distributed and $\boldsymbol{x}_{t+1}|\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{R}_\omega)$, we have that $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}$ are jointly Gaussian distributed. Similarly, since $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}$ are jointly Gaussian and $\boldsymbol{y}_t|(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \sim \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{I})$, we have that $\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{x}_{t+1}$ are jointly Gaussian distributed. Following the same idea, we can show that $\boldsymbol{x}_{t_0}, \boldsymbol{x}_{t_0+1}, \cdots, \boldsymbol{x}_t, \boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1}, \cdots, \boldsymbol{y}_t$ are jointly Gaussian distributed and thus $\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1}, \cdots, \boldsymbol{y}_t$ are jointly Gaussian distributed. Therefore, $\boldsymbol{\mu}_t^*$ follows a Gaussian distribution.

To find the limiting distribution of $\boldsymbol{\mu}_t^*$, we first consider the limiting expectation of $\boldsymbol{\mu}_t^*$. We have that

$$\mathbb{E}_{t_0}[\boldsymbol{\mu}_t^*|\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*]$$
$$= \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\big)^{t-t_0}\boldsymbol{\mu}_{t_0}^*$$
$$+ \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\big)^{t-t_0-1}\boldsymbol{\Sigma}^*\boldsymbol{A}\boldsymbol{x}_{t_0} + \cdots$$
$$+ (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^{t-t_0-1}\boldsymbol{x}_{t_0} + \boldsymbol{\Sigma}^*\boldsymbol{A}^{t-t_0}\boldsymbol{x}_{t_0}.$$

Since there exists $\delta > 0$ such that $\|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\| < 1 - \delta$ and $\|A\| < 1 - \delta$, we have that

$$\|\mathbb{E}_{t_0}[\mu_t^* | x_{t_0}, y_{t_0}, \mu_{t_0}^*]\|$$

$$\overset{(a)}{\leq} \|((A\Sigma^*A^\top + R_\omega + I)^{-1}A)^{t-t_0}\mu_{t_0}^*\|$$
$$+ \|((A\Sigma^*A^\top + R_\omega + I)^{-1}A)^{t-t_0-1}\Sigma^*Ax_{t_0}\| + \cdots$$
$$+ \|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\Sigma^*A^{t-t_0-1}x_{t_0}\|$$
$$+ \|\Sigma^*A^{t-t_0}x_{t_0}\|$$

$$\overset{(b)}{\leq} \|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\|^{t-t_0}\|\mu_{t_0}^*\|$$
$$+ \|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\|^{t-t_0-1}\|\Sigma^*\|\|A\|\|x_{t_0}\| + \cdots$$
$$+ \|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\|\|\Sigma^*\|\|A\|^{t-t_0-1}\|x_{t_0}\|$$
$$+ \|\Sigma^*\|\|A\|^{t-t_0}\|x_{t_0}\|$$
$$\leq (1-\delta)^{t-t_0}\|\mu_{t_0}^*\| + (t-t_0)(1-\delta)^{t-t_0}\|\Sigma^*\|\|x_{t_0}\|,$$

where (a) is from the triangle inequality of norms and (b) is due to the submultiplicative of the operator norm. We then have that $\lim_{t\to\infty}\|\mathbb{E}_{t_0}[\mu_t^*|x_{t_0}, y_{t_0}, \mu_{t_0}^*]\| = 0$. Therefore, we have that $\lim_{t\to\infty}\mathbb{E}_{t_0}[\mu_t^*|x_{t_0}, y_{t_0}, \mu_{t_0}^*]$ exists and is independent of the initial state $\{x_{t_0}, y_{t_0}, \mu_{t_0}^*\}$.

We then consider the covariance matrix of μ_t^*. Given the initial state $\{x_{t_0}, y_{t_0}, \mu_{t_0}^*\}$, let

$$\Sigma_{\mu_t^*} = \mathbb{E}_{t_0}\Big[(\mu_t^* - \mathbb{E}_{t_0}[\mu_t^*|x_{t_0}, y_{t_0}, \mu_{t_0}^*])$$
$$(\mu_t^* - \mathbb{E}_{t_0}[\mu_t^*|x_{t_0}, y_{t_0}, \mu_{t_0}^*])^\top | x_{t_0}, y_{t_0}, \mu_{t_0}^*\Big],$$
$$\Sigma_{ij} = \mathbb{E}_{t_0}\Big[(y_i - \mathbb{E}_{t_0}[y_i|x_{t_0}, y_{t_0}, \mu_{t_0}^*])$$
$$(y_j - \mathbb{E}_{t_0}[y_j|x_{t_0}, y_{t_0}, \mu_{t_0}^*])^\top | x_{t_0}, y_{t_0}, \mu_{t_0}^*\Big]. \quad (59)$$

From (57) and (59), we have that

$$\Sigma_{\mu_t^*} = \sum_{i=t_0}^{t-1}\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}$$
$$\Sigma^*\Sigma_{(t-i+t_0)(t-i+t_0)}\Big(\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}\Sigma^*\Big)^\top$$
$$+ 2\sum_{i=t_0}^{t-1}\sum_{j=t_0}^{i-1}\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}$$
$$\Sigma^*\Sigma_{(t-i+t_0)(t-j+t_0)}\Big(\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{j-t_0}\Sigma^*\Big)^\top. \quad (60)$$

We first show that $\|\Sigma_{ij}\|$ is bounded for any $i, j > t_0$. When $i = j$, we have that

$$\|\Sigma_{ii}\|$$
$$= \|I + R_\omega + A^\top R_\omega A + \cdots + (A^\top)^{(i-t_0-1)}R_\omega A^{i-t_0-1}\|$$
$$\leq \|I\| + \|R_\omega\| + \cdots + \|A^\top\|^{i-t_0-1}\|R_\omega\|\|A\|^{i-t_0-1}$$
$$= \|I\| + \big(1 + \|A\|^2 + \cdots + \|A\|^{2(i-t_0)}\big)\|R_\omega\|$$
$$= \|I\| + \frac{1 - \|A\|^{2(i-t_0)}}{1 - \|A\|^2}\|R_\omega\|$$
$$\leq \|I\| + \frac{1}{1 - \|A\|^2}\|R_\omega\|, \quad (61)$$

where the last inequality is due to the fact that there exists $0 < \delta < 1$ such that $\|A\| < 1 - \delta$. When $i \neq j$, we have that

$$\|\Sigma_{ij}\| = \Big\|\mathbb{E}_{t_0}\Big[(y_i - \mathbb{E}_{t_0}[y_i|x_{t_0}, y_{t_0}, \mu_{t_0}^*])$$
$$(y_j - \mathbb{E}_{t_0}[y_j|x_{t_0}, y_{t_0}, \mu_{t_0}^*])^\top | x_{t_0}, y_{t_0}, \mu_{t_0}^*\Big]\Big\|$$
$$\leq \frac{1}{2}\big(\|\Sigma_{ii}\| + \|\Sigma_{jj}\|\big)$$
$$\leq \|I\| + \frac{1}{1 - \|A\|^2}\|R_\omega\|, \quad (62)$$

where the first inequality is due to the fact that for any two vectors $u, v \in \mathbb{R}^K$, $(u - v)(u - v)^\top$ is positive semi-definite, and thus $uu^\top + vv^\top - 2uv^\top \succeq 0$. Since the operator norm equals to the largest eigenvalue for the positive semi-definite matrix, we have that $\|uv^\top\| \leq \frac{1}{2}\|uu^\top + vv^\top\| \leq \frac{1}{2}(\|uu^\top\| + \|vv^\top\|)$.

To show that $\Sigma_{\mu_t^*}$ converges, it suffices to show that for any $\epsilon > 0$, there exists an integer $T \geq t_0$ such that for any $s, t > T$, $\|\Sigma_{\mu_s^*} - \Sigma_{\mu_t^*}\| < \epsilon$. Without loss of generality, we assume that $s > t$. We then have that

$$\|\Sigma_{\mu_s^*} - \Sigma_{\mu_t^*}\|$$
$$\leq \Big\|\sum_{i=t}^{s-1}\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}\Sigma^*$$
$$\Sigma_{(s-i+t_0)(s-i+t_0)}\Big(\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}\Sigma^*\Big)^\top\Big\|$$
$$+ 2\Big\|\sum_{i=t}^{s-1}\sum_{j=t_0}^{i-1}\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{i-t_0}\Sigma^*$$
$$\Sigma_{(s-i+t_0)(s-j+t_0)}\Big(\big((A\Sigma^*A^\top + R_\omega + I)^{-1}A\big)^{j-t_0}\Sigma^*\Big)^\top\Big\|,$$

where the inequality is from the triangle inequality of operator norm. To simplify the notation, let $b = \|(A\Sigma^*A^\top + R_\omega + I)^{-1}A\| < 1 - \delta$, $\|\widehat{\Sigma}\| = \|I\| + \frac{1}{1 - \|A\|^2}\|R_\omega\|$. It then follows that

$$\|\Sigma_{\mu_s^*} - \Sigma_{\mu_t^*}\|$$
$$\leq \sum_{i=T+1}^{s-1}b^{2(i-t_0)}\|\Sigma^*\|^2\|\Sigma_{(s-i+t_0)(s-i+t_0)}\|$$
$$+ 2\sum_{i=T+1}^{s-1}\sum_{j=t_0}^{i-1}b^{(i+j-2t_0)}\|\Sigma^*\|^2\|\Sigma_{(s-i+t_0)(s-j+t_0)}\|$$
$$\leq \sum_{i=T+1}^{s-1}b^{2(i-t_0)}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$
$$+ 2\sum_{i=T+1}^{s-1}\sum_{j=t_0}^{i-1}b^{(i+j-2t_0)}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$
$$= \frac{b^{2(T+1-t_0)}(1 - b^{2(s-T-1)})}{1 - b^2}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$
$$+ 2\sum_{i=T+1}^{s-1}\frac{b^{(i-t_0)}(1 - b^{(i-t_0)})}{1 - b}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$
$$\leq \frac{b^{2(T+1-t_0)}}{1 - b^2}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$
$$+ 2\frac{b^{(T+1-t_0)}(1 - b^{(s-T-1)})}{(1 - b)^2}\|\Sigma^*\|^2\|\widehat{\Sigma}\|$$

$$\le \frac{b^{2(T+1-t_0)}}{1-b^2}\|\mathbf{\Sigma}^*\|^2\|\widehat{\mathbf{\Sigma}}\| + 2\frac{b^{(T+1-t_0)}}{(1-b)^2}\|\mathbf{\Sigma}^*\|^2\|\widehat{\mathbf{\Sigma}}\|. \quad (63)$$

Since there exists $0 < \delta < 1$ such that $b < 1 - \delta$, for any $\epsilon > 0$, there exists an integer T such that for any $s, t > T$, $\|\mathbf{\Sigma}_{\boldsymbol{\mu}_s^*} - \mathbf{\Sigma}_{\boldsymbol{\mu}_t^*}\| < \epsilon$. Therefore, $\mathbf{\Sigma}_{\boldsymbol{\mu}_t^*}$ converges as $t \to \infty$.

Therefore, given the initial state $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$, the limiting covariance matrix of $\boldsymbol{\mu}_t^*$ exists. Moreover, from the definition of $\mathbf{\Sigma}_{\boldsymbol{\mu}_t^*}$ in (60), $\mathbf{\Sigma}_{\boldsymbol{\mu}_t^*}$ is independent of the initial state $\{\boldsymbol{x}_{t_0}, \boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}^*\}$. Since the limiting distribution of $\boldsymbol{\mu}_t^*$ exists and is independent of the initial state, from the definition of stationary distribution of Markov chain [53], $\boldsymbol{\mu}_t^*$ has a stationary distribution, which is a Gaussian distribution.

Step 3. We then show that $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain and its stationary distribution exists and is unique. We have that

$$p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*, \cdots, \boldsymbol{\mu}_{t-1}^*)$$
$$= \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t, \boldsymbol{\mu}_{t_0}^*)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*)}$$
$$\cdot \frac{p_{t_0}(\boldsymbol{\mu}_{t_0+1}^*, \cdots, \boldsymbol{\mu}_{t-1}^*|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t, \boldsymbol{\mu}_{t_0}^*)}{p_{t_0}(\boldsymbol{\mu}_{t_0+1}^*, \cdots, \boldsymbol{\mu}_{t-1}^*|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*)}$$
$$\overset{(a)}{=} \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t, \boldsymbol{\mu}_{t_0}^*)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*)}$$
$$= p_{t_0}(\boldsymbol{y}_t|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*)$$
$$\overset{(b)}{=} \frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}}$$
$$\cdot \frac{1}{\sqrt{\det((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}$$
$$\cdot \exp\Big(-\frac{1}{2}\Big((\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*)^\top(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*)$$
$$+ \boldsymbol{y}_t^\top\boldsymbol{y}_t - \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*) + \boldsymbol{y}_t\big)^\top$$
$$\cdot \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I}\big)^{-1}$$
$$\cdot \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*) + \boldsymbol{y}_t\big)\Big)\Big), \quad (64)$$

which depends only on $\boldsymbol{\mu}_{t-1}^*$, and (a) is due to the fact that conditioning on $(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})$, $(\boldsymbol{\mu}_{t_0+1}^*, \cdots, \boldsymbol{\mu}_{t-1}^*)$ is independent of $\boldsymbol{y}_t$ and (b) is from the update rule of the forward variable in Lemma 1. We then have that

$$p_{t_0}(\boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t_0}^*, \cdots, \boldsymbol{\mu}_{t-1}^*)$$
$$= p_{t_0}(\boldsymbol{y}_t|\boldsymbol{\mu}_{t-1}^*)p_{t_0}(\boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*, \boldsymbol{y}_t)$$
$$= p_{t_0}(\boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{y}_{t-1}, \boldsymbol{\mu}_{t-1}^*). \quad (65)$$

Therefore, $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is a Markov chain.

Since in **Step 2**, we show that the stationary distribution of $\boldsymbol{\mu}_t^*$ exists, we have that

$$\int \big(\pi(\boldsymbol{\mu}_{t-1}^*)p_{t_0}(\boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*)d\boldsymbol{\mu}_{t-1}^*\big)$$
$$\cdot p_{t_0}(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_t^*)p_{t_0}(\boldsymbol{\mu}_{t+1}^*|\boldsymbol{\mu}_t, \boldsymbol{y}_{t+1})d\boldsymbol{\mu}_t^*\boldsymbol{y}_t$$
$$= \int \pi(\boldsymbol{\mu}_t^*)p_{t_0}(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_t^*)p_{t_0}(\boldsymbol{\mu}_{t+1}^*|\boldsymbol{\mu}_t, \boldsymbol{y}_{t+1})d\boldsymbol{\mu}_t^*$$
$$= \int \pi(\boldsymbol{\mu}_t^*)p_{t_0}(\boldsymbol{y}_{t+1}, \boldsymbol{\mu}_{t+1}^*|\boldsymbol{\mu}_t^*)d\boldsymbol{\mu}_t^*. \quad (66)$$

Therefore, $\int \pi(\boldsymbol{\mu}_{t-1}^*)p_{t_0}(\boldsymbol{y}_t, \boldsymbol{\mu}_t^*|\boldsymbol{\mu}_{t-1}^*)d\boldsymbol{\mu}_{t-1}^*$ is a stationary distribution of $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$. Following the same techniques as in the proof of Lemma 3, we have that $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is π-irreducible. Therefore, the stationary distribution of $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is unique. $\qquad\square$

APPENDIX E
PROOF OF THEOREM 1

Proof: **Step 1.** We first consider a special case where $\mathbf{\Sigma}_{t_0} = \mathbf{\Sigma}^*$. We have that

$$\log p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)$$
$$= \log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})} + \log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-2})}$$
$$+ \cdots + \log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1})}{p_{t_0}^*(\boldsymbol{y}_{t_0})} + \log p_{t_0}^*(\boldsymbol{y}_{t_0}) \quad (67)$$

and

$$\log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}$$
$$= \log \frac{\int \alpha_t(\boldsymbol{x}_t)d\boldsymbol{x}_t}{\int \alpha_{t-1}(\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1}}$$
$$= \log\Big(\frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}}$$
$$\cdot \frac{1}{\sqrt{\det((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}\Big)$$
$$- \frac{1}{2}\Big((\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*)^\top(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*)$$
$$+ \boldsymbol{y}_t^\top\boldsymbol{y}_t - \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*) + \boldsymbol{y}_t\big)^\top$$
$$\cdot \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I}\big)^{-1}$$
$$\cdot \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*) + \boldsymbol{y}_t\big)\Big), \quad (68)$$

where the second equality is from the updated rule of $\alpha_t(\boldsymbol{x}_t)$ in (12).

Since $\boldsymbol{\mu}_t^* = (\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\boldsymbol{\mu}_{t-1}^* + \mathbf{\Sigma}^*\boldsymbol{y}_t$, we replace $\boldsymbol{A}\boldsymbol{\mu}_{t-1}^*$ by $(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu}_t^* - \mathbf{\Sigma}^*\boldsymbol{y}_t)$. Therefore, $\log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}$ can be equivalently written as a function of only $\boldsymbol{\mu}_t^*, \boldsymbol{y}_t$. Let $h(\boldsymbol{\mu}_t^*, \boldsymbol{y}_t) = \log \frac{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}$. We have the following explicit expression for $h(\boldsymbol{\mu}_t^*, \boldsymbol{y}_t)$,

$$h(\boldsymbol{\mu}_t^*, \boldsymbol{y}_t)$$
$$= \log\Big(\frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}} \quad\quad (69)$$
$$\cdot \frac{1}{\sqrt{\det((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}\Big)$$
$$- \frac{1}{2}\Big(((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu}_t^* - \mathbf{\Sigma}^*\boldsymbol{y}_t))^\top$$
$$\cdot (\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu}_t^* - \mathbf{\Sigma}^*\boldsymbol{y}_t))$$
$$+ \boldsymbol{y}_t^\top\boldsymbol{y}_t - \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})$$
$$\cdot (\boldsymbol{\mu}_t^* - \mathbf{\Sigma}^*\boldsymbol{y}_t)) + \boldsymbol{y}_t\big)^\top \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I}\big)^{-1}$$
$$\cdot \big((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}((\boldsymbol{A}\mathbf{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})(\boldsymbol{\mu}_t^* - \mathbf{\Sigma}^*\boldsymbol{y}_t))$$
$$+ \boldsymbol{y}_t\big)\Big) \quad (70)$$

Since $\pi(\boldsymbol{\mu}_t^*), \pi(\boldsymbol{y}_t)$ are Gaussian distributions and $h(\boldsymbol{\mu}, \boldsymbol{y})$ is a quadratic function of $\boldsymbol{\mu}, \boldsymbol{y}$, it can be easily verified that $\mathbb{E}_\pi[|h(\boldsymbol{\mu}, \boldsymbol{y})|] < \infty$. We further note that $\mathbb{R}^{2K}$ is a state space with countably generated σ-algebra [52] and $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$ is π-irreducible and aperiodic. Therefore, from the ergodic theorem of Markov chain [52], [53], we have that under $\mathbb{P}_{t_0}$,

$$\lim_{t \to \infty} \frac{1}{t} \log p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) = \mathbb{E}_\pi[h(\boldsymbol{\mu}, \boldsymbol{y})] \quad (71)$$

almost surely. Moreover, the convergence result does not depend on the initial state of $\{\boldsymbol{y}_t, \boldsymbol{\mu}_t^*\}_{t=t_0}^\infty$.

Step 2. We then show that $\lim_{t \to \infty} \frac{1}{t} \log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) = \mathbb{E}_\pi[h(\boldsymbol{\mu}, \boldsymbol{y})]$ almost surely under $\mathbb{P}_{t_0}$ by showing that

$$\lim_{t \to \infty} \frac{1}{t} \big(\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) - \log p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) \big) = 0 \quad (72)$$

almost surely under $\mathbb{P}_{t_0}$.

We have that

$$\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)$$
$$= \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})} + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-2})}$$
$$+ \cdots + \log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1})}{p_{t_0}(\boldsymbol{y}_{t_0})} + p_{t_0}(\boldsymbol{y}_{t_0}) \quad (73)$$

and

$$\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_t)}{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t-1})}$$
$$= \log \Big(\frac{\frac{1}{\sqrt{(2\pi)^K \det(\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)}}}{\frac{1}{\sqrt{\det((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})}}} \Big)$$
$$- \frac{1}{2} \Big((\boldsymbol{A}\boldsymbol{\mu}_{t-1})^\top (\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} (\boldsymbol{A}\boldsymbol{\mu}_{t-1})$$
$$+ \boldsymbol{y}_t^\top \boldsymbol{y}_t - ((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}) + \boldsymbol{y}_t)^\top$$
$$\cdot ((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{I})^{-1}$$
$$((\boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_{t-1}) + \boldsymbol{y}_t) \Big). \quad (74)$$

To show that $\lim_{t \to \infty} \frac{1}{t} \big(\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) - \log p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) \big) = 0$. It suffices to show the limit sum of each term in (74) converges. Here we provide the proof of $\lim_{t \to \infty} \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \big((\boldsymbol{A}\boldsymbol{\mu}_i)^\top (\boldsymbol{A}\boldsymbol{\Sigma}_{i-1}\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i) - (\boldsymbol{A}\boldsymbol{\mu}_i^*)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*) \big) = 0$. The rest of the terms can be proved using the same techniques. Let

$$(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} = (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_t^1,$$
$$(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}$$
$$= (\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_t^2,$$
$$(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} = (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{\Lambda}_t^3.$$

Since $\boldsymbol{\Sigma}_t$ converges to $\boldsymbol{\Sigma}^*$ as $t \to \infty$ and $(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}, (\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}, (\boldsymbol{A}\boldsymbol{\Sigma}_t\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}$ are continuous in $\boldsymbol{\Sigma}_t$, we have that $\lim_{t \to \infty} \|\boldsymbol{\Lambda}_t^1\| = 0$, $\lim_{t \to \infty} \|\boldsymbol{\Lambda}_t^2\| = 0$ and $\lim_{t \to \infty} \|\boldsymbol{\Lambda}_t^3\| = 0$. Therefore, $\|\boldsymbol{\Lambda}_t^1\|, \|\boldsymbol{\Lambda}_t^2\|, \|\boldsymbol{\Lambda}_t^3\|$ are uniformly upper bounded for

any t. Denote by $\|\boldsymbol{\Lambda}^1\|, \|\boldsymbol{\Lambda}^2\|, \|\boldsymbol{\Lambda}^3\|$ the upper bound of $\|\boldsymbol{\Lambda}_t^1\|, \|\boldsymbol{\Lambda}_t^2\|, \|\boldsymbol{\Lambda}_t^3\|$ respectively. Given the initial $\boldsymbol{\mu}_{t_0}$, from (15), we have the following representation for $\boldsymbol{\mu}_t$:

$$\boldsymbol{\mu}_{t_0+1} = \boldsymbol{\Lambda}_{t_0+1}^1 \boldsymbol{A}\boldsymbol{\mu}_{t_0} + \boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1} + \boldsymbol{\mu}_{t_0+1}^*,$$
$$\boldsymbol{\mu}_{t_0+2} = \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1 \big) \boldsymbol{A}$$
$$(\boldsymbol{\Lambda}_{t_0+1}^1 \boldsymbol{A}\boldsymbol{\mu}_{t_0} + \boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1} + \boldsymbol{\mu}_{t_0+1}^*) +$$
$$\big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^2 \big) \boldsymbol{y}_{t_0+2}$$
$$= \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1 \big) \boldsymbol{A}$$
$$(\boldsymbol{\Lambda}_{t_0+1}^1 \boldsymbol{A}\boldsymbol{\mu}_{t_0} + \boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1}) + \boldsymbol{\Lambda}_{t_0+2}^2 \boldsymbol{y}_{t_0+2} + \boldsymbol{\mu}_{t_0+2}^*$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$\boldsymbol{\mu}_{t_0+t-1} = \boldsymbol{\Lambda}_{t_0+t-1}^2 \boldsymbol{y}_{t_0+t-1}$$
$$+ \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+t-1}^1 \big) \boldsymbol{A}\boldsymbol{\Lambda}_{t_0+t-2}^2 \boldsymbol{y}_{t_0+t-2}$$
$$+ \big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+t-1}^1 \big) \boldsymbol{A}$$
$$\big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+t-2}^1 \big) \boldsymbol{A}\boldsymbol{\Lambda}_{t_0+t-3}^2 \boldsymbol{y}_{t_0+t-3}$$
$$+ \cdots + \prod_{i=t_0+2}^{t_0+t-1} \big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_i^1)\boldsymbol{A} \big)$$
$$\cdot (\boldsymbol{\Lambda}_{t_0+1}^1 \boldsymbol{A}\boldsymbol{\mu}_{t_0} + \boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1}) + \boldsymbol{\mu}_{t_0+t-1}^*. \quad (75)$$

Let $\boldsymbol{Y}_t = (\boldsymbol{y}_{t_0}, \boldsymbol{y}_{t_0+1}, \cdots, \boldsymbol{y}_{t_0+t-1})$ and denote $\boldsymbol{\mu}_t$ as $F_t(\boldsymbol{Y}_t) + \boldsymbol{\mu}_t^*$. We then have that

$$\frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \Big((\boldsymbol{A}\boldsymbol{\mu}_i)^\top ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{\Lambda}_i^3)(\boldsymbol{A}\boldsymbol{\mu}_i)$$
$$- (\boldsymbol{A}\boldsymbol{\mu}_i^*)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*) \Big)$$
$$= \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \Big((\boldsymbol{A}F_i(\boldsymbol{Y}_i) + \boldsymbol{A}\boldsymbol{\mu}_i^*)^\top ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{\Lambda}_i^3)$$
$$\cdot (\boldsymbol{A}F_i(\boldsymbol{Y}_i) + \boldsymbol{A}\boldsymbol{\mu}_i^*) - (\boldsymbol{A}\boldsymbol{\mu}_i^*)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*) \Big)$$
$$= \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} \Big((\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}F_i(\boldsymbol{Y}_i))$$
$$+ (\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top \boldsymbol{\Lambda}_i^3(\boldsymbol{A}F_i(\boldsymbol{Y}_i)) + (\boldsymbol{A}\boldsymbol{\mu}_i^*)^\top \boldsymbol{\Lambda}_i^3(\boldsymbol{A}\boldsymbol{\mu}_i^*)$$
$$+ 2(\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*)$$
$$+ 2(\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top \boldsymbol{\Lambda}_i^3(\boldsymbol{A}\boldsymbol{\mu}_i^*) \Big). \quad (76)$$

In the following, we will show that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} (\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}F_i(\boldsymbol{Y}_i))$$
$$= 0$$

and

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=t_0}^{t_0+t-1} (\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*) = 0.$$

The convergence of the rest of terms can be proved using the same techniques. To simplify the notation, we set $\boldsymbol{\mu}_{t_0} = \boldsymbol{0}$. The proof for any arbitrary $\boldsymbol{\mu}_{t_0}$ can also be derived similarly. For $(\boldsymbol{A}F_i(\boldsymbol{Y}_i))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}F_i(\boldsymbol{Y}_i)), i = t_0, \cdots, t$, we have that

$$(\boldsymbol{A}F_{t_0+1}(\boldsymbol{Y}_{t_0+1}))^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}F_{t_0+1}(\boldsymbol{Y}_{t_0+1}))$$

$$= (\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1})^\top \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} \boldsymbol{A}\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1},$$

$$\big(\boldsymbol{A}F_{t_0+2}(\boldsymbol{Y}_{t_0+2})\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_{t_0+2}(\boldsymbol{Y}_{t_0+2})\big)$$
$$= \Big(\big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1)\boldsymbol{A}(\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1})$$
$$+ \boldsymbol{\Lambda}_{t_0+2}^2 \boldsymbol{y}_{t_0+2}\Big)^\top \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\boldsymbol{A}$$
$$\cdot \Big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1)\boldsymbol{A}(\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1})$$
$$+ \boldsymbol{\Lambda}_{t_0+2}^2 \boldsymbol{y}_{t_0+2}\Big),$$

$$\big(\boldsymbol{A}F_{t_0+3}(\boldsymbol{Y}_{t_0+3})\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_{t_0+3}(\boldsymbol{Y}_{t_0+3})\big)$$
$$= \Big(\big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+3}^1)\boldsymbol{A}$$
$$((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1)\boldsymbol{A}(\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1})$$
$$+ ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+3}^1)\boldsymbol{A}(\boldsymbol{\Lambda}_{t_0+2}^2 \boldsymbol{y}_{t_0+2})$$
$$+ \boldsymbol{\Lambda}_{t_0+3}^2 \boldsymbol{y}_{t_0+3}\Big)^\top \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\boldsymbol{A}$$
$$\cdot \Big(((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+3}^1)\boldsymbol{A}$$
$$\cdot ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+2}^1)\boldsymbol{A}(\boldsymbol{\Lambda}_{t_0+1}^2 \boldsymbol{y}_{t_0+1})$$
$$+ ((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_{t_0+3}^1)\boldsymbol{A}$$
$$\cdot (\boldsymbol{\Lambda}_{t_0+2}^2 \boldsymbol{y}_{t_0+2}) + \boldsymbol{\Lambda}_{t_0+3}^2 \boldsymbol{y}_{t_0+3}\Big)$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots \quad (77)$$

We note that there exists a $0 < \delta < 1$ such that for any $i \geq t_0$, $\big\|((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1} + \boldsymbol{\Lambda}_i^1)\boldsymbol{A}\big\| = \big\|(\boldsymbol{A}\boldsymbol{\Sigma}_i\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\boldsymbol{A}\big\| \leq \big\|(\boldsymbol{A}\boldsymbol{\Sigma}_i\boldsymbol{A}^\top + \boldsymbol{R}_\omega + \boldsymbol{I})^{-1}\big\| \|\boldsymbol{A}\| \leq (1-\delta)$. Let $B = 1 - \delta$ and $C = \big\|\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\boldsymbol{A}\big\|$. We then have that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1} \Big\|\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)\Big\|$$

$$\leq \lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1} \big\|F_i(\boldsymbol{Y}_i)\big\|^2 \big\|\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\boldsymbol{A}\big\|$$

$$\leq \lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1} C\big\|F_i(\boldsymbol{Y}_i)\big\|^2$$

$$\leq \lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1} C\bigg(\sum_{j=t_0}^{i} B^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|\bigg)^2$$

$$= \lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i}\sum_{k=t_0}^{i} CB^{2i-j-k}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|\|\boldsymbol{\Lambda}_k^2\|\|\boldsymbol{y}_k\|$$

$$\leq \lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i}\sum_{k=t_0}^{i} CB^{2i-j-k}\big(\|\boldsymbol{\Lambda}_j^2\|^2\|\boldsymbol{y}_j\|^2$$
$$+ \|\boldsymbol{\Lambda}_k^2\|^2\|\boldsymbol{y}_k\|^2\big), \qquad (78)$$

where the first inequality is due to the submultiplicative of operator norm, and the third inequality is from the explicit expression of $F_i(\boldsymbol{Y}_i)$ and the triangle inequality and the submultiplicative of operator norm. We then consider the coefficient of the sum of all terms containing $\|\boldsymbol{\Lambda}_{t_0}^2\|^2\|\boldsymbol{y}_{t_0}\|^2$,

denoted by $Co(\|\boldsymbol{\Lambda}_{t_0}^2\|^2\|\boldsymbol{y}_{t_0}\|^2)$. We have that

$$Co(\|\boldsymbol{\Lambda}_{t_0}^2\|^2\|\boldsymbol{y}_{t_0}\|^2) = \lim_{t\to\infty}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} 2CB^{2i-t_0-j}$$

$$\leq \lim_{t\to\infty}\sum_{i=t_0}^{t_0+t-1}\frac{2CB^{i-t_0}}{1-B}$$

$$\leq \frac{2C}{(1-B)^2}. \qquad (79)$$

Similarly, we can show that the coefficient of $\|\boldsymbol{\Lambda}_{t_0+1}^2\|^2\|\boldsymbol{y}_{t_0+1}\|^2, \cdots, \|\boldsymbol{\Lambda}_{t_0+t-1}^2\|^2\|\boldsymbol{y}_{t_0+t-1}\|^2$ are not larger than $\frac{2C}{(1-B)^2}$. Since $\lim_{t\to\infty}\|\boldsymbol{\Lambda}_t^2\|^2 = 0$, for any $\epsilon > 0$, there exists an integer T such that for any $t > T$, $\|\boldsymbol{\Lambda}_t^2\|^2 < \epsilon$. We then have that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)$$

$$\leq \lim_{t\to\infty}\frac{1}{t}2C\frac{\|\boldsymbol{\Lambda}_{t_0}^2\|^2\|\boldsymbol{y}_{t_0}\|^2 + \cdots + \|\boldsymbol{\Lambda}_{t_0+t-1}^2\|^2\|\boldsymbol{y}_{t_0+t-1}\|^2}{(1-B)^2}$$

$$\leq \lim_{t\to\infty}\frac{1}{t}2C\bigg(\frac{\|\boldsymbol{\Lambda}_{t_0}^2\|^2\|\boldsymbol{y}_{t_0}\|^2 + \cdots + \|\boldsymbol{\Lambda}_T^2\|^2\|\boldsymbol{y}_T\|^2}{(1-B)^2}$$
$$+ \frac{\epsilon\big(\|\boldsymbol{y}_{T+1}\|^2 + \cdots + \|\boldsymbol{y}_{t_0+t-1}\|^2\big)}{(1-B)^2}\bigg)$$

$$\leq \frac{2\epsilon C\mathbb{E}_\pi[\|\boldsymbol{y}\|^2]}{(1-B)^2}, \qquad (80)$$

where the last inequality is from the ergodic theorem of Markov chain [52], [53].

Since $\mathbb{E}_\pi[\|\boldsymbol{y}\|^2]$ is bounded, by letting $\epsilon \to 0$, we have that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big\|\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)\big\|$$
$$= 0. \qquad (81)$$

Therefore, we have that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)$$
$$= \boldsymbol{0}. \qquad (82)$$

We then consider $\lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*)$. We have that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big\|\big(\boldsymbol{A}F_i(\boldsymbol{Y}_i)\big)^\top (\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*)\big\|$$

$$\leq \lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\big\|F_i(\boldsymbol{Y}_i)\big\|\big\|\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}\boldsymbol{A}\big\|\|\boldsymbol{\mu}_i^*\|$$

$$\leq \lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1} C\big\|F_i(\boldsymbol{Y}_i)\big\|\|\boldsymbol{\mu}_i^*\|$$

$$= \lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|\|\boldsymbol{\mu}_i^*\|$$

$$\leq \lim_{t\to\infty}\frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\big(\|\boldsymbol{y}_j\|^2 + \|\boldsymbol{\mu}_i^*\|^2\big). \qquad (83)$$

It suffices to show that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2 = 0 \tag{84}$$

and

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{\mu}_i^*\|^2 = 0. \tag{85}$$

We have that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2$$

$$= \lim_{t\to\infty} \frac{1}{t}\sum_{j=t_0}^{t_0+t-1}\sum_{i=j}^{t_0+t-1} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2$$

$$\leq \lim_{t\to\infty} \frac{1}{t}\sum_{j=t_0}^{t_0+t-1} \frac{C}{1-B}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2. \tag{86}$$

For any $\epsilon > 0$, there exists an integer T such that for any $t > T$, $\|\boldsymbol{\Lambda}_t^2\| < \epsilon$. It then follows that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{j=t_0}^{t_0+t-1} \frac{C}{1-B}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2$$

$$\leq \lim_{t\to\infty} \frac{1}{t}C\Big(\frac{\|\boldsymbol{\Lambda}_{t_0}^2\|\|\boldsymbol{y}_{t_0}\|^2 + \cdots + \|\boldsymbol{\Lambda}_T^2\|\|\boldsymbol{y}_T\|^2}{1-B}$$

$$+ \frac{\epsilon\|\boldsymbol{y}_{T+1}\|^2 + \cdots + \epsilon\|\boldsymbol{y}_{t_0+t-1}\|^2}{1-B}\Big)$$

$$\leq \frac{\epsilon C \mathbb{E}_\pi[\|\boldsymbol{y}\|^2]}{1-B}. \tag{87}$$

Since $\mathbb{E}_\pi[\|\boldsymbol{y}\|^2]$ is bounded, by letting $\epsilon \to 0$, we have that $\lim_{t\to\infty} \frac{1}{t}\sum_{j=t_0}^{t_0+t-1} \frac{C}{1-B}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{y}_j\|^2 = 0$.

Consider $\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{\mu}_i^*\|^2$. We have that $\lim_{t\to\infty} B^t = 0$ and $\lim_{t\to\infty} \|\boldsymbol{\Lambda}_t^2\| = 0$. Therefore, for any $\epsilon > 0$, there exists S such that for $s > S$, $\|\boldsymbol{\Lambda}_{s-1}^2\| < \epsilon$ and $B^{s-1} < \epsilon$. Let $T = 2S$. For any $t > t_0 + T$, we have that

$$\sum_{j=t_0}^{t} CB^{t-j}\|\boldsymbol{\Lambda}_j^2\|$$

$$= C\big(\|\boldsymbol{\Lambda}_t^2\| + B\|\boldsymbol{\Lambda}_{t-1}^2\| + \cdots + B^{t-t_0}\|\boldsymbol{\Lambda}_{t_0}^2\|\big)$$

$$\leq C\big(\epsilon + B\epsilon + \cdots + B^{\lfloor\frac{t-t_0}{2}\rfloor}\epsilon + B^{\lfloor\frac{t-t_0}{2}\rfloor+1}\|\boldsymbol{\Lambda}_{t-\lfloor\frac{t-t_0}{2}\rfloor-1}^2\|$$

$$+ \cdots + B^{t-t_0}\|\boldsymbol{\Lambda}_{t_0}^2\|\big)$$

$$\leq C\Big(\frac{\epsilon}{1-B} + \frac{B^{\lfloor\frac{t-t_0}{2}\rfloor+1}}{1-B}\|\boldsymbol{\Lambda}^2\|\Big)$$

$$\leq \frac{C(1+\|\boldsymbol{\Lambda}^2\|)}{1-B}\epsilon, \tag{88}$$

where for the second inequality, we compute the sum of the first half and the second half respectively and use the fact that $\|\boldsymbol{\Lambda}_t^2\| \leq \|\boldsymbol{\Lambda}^2\|$ for any t. Since $\frac{C(1+\|\boldsymbol{\Lambda}^2\|)}{1-B}$ is bounded, the coefficient of $\|\boldsymbol{\mu}_i^*\|$ will converge to zero as $t \to \infty$. Therefore, following the same steps as in (80), we have that

$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\sum_{j=t_0}^{i} CB^{i-j}\|\boldsymbol{\Lambda}_j^2\|\|\boldsymbol{\mu}_i^*\|^2 = 0$. Therefore, we have that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\Big((\boldsymbol{A}\boldsymbol{\mu}_i)^\top\big((\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1} + \boldsymbol{\Lambda}_i^3\big)(\boldsymbol{A}\boldsymbol{\mu}_i)$$

$$- (\boldsymbol{A}\boldsymbol{\mu}_i^*)^\top(\boldsymbol{A}\boldsymbol{\Sigma}^*\boldsymbol{A}^\top + \boldsymbol{R}_\omega)^{-1}(\boldsymbol{A}\boldsymbol{\mu}_i^*)\Big) = 0, \tag{89}$$

and thus

$$\lim_{t\to\infty} \frac{1}{t}\big(\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})$$

$$- \log p_{t_0}^*(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})\big) = 0. \tag{90}$$

Step 3. Since the observations are independent under $\mathbb{P}_\infty$, we have that $\lim_{t\to\infty} \frac{1}{t}\log p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) = \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t_0+t-1}\log p_\infty(\boldsymbol{y}_i)$. Since $\log p_\infty(\boldsymbol{y})$ is a quadratic function of $\boldsymbol{y}$, we have that $\mathbb{E}_\pi[|\log p_\infty(\boldsymbol{y})|] < \infty$. From the ergodic theorem of Markov chain [52], [53], we have that under $\mathbb{P}_{t_0}$,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=t_0}^{t_0+t-1}\log p_\infty(\boldsymbol{y}_i) = \mathbb{E}_\pi[\log p_\infty(\boldsymbol{y})] \tag{91}$$

almost surely.

It then follows that under $\mathbb{P}_{t_0}$,

$$\lim_{t\to\infty} \frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}$$

$$= \lim_{t\to\infty} \frac{1}{t}\Big(\log p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1}) - \sum_{i=t_0}^{t_0+t-1}\log p_\infty(\boldsymbol{y}_i)\Big)$$

$$= \mathbb{E}_\pi[h(\boldsymbol{\mu}, \boldsymbol{y})] - \mathbb{E}_\pi[\log p_\infty(\boldsymbol{y})] \tag{92}$$

almost surely.

Let

$$\mathcal{K} = \mathbb{E}_\pi[h(\boldsymbol{\mu}, \boldsymbol{y})] - \mathbb{E}_\pi[\log p_\infty(\boldsymbol{y})]. \tag{93}$$

We then have that for any initial state $\boldsymbol{y}_{t_0}, \boldsymbol{\mu}_{t_0}$, $\lim_{t\to\infty} \frac{1}{t}\log \frac{p_{t_0}(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})}{p_\infty(\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{t_0+t-1})} = \mathcal{K}$ under $\mathbb{P}_{t_0}$ almost surely.

From (68), we have that given $\boldsymbol{\mu}_t^*$, the distribution of $\boldsymbol{y}_t$ under $\mathbb{P}_{t_0}$ can be fully specified, denoted by $p_{t_0}^*(\boldsymbol{y}_t|\boldsymbol{\mu}_t^*)$. Therefore, we have that $\pi(\boldsymbol{\mu}, \boldsymbol{y}) = \pi(\boldsymbol{\mu})p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu})$. Note that $h(\boldsymbol{\mu}, \boldsymbol{y}) = \log p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu})$. We then have that

$$\mathcal{K} = \mathbb{E}_\pi\Big[\log \frac{p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu})}{p_\infty(\boldsymbol{y})}\Big]$$

$$= \mathbb{E}_\pi\Big[\log \frac{\pi(\boldsymbol{\mu})p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu})}{\pi(\boldsymbol{\mu})p_\infty(\boldsymbol{y})}\Big]$$

$$= D\big(\pi(\boldsymbol{\mu})p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu})\|\pi(\boldsymbol{\mu})p_\infty(\boldsymbol{y})\big)$$

$$\geq 0, \tag{94}$$

where $D(\cdot\|\cdot)$ denotes the KL-Divergence between two distributions and the equality holds if and only if $\pi(\boldsymbol{\mu})p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu}) \neq \pi(\boldsymbol{\mu})p_\infty(\boldsymbol{y})$. From (68), we have that $\pi(\boldsymbol{\mu})p_{t_0}^*(\boldsymbol{y}|\boldsymbol{\mu}) \neq \pi(\boldsymbol{\mu})p_\infty(\boldsymbol{y})$. Therefore, we have that $\mathcal{K} > 0$.

Therefore, for any $\eta > 0$,

$$\lim_{t\to\infty} \sup_{t_0\geq 1} \text{esssup} \mathbb{P}_{t_0}\Big\{\max_{k\leq t}\sum_{i=t_0}^{t_0+k-1}\log \frac{p_{t_0}(\boldsymbol{y}_i|\boldsymbol{y}_{t_0}, \cdots, \boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i)}$$

$$\geq \mathcal{K}(1+\eta)t\big|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}\Big\} = 0. \tag{95}$$

Then (20) follows from [24, Theorem 1]. □

APPENDIX F
PROOF OF THEOREM 3

Proof: Let $\sigma_0 = 0$ and define the stopping times

$$\sigma_{m+1} = \inf\Big\{t > \sigma_m : \sum_{i=\sigma_m+1}^{t} \log \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$
$$\leq 0\Big\}, \text{for } m \geq 0. \tag{96}$$

We have that

$$\mathbb{E}_\infty\Bigg[\prod_{i=\sigma_m+1}^{t+1} \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \text{ for some } t > \sigma_m\Big|\mathcal{F}_t\Bigg]$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$
$$\times \mathbb{E}_\infty\Bigg[\frac{p_1(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}{p_\infty(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}\Big|\mathcal{F}_t\Bigg]$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$
$$\times \int p_1(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)d\boldsymbol{y}_{t+1}$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}. \tag{97}$$

Therefore, $\{\prod_{i=\sigma_m+1}^{t} \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}, \mathcal{F}_t, t > \sigma_m\}$ is a martingale under the pre-change distribution with mean 1. Therefore, from Doob's submartingale inequality [61] and the optional sampling theorem [61], we have that

$$\mathbb{P}_\infty\Big\{\sum_{i=\sigma_m+1}^{t} \log \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \geq c$$
$$\text{for some } t > \sigma_m\big|\mathcal{F}_{\sigma_m}\Big\} \leq e^{-c}. \tag{98}$$

Let $M = \inf\{m \geq 0 : \sigma_m < \infty$ and $\sum_{i=\sigma_m+1}^{t} \log \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \geq c$ for some $t > \sigma_m\}$. We have that

$$\mathbb{P}_\infty(M \geq m+1|\mathcal{F}_{\sigma_m})$$
$$= \mathbb{P}_\infty\Big\{\sum_{i=\sigma_m+1}^{t} \log \frac{p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} < c$$
$$\text{for all } t > \sigma_m\big|\mathcal{F}_{\sigma_m}\Big\}$$
$$\geq 1 - e^{-c}. \tag{99}$$

We then have that

$$\mathbb{P}_\infty(M > m) = \mathbb{E}_\infty\big[\mathbb{P}_\infty(M \geq m+1|\mathcal{F}_{\sigma_m})\mathbb{1}_{\{M \geq m\}}\big]$$
$$\geq (1 - e^{-c})\mathbb{P}_\infty(M > m-1)$$
$$\geq (1 - e^{-c})^2 \mathbb{P}_\infty(M > m-2)$$
$$\geq (1 - e^{-c})^m \mathbb{P}_\infty(M > 0)$$
$$\geq (1 - e^{-c})^m. \tag{100}$$

It then follows that

$$\mathbb{E}_\infty[\tau_c^*] \geq \mathbb{E}_\infty[M] \geq \sum_{m=0}^{\infty} \mathbb{P}_\infty(M > m)$$
$$\geq \sum_{m=0}^{\infty} (1 - e^{-c})^m = e^c. \tag{101}$$

Let $c = \log\gamma$, we have that $\mathbb{E}_\infty[\tau_c^*] \geq \gamma$.

Define $h_t(\boldsymbol{\mu},\boldsymbol{y})$ by replacing $\boldsymbol{\Sigma}^*$ with $\boldsymbol{\Sigma}_t$ in $h(\boldsymbol{\mu},\boldsymbol{y})$. We then have that $\log L_t = \sum_{i=1}^{t} h_i(\boldsymbol{\mu}_i,\boldsymbol{y}_i) - \log p_\infty(\boldsymbol{y}_i)$. From (92), we have that for $0 < \delta < 1$ and any initial $\boldsymbol{\mu}_t, \boldsymbol{y}_t$,

$$\lim_{n\to\infty} \sup_{t\geq t_0\geq 1} \text{esssup}\mathbb{P}_{t_0}\Big\{n^{-1}\sum_{i=t}^{t+n-1} h_i(\boldsymbol{y}_i,\boldsymbol{\mu}_i) - \log p_\infty(\boldsymbol{y}_i)$$
$$< \mathcal{K} - \delta|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1}\Big\} = 0. \tag{102}$$

This implies that

$$\sup_{t\geq t_0\geq 1} \text{esssup}\mathbb{P}_{t_0}\Big\{\sum_{i=t}^{t+n_c-1} h_i(\boldsymbol{y}_i,\boldsymbol{\mu}_i) - \log p_\infty(\boldsymbol{y}_i)$$
$$< c|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t-1}\Big\} \leq \delta \tag{103}$$

for all large c, where n_c is the largest integer $\leq (1-\delta)^{-1}\mathcal{K}^{-1}c$. We then have that for any $t_0 \geq 1$ and $l \geq 1$,

$$\text{esssup}\mathbb{P}_{t_0}\{\tau_c^* - t_0 > ln_c|\mathcal{F}_{t_0-1}\}$$
$$\leq \text{esssup}\mathbb{P}_{t_0}\Big\{\sum_{i=t_0+(j-1)n_c}^{t_0+jn_c-1} h_i(\boldsymbol{y}_i,\boldsymbol{\mu}_i) - \log p_\infty(\boldsymbol{y}_i) < c$$
$$\text{for all } 1 \leq j \leq l|\mathcal{F}_{t_0-1}\Big\}$$
$$\leq \delta^l. \tag{104}$$

Therefore,

$$\sup_{t_0\geq 1} \text{esssup}\mathbb{E}_{t_0}[n_c^{-1}(\tau_c^* - t_0)^+|\mathcal{F}_{t_0-1}] \leq \sum_{l=0}^{\infty} \delta^l = (1-\delta)^{-1}.$$

Let $c = \log\gamma$ and $\gamma \to \infty$. Since $n_c \sim (1-\delta)^{-1}\mathcal{K}^{-1}c$, we have that

$$\sup_{t_0\geq 1} \text{esssup}\mathbb{E}_{t_0}[(\tau_c^* - t_0)^+|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{t_0-1}] \leq \frac{\log\gamma}{\mathcal{K}}(1 + o(1)).$$

This completes the proof. □

APPENDIX G
PROOF OF THEOREM 7

Proof: Define $p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})$ by replacing $\boldsymbol{A}, \boldsymbol{R}_\omega, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ in $p_1(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})$ with $\widehat{\boldsymbol{A}}_i, \widehat{\boldsymbol{R}}_{\omega,i}, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i$. Note that the OGA-CuSum can be equivalently written as

$$\tau_{\text{OGA}} = \inf\Big\{t : \max_{1\leq k\leq t} \sum_{i=k}^{t} \log \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \geq c\Big\}. \tag{105}$$

Let $\sigma_0 = 0$ and define the stopping times

$$\sigma_{m+1} = \inf\Big\{t > \sigma_m : \sum_{i=\sigma_m+1}^{t} \log \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_\infty(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$

$$\leq 0\Big\}, \text{ for } m \geq 0. \tag{106}$$

We have that

$$\mathbb{E}_{\infty}\left[\prod_{i=\sigma_m+1}^{t+1} \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \text{ for some } t > \sigma_m \Big| \mathcal{F}_t\right]$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$

$$\times \mathbb{E}_{\infty}\left[\frac{p_{\widehat{\theta}_{t+1}}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}{p_{\infty}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)}\Big|\mathcal{F}_t\right]$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}$$

$$\times \int p_{\widehat{\theta}_{t+1}}(\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_t)d\boldsymbol{y}_{t+1}$$

$$= \prod_{i=\sigma_m+1}^{t} \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}. \tag{107}$$

Therefore, $\{\prod_{i=\sigma_m+1}^{t} \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}, \mathcal{F}_t, t > \sigma_m\}$ is a martingale under the pre-change distribution with mean 1. Therefore, from Doob's submartingale inequality [61] and the optional sampling theorem [61], we have that

$$\mathbb{P}_{\infty}\Big\{\sum_{i=\sigma_m+1}^{t} \log \frac{p_{\widehat{\theta}_i}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})}{p_{\infty}(\boldsymbol{y}_i|\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{i-1})} \geq c$$

$$\text{for some } t > \sigma_m \Big| \mathcal{F}_{\sigma_m}\Big\} \leq e^{-c}. \tag{108}$$

Theorem 7 can then be proved following the same techniques as in the proof of Theorem 3. $\qquad\square$

REFERENCES

[1] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] V. V. Veeravalli and T. Banerjee, "Quickest change detection," in *Academic Press Library in Signal Processing, Array Signal Processing*, vol. 3. Amsterdam, The Netherlands: Elsevier, 2013, pp. 209–256.

[3] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Boca Raton, FL, USA: CRC Press, 2014.

[4] 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.

[5] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, nos. 1–2, pp. 100–115, 1954.

[6] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, no. 1, pp. 206–227, Mar. 1985.

[7] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, Dec. 1986.

[8] Y. Ritov, "Decision theoretic optimality of the cusum procedure," *Ann. Statist.*, vol. 18, no. 3, pp. 1464–1469, Sep. 1990.

[9] A. Tartakovsky, "Asymptotic properties of CUSUM and Shiryaev's procedures for detecting a change in a nonhomogeneous Gaussian process," *Math. Methods Statist.*, vol. 4, no. 4, pp. 389–404, 1995.

[10] I. Y. H. Gu, M. H. J. Bollen, and E. Styvaktakis, "The use of time-varying AR models for the characterization of voltage disturbances," in *Proc. IEEE Power Eng. Soc. Winter Meeting. Conf.*, Feb. 2000, pp. 2943–2948.

[11] D. Sidorov, D. Panasetsky, and V. Šmádl, "Non-stationary autoregressive model for on-line detection of inter-area oscillations in power systems," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Eur. (ISGT Europe)*, Oct. 2010, pp. 1–5.

[12] S. Li and X. Wang, "Monitoring disturbances in smart grids using distributed sequential change detection," in *Proc. 5th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2013, pp. 432–435.

[13] S. Li and X. Wang, "Cooperative change detection for voltage quality monitoring in smart grids," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 1, pp. 86–99, Jan. 2016.

[14] M. N. Kurt, Y. Yilmaz, and X. Wang, "Distributed quickest detection of cyber-attacks in smart grid," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2015–2030, 2018.

[15] L. Chen, S. Li, and X. Wang, "Quickest fault detection in photovoltaic systems," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1835–1847, May 2018.

[16] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. Berlin, Germany: Springer, 2007.

[17] B. Yakir, *Optimal Detection of a Change in Distribution When the Observations Form a Markov Chain With a Finite State Space* (Institute of Mathematical Statistics), vol. 23. Hayward, CA, USA: IMS, 1994, pp. 346–358.

[18] A. M. Polansky, "Detecting change-points in Markov chains," *Comput. Statist. Data Anal.*, vol. 51, no. 12, pp. 6013–6026, Aug. 2007.

[19] 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.

[20] B. Darkhovsky, "Change-point problem for high-order Markov chain," *Sequential Anal.*, vol. 30, no. 1, pp. 41–51, Jan. 2011.

[21] J.-G. Xian, D. Han, and J.-Q. Yu, "Online change detection of Markov chains with unknown post-change transition probabilities," *Commun. Statist. Theory Methods*, vol. 45, no. 3, pp. 597–611, Feb. 2016.

[22] H. Chen, J. Tang, and A. Gupta, "Change detection of Markov kernels with unknown pre and post change kernel," in *Proc. IEEE 61st Conf. Decis. Control (CDC)*, Dec. 2022, pp. 4814–4820.

[23] J. J. Ford, J. M. Kennedy, C. Tompkins, J. James, and A. McFadyen, "Exactly optimal quickest change detection of Markov chains," *IEEE Control Syst. Lett.*, vol. 7, pp. 2749–2754, 2023.

[24] 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.

[25] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou, "Data-driven quickest change detection in Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[26] B. Chen and P. Willett, "Detection of hidden Markov model transient signals," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 4, pp. 1253–1268, Oct. 2000.

[27] L. Gereneser and G. Molnar-Saska, "Change detection of hidden Markov models," in *Proc. 43rd IEEE Conf. Decis. Control (CDC)*, vol. 2, Jul. 2004, pp. 1754–1758.

[28] S. Dayanik and C. Goulding, "Sequential detection and identification of a change in the distribution of a Markov-modulated random sequence," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3323–3345, Jul. 2009.

[29] C.-D. Fuh and Y. Mei, "Quickest change detection and Kullback–Leibler divergence for two-state hidden Markov models," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4866–4878, Sep. 2015.

[30] C.-D. Fuh and A. G. Tartakovsky, "Asymptotic Bayesian theory of quickest change detection for hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 511–529, Jan. 2019.

[31] J. J. Ford, J. James, and T. L. Molloy, "Exactly optimal Bayesian quickest change detection for hidden Markov models," *Automatica*, vol. 157, Nov. 2023, Art. no. 111232.

[32] P. Bougerol, "Théorèmes limite pour les systèmes linéaires à coefficients Markoviens," *Probab. Theory Rel. Fields*, vol. 78, no. 2, pp. 193–221, Jun. 1988.

[33] M. Bagshaw and R. A. Johnson, "Sequential procedures for detecting parameter changes in a time-series model," *J. Amer. Stat. Assoc.*, vol. 72, no. 359, pp. 593–597, Sep. 1977.

[34] R. A. Davis, D. Huang, and Y.-C. Yao, "Testing for a change in the parameter values and order of an autoregressive model," *Ann. Statist.*, vol. 23, no. 1, pp. 282–304, Feb. 1995.

[35] E. Gombay, "Change detection in autoregressive time series," *J. Multivariate Anal.*, vol. 99, no. 3, pp. 451–464, Mar. 2008.

[36] J. Kang and S. Lee, "Parameter change test for Poisson autoregressive models," *Scandin. J. Statist.*, vol. 41, no. 4, pp. 1136–1152, Dec. 2014.

[37] N. Hilgert, G. Verdier, and J.-P. Vila, "Change detection for uncertain autoregressive dynamic models through nonparametric estimation," *Stat. Methodol.*, vol. 33, pp. 96–113, Dec. 2016.

[38] F. Akashi, H. Dette, and Y. Liu, "Change-point detection in autoregressive models with no moment assumptions," *J. Time Ser. Anal.*, vol. 39, no. 5, pp. 763–786, Sep. 2018.

[39] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[40] D. Kalus, M. Muma, and A. M. Zoubir, "Distributed robust change point detection for autoregressive processes with an application to distributed voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 3906–3910.

[41] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, Dec. 1971.

[42] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *Ann. Statist.*, vol. 23, no. 1, pp. 255–271, Feb. 1995.

[43] 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.

[44] 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.

[45] Y. Mei, "Sequential change-point detection when unknown parameters are present in the pre-change distribution," *Ann. Statist.*, vol. 34, no. 1, pp. 92–122, Feb. 2006.

[46] B. Brodsky and B. Darkhovsky, "Asymptotically optimal sequential change-point detection under composite hypotheses," in *Proc. 44th IEEE Conf. Decis. Control*, Dec. 2005, pp. 7347–7351.

[47] S. Pergamenchtchikov and A. G. Tartakovsky, "Asymptotically optimal pointwise and minimax change-point detection for general stochastic models with a composite post-change hypothesis," *J. Multivariate Anal.*, vol. 174, Nov. 2019, Art. no. 104541.

[48] Y. Cao, L. Xie, Y. Xie, and H. Xu, "Sequential change-point detection via online convex optimization," *Entropy*, vol. 20, no. 2, p. 108, Feb. 2018.

[49] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of quickest change detection," *Theory Probab. Appl.*, vol. 49, no. 3, pp. 458–497, Jan. 2005.

[50] M. Baron and A. G. Tartakovsky, "Asymptotic optimality of change-point detection schemes in general continuous-time models," *Sequential Anal.*, vol. 25, no. 3, pp. 257–296, Sep. 2006.

[51] A. G. Tartakovsky, "On asymptotic optimality in sequential changepoint detection: Non-iid case," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3433–3450, Jun. 2017.

[52] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probab. Surveys*, vol. 1, pp. 20–71, Jan. 2004.

[53] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer, 2012.

[54] M. Raginsky, R. M. Willett, C. Horn, J. Silva, and R. F. Marcia, "Sequential anomaly detection in the presence of noise and limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5544–5562, Aug. 2012.

[55] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*. London, U.K.: Springer, 2002.

[56] A. G. Tartakovsky, "Asymptotic performance of a multichart CUSUM test under false alarm probability constraint," in *Proc. 44th IEEE Conf. Decis. Control*, Dec. 2005, pp. 320–325.

[57] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: A new criterion for autoregression," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4024–4043, Jun. 2018.

[58] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "SMC2: An efficient algorithm for sequential analysis of state space models," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 75, no. 3, pp. 397–426, Jun. 2013.

[59] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[60] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, "Bayesian model comparison with the Hyvärinen score: Computation and consistency," *J. Amer. Stat. Assoc.*, vol. 114, no. 528, pp. 1826–1837, Oct. 2019.

[61] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

Zhongchang Sun (Student Member, IEEE) received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University at Buffalo, The State University of New York. His research interests include hypothesis testing, quickest change detection, distributional robust optimization, and reinforcement learning.

Shaofeng Zou (Member, IEEE) received the B.E. degree (Hons.) from Shanghai Jiao Tong University, Shanghai, China, in 2011, and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2016. From 2016 to 2018, he was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign. He is currently an Assistant Professor with the Department of Electrical Engineering, University at Buffalo, The State University of New York. His research interests include reinforcement learning, machine learning, statistical signal processing, and information theory. He received the National Science Foundation CAREER Award in 2024, the National Science Foundation CRII Award in 2019, and the 2023 AAAI Distinguished Paper Award. He has been serving as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING since 2023.