Received 10 January 2024; revised 27 March 2024; accepted 21 April 2021; date of current version 21 April 2024. Sonia Martinez

Model-Free Change Point Detection for Mixing Processes

HAO CHEN[1] (Student Member, IEEE), ABHISHEK GUPTA[1] (Member, IEEE), YIN SUN[2] (Member, IEEE), NESS SHROFF[1] (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA

[2]Department of Electrical and Computer Engineering, Auburn University Auburn, AL 36849, USA

CORRESPONDING AUTHOR: HAO CHEN (e-mail: chen.6945@osu.edu)

This work was supported by Cisco Systems and the U.S. National Science Foundation

ABSTRACT This paper considers the change point detection problem under dependent samples. In particular, we provide performance guarantees for the MMD-CUSUM test under exponentially α, β, and fast ϕ-mixing processes, which significantly expands its utility beyond the i.i.d. and Markovian cases used in previous studies. We obtain lower bounds for average-run-length (ARL) and upper bounds for average-detection-delay (ADD) in terms of the threshold parameter. We show that the MMD-CUSUM test enjoys the same level of performance as the i.i.d. case under fast ϕ-mixing processes. The MMD-CUSUM test also achieves strong performance under exponentially α/β-mixing processes, which are significantly more relaxed than existing results. The MMD-CUSUM test statistic adapts to different settings without modifications, rendering it a completely data-driven, dependence-agnostic change point detection scheme. Numerical simulations are provided at the end to evaluate our findings.

INDEX TERMS Change point detection, kernel method, mixing processes

I. INTRODUCTION

Change point detection studies the problem of monitoring for abrupt changes in the statistical properties of an observation sequence, which has been widely considered in the literature [1, 2, 3, 4]. Change point detection has a diverse application that spans many areas, including cybersecurity, network intrusion detection, automated fault monitoring, factory quality control, etc. In many of these application scenarios, one may face various challenges, such as complex unknown dynamics, noisy non-i.i.d observations, and unknown preand post-change distributions. Ideally, a completely datadriven method with very few distributional assumptions (independence, density functions, etc.) would be preferred. The goal of this paper is to study the change point detection problem under a completely data-driven setting. To tackle this problem, we employ the MMD-CUSUM statistic proposed in [5] and analyze its performance under three common mixing conditions, namely α, β, and ϕ-mixing.

The MMD-CUSUM statistic is an extension of the wellknown CUSUM statistic [6] with the maximum mean discrepancy (MMD). MMD has wide adoption in statistical twosample tests [7] and the training of generative adversarial networks [8]. As a probability distance, MMD can be easily estimated from samples on general domains (continuous or discrete) without the need for a density function. Thus, it is well suited for change point detection under the completely data-driven setting where pre- and post-change distributions can be unknown. Additionally, kernel methods have wide compatibility [9, 10] due to the diversity of kernel functions with different data structures, such as discrete data, continuous data, graphical data, etc. Thus, the kernel base method has vast application potential in designing completely datadriven change point detection schemes. In particular, the sequential testing procedures using the maximum mean discrepancy (MMD) have sparked some research interests lately [11, 12, 13, 14, 5]. Most of the existing studies focus on studying the properties of the MMD-based procedures under the i.i.d. case. For continuous state space Markov chains, the MMD-CUSUM test is proposed in

[5] for uniformly ergodic Markov chains, which is known to be hard to satisfy in practice.

Thus, more relaxed assumptions need to be considered to meet the demands of the completely data-driven setting. The main challenge in generalizing the performance analysis of MMD-CUSUM lies in the dependence of samples. Our proposal assumes the mixing property of the stochastic processes generated by the dynamic system. Mixing measures the dependence in the process by its definition [15], and it is widely considered in extending various results in probability theory to dependent time series [16, 17, 18]. Thus, establishing the performance bounds under various mixing conditions is a natural choice. Furthermore, the mixing conditions we assume highlight the fundamental limit for MMD-CUSUM to achieve a good performance; that is, the speed and strength of the mixing condition the processes satisfy.

In the current paper, we analyze the performance of MMD-CUSUM under three common mixing conditions, namely α, β, and ϕ-mixing. We provide bounds on the average-run-length (ARL) and average-detection-delay (ADD) which are the common performance metrics [19]. ARL characterizes how frequently the false alarm occurs and ADD characterizes the quickness of the reaction. As outlined in [20], the information-theoretic lower bounds are $O(\exp(b))$ for ARL and $O(b)$ for ADD for large $b > 0$, where b is the threshold parameter. We show that under the fast ϕmixing condition, the MMD-CUSUM achieves these lower bounds and thus is order optimal. Under the exponential α/β-mixing, ADD is bounded by $O(b)$ where ARL is bounded by $O(\exp b^{\gamma/(\gamma+1)})$, where $\gamma > 0$ controls the mixing speed (more details in IV).

The rest of the paper is organized as follows. Section II introduces the necessary background about reproducing kernel Hilbert space and mixing processes. Section III states the problem setting for online change point detection and introduces the MMD-CUSUM test statistic. Section IV establishes the main results of this paper. Section V presents the experiments of the MMD-CUSUM test on synthetic datasets. We conclude the paper with discussions of the limitations and future work in Section VI and VII.

A. Related works

Continuous efforts have been made to adapt the kernel two-sample test to a sequential setting, i.e., change point detection. Early work has been focused on detection change in a stream of i.i.d. samples [11, 12, 13, 14]. In [11, 12], the authors developed a Shewhart chart-type [21] procedure that maintains a running estimate of the MMD between a set of curated reference data and incoming samples within a fixed sliding window. Analysis shows strong performance guarantees with an $O(\exp(b^2))$ average-run-length (ARL) and an $O(b)$ average detection delay (ADD), where b is the threshold. However, testing schemes with sliding windows suffer from loss of information as older samples are discarded. To maintain history information, kernel-based CUSUM-type statistics were proposed in [14] with an $O(\exp(b))$ averagerun-length (ARL) and an $O(b)$ average detection delay (ADD). In [13], the authors devised a neural network-based kernel selection strategy that finds a kernel whose MMD can best separate the nominal distribution from an adversarial one. The testing scheme is to estimate the MMD with the selected kernel on two adjacent sliding windows. Empirical studying shows promising performance, albeit without theoretical guarantees.

The analysis of the above methods is based heavily on the i.i.d. assumption. Their technique and results do not carry over naturally to the non-i.i.d. case. Due to the ubiquity of time series data in machine learning, signal processing, economics, and dynamic systems, the i.i.d. assumption limits the application of these methods. More recently, researchers have been adapting the kernel-based change point detection to dependent data. In [5], the MMD-CUSUM test is proposed and analyzed under the setting of uniformly ergodic Markov chains on general state space. Recently, [22] extended the analysis of MMD-CUSUM to noisy observations of uniformly ergodic Markov chains, i.e., hidden Markov models (HMM). Both cases are special cases of ϕ-mixing processes [15]. In fact, we show that the same performance can be obtained even when the Markovian and HMM structures are ignored. In other words, the Markov chain and HMM assumptions are not necessary for the performance of the MMD-CUSUM test. Our work even extends to the α/βmixing processes, which have never been considered for the MMD-CUSUM test previously.

More broadly, our study falls under the umbrella of the quickest change detection (QCD) theory [23]. Studies on the QCD problem can be split into two categories: the Bayesian and minimax formulation, depending on the assumption of the change point. The Bayesian formulation, pioneered by [24, 25], places a prior on the distribution of the change point (usually a geometric distribution). Whereas the minimax formulation, first considered by [26], assumes the change point is unknown and deterministic. Under both formulations, the different notions of detection delay are minimized while constrained on the probability of false alarm or the false alarm rate (1/ARL). A well-known Bayesian QCD formulation is Shiryaev's problem [24], which seeks the stopping rule that minimizes the average detection delay (under the change point prior) while constrained on the probability of false alarm. The minimax formulations include Lorden's problem [26] and Pollak's problem [27], where the former minimizes the worst-case average delay and the latter

minimizes the conditional average delay while both contained on the false alarm rate.

Although the CUSUM statistic was first proposed as a heuristic for the minimax formulation under i.i.d. setting by [6], strong optimality properties have been shown for CUSUM statistic under various settings. Under the i.i.d. setting, exact optimality was shown by [28, 29] for Lorden's problem. For general non-i.i.d. settings, [20] has shown that an extension of the CUSUM statistic achieves the information-theoretic lower bound on the conditional average delay (as well as the worst case delay) asymptotically as the false alarm rate goes to 0.

However, the optimality result mentioned previously requires specific knowledge of the pre-and post-change distributions. Furthermore, the QCD problems are intractable for general stochastic processes due to the lack of problem structure [20]. Thus, the numerous studies on QCD for noni.i.d settings [1, 20, 30, 31, 2, 19, 3, 32, 33, 34, 35, 36] cannot be easily converted to the completely data-driven setting.

B. Contributions

As a non-parametric model-free change point detection procedure, the MMD-CUSUM test exhibits great potential in completely data-driven applications where distributional assumptions may be difficult to verify. Our performance guarantees under general mixing conditions establish its robustness under dependent samples and further strengthen its capability as a model-free testing scheme. The mixing conditions considered in this paper not only subsume the i.i.d., Markov chain, and HMM settings but also greatly expand beyond those appearing in previous studies on the performance of the MMD-CUSUM test. Our results indicate that the Markovian or HMM structures are not necessary for the strong performance of the MMD-CUSUM test.

Additionally, we provide the first performance guarantee for the MMD-CUSUM test under exponentially α/β and fast ϕ-mixing processes. Note that stationary exponentially βmixing processes include the geometrically ergodic Markov chains as a special case, which violates Doeblin's condition [37, page 402]. In stark contrast, Doeblin's condition is the core assumption for the performance analysis of the MMDCUSUM test in [5] and [22].

II. BACKGROUND

In this section, we introduce the necessary background for our discussion. Section A collects the usual facts about reproducing kernel Hilbert space (RKHS) and maximum mean discrepancy (MMD). Section B presents the two notions of mixing used to obtain the main results. Our standard reference is [38] for RKHS and [15] for mixing processes.

A. RKHS and MMD

Let $(\mathrm{X},\mathrm{X},\mathrm{P})$ be a measure space with Borel σ-algebra X and σ-finite measure P. Let $\mathrm{P}(\mathrm{X})$ denote the set of probability measures over the σ-algebra X. The supremum norm of f is written as $\|f\|_\infty := \sup_{x\in\mathrm{X}} |f(x)|$ and its span is written as $\mathrm{span}(f) := \sup_{x,x'\in\mathrm{X}} |f(x) - f(x')|$.

A *reproducing kernel Hilbert space* (RKHS) $\mathrm{H}(\mathrm{X})$ on X with kernel $k : \mathrm{X}\times\mathrm{X} \to \mathrm{R}$ is a Hilbert space of real-valued functions on X equipped with inner product $\langle\cdot,\cdot\rangle_{\mathrm{H}(\mathrm{X})}$. The corresponding Hilbert space norm $\|f\|^2_{\mathrm{H}(\mathrm{X})} = \|\|$ The kernel function k satisfies the reproducing property:

$$k(x,\cdot) \in \mathrm{H}(\mathrm{X}) \text{ and } \langle f(\cdot), k(x,\cdot)\rangle_{\mathrm{H}(\mathrm{X})} = f(x), \qquad \text{for } x \in \mathrm{X}.$$

The current paper relies on a particular application of RKHS — Hilbert space embeddings of probability measure. The Hilbert space embedding of μ under k is written as

$$\mathrm{U}(\mu)(\cdot) := \int_{\mathrm{X}} k(x,\cdot)\,d\mu,$$

where $\mathrm{U}(\mu)$ is also called the kernel mean embedding of μ. Suppose $\nu \in \mathrm{P}(\mathrm{X})$ is another probability measure. One can define a distance function between μ and ν using the Hilbert space metric between $\mathrm{U}(\mu)$ and $\mathrm{U}(\nu)$

$$\mathrm{MMD}_k(\mu,\nu) = \|\mathrm{U}(\mu) - \mathrm{U}(\nu)\|_{\mathrm{H}(\mathrm{X})},$$ which is known as the *maximum mean discrepancy* (MMD) [7]. The kernel k such that $\mathrm{MMD}_k(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$ for all $\mu,\nu \in \mathrm{P}(\mathrm{X})$ is call a *characteristic kernel* [39]. MMD_k with a characteristic kernel k is a metric on $\mathrm{P}(\mathrm{X})$.

MMD enjoys a computational advantage, compared with other probability distance functions, such as KL divergence [40] and total variation metric (Definition 7), that allows it to be easily estimated empirically for distributions on general domains [9, 10].

Let $X_i \sim \mu$ and $X_j' \sim \nu$ for $i = 1,\cdots,m$ and $j = 1,\cdots,n$. Define their empirical measures as $\hat{\mu}_m, \hat{\nu}_n$, respectively. The consistent estimation of the squared MMD is

$$\mathrm{MMD}_k^2(\hat{\mu}_m,\hat{\nu}_n) = \frac{1}{m^2}\sum_{1\le i,j\le N} k(X_i,X_j) + \frac{1}{n^2}\sum_{1\le i,j\le M} k(X_i',X_j') - \frac{2}{mn}\sum_{i,j} k(X_i,X_j'),$$

This was first used by [7] to propose the kernel two-sample test, and it is the core component of the MMD-CUSUM test studied in the current paper.

Throughout the paper, we assume the kernel k is *real-valued, measurable, characteristic, and bounded*, i.e., $\sup_{x\in\mathrm{X}} k(x,x) = k < \infty$. The boundedness ensures MMD_k is well-defined.

B. Mixing processes

The definitions of the mixing process require the following necessary notations. Consider the space of X-valued doubly infinite stochastic processes as $(\mathrm{X}^{\infty},\mathcal{X}^{\infty},\tilde{\mathrm{P}})$ where the indices of a process $X = \{X_i\}_{i\in\mathbb{Z}} \in \mathrm{X}^{\infty}$ are allowed to be $-\infty$ and ∞. For each index $t \in \mathbb{Z}$, let $\mathcal{X}_{t}^{\infty}$ denote the σ-algebra generated by $\{X_i\}_{i=t}^{\infty}$ and $\mathcal{X}_{-\infty}^{i}$ is the σalgebra generated by $\{X_t\}_{t=-\infty}^{i}$. We use $\mathcal{X}_{\tau}^{\tau'}$ to denote the σ-algebra generated by $\{X_i\}_{i=-\infty}^{\tau} \cup \{X_i\}_{i=\tau'+1}^{\infty}$. The marginal probability measure on $\{X_i\}_{i=t}^{\infty}$ is written as $\tilde{\mathrm{P}}_{t}^{\infty}$ and the joint probability measure on $\{X_i\}_{i=-\infty}^{t}$ as $\tilde{\mathbb{P}}_{-\infty}^{t}$.

With these notations, we have the definitions of α, β, and ϕ-mixing coefficients following [41, 15].

Definition 1 (α-mixing coefficient). The *α-mixing coefficient* [42] of a stationary process X is defined as $\alpha(n) = \sup_t \sup_{A\in\mathrm{X}_{-\infty}^{t},B\in\mathrm{X}_{t+n}^{\infty}} |\tilde{\mathrm{P}}(A\cap B) - \tilde{\mathrm{P}}(A)\cdot\tilde{\mathrm{P}}(B)|$.

X is called *α-mixing* if $\alpha(n) \to 0$ as $n \to \infty$.

The following β-mixing coefficient provides a stronger notion of decaying dependence. It can be shown that $2\alpha(n) \le \beta(n)$ [41].

Definition 2 (β-mixing coefficient). The *β-mixing coefficient* [16] of a stochastic process X is defined as $\beta(n) = \sup_t \sup_{C\in\mathcal{X}_{t+n}^{\infty}} |\tilde{\mathrm{P}}(C) - \tilde{\mathrm{P}}_{-\infty}^{t}\otimes\tilde{\mathrm{P}}_{t+1}^{\infty}(C)|$.

X is called *β-mixing* if $\beta(n) \to 0$ as $n \to \infty$. The β-mixing coefficient can be equivalently written as

$$\beta(n) = \sup_{t} \mathbb{E}_{\tilde{\mathbb{P}}_{-\infty}^{t}}\left[\sup_{C\in\mathcal{X}_{t+n}^{\infty}} |\tilde{\mathbb{P}}_{t+n}^{\infty}(C|\mathcal{X}_{-\infty}^{t}) - \tilde{\mathbb{P}}_{t+n}^{\infty}(C)|\right].$$

Comparing the second definition of β-mixing with the following definition of ϕ-mixing, we can see that $\beta(n) \le \phi(n)$.

Definition 3 (ϕ-mixing coefficient). The *ϕ-mixing coefficient* [18] of a stochastic process X is defined as

$$\phi(n) = \sup_{t} \sup_{B\in\mathcal{X}_{-\infty}^{t}} \left[\sup_{C\in\mathcal{X}_{t+n}^{\infty}} |\tilde{\mathbb{P}}_{t+n}^{\infty}(C|B) - \tilde{\mathbb{P}}_{t+n}^{\infty}(C)|\right].$$

X is called *ϕ-mixing* if $\phi(n) \to 0$ as $n \to \infty$.

We say X is *stationary* with respect to $\mu \in \mathrm{P}(\mathrm{X})$ if the one-dimensional marginal probability of X_i equals μ for $\forall i \in \mathbb{Z}$. For stationary processes, the supremum over t in the above definitions can be ignored, and one can set $t = 0$ without loss of generality. To maintain the simplicity of the presentation, we focus on *stationary* stochastic processes with α, β, and ϕ-mixing properties in the sequel. However, the results put forward in the current paper can be extended to asymptotically stationary processes, which is discussed in Section VI.

The decay rates of the mixing coefficients play an important role in our discussion. The following definitions introduce the *exponential α/β-mixing* condition and *fast ϕmixing*, which are used throughout the paper.

Definition 4 (exponential α/β-mixing). X is said to be exponential α or β-mixing, if the α or β-mixing coefficient satisfies

$$\alpha(n) \le \bar{\alpha}\exp(-cn^{\gamma}), \quad n \ge 1,$$
$$\text{or} \quad \beta(n) \le \bar{\beta}\exp(-cn^{\gamma}), \quad n \ge 1,$$

for $\bar{\alpha},\bar{\beta},\gamma,c > 0$.

Definition 5. [fast ϕ-mixing] X is said to be fast ϕ-mixing, if the ϕ-mixing coefficient satisfies

$$\Phi := \sum_{n=0}^{\infty} \phi(n) < \infty.$$

An exponentially decaying ϕ-mixing coefficient is certainly summable and thus is covered under the above definition. Definition 4 and 5 form the basic assumption on the mixing processes studied in the current paper.

To bridge the notions of mixing with RKHS, it is convenient to consider the following kernel mixing coefficient introduced in [43].

Definition 6 (kernel mixing coefficient). Let X be a stationary process with distribution μ. For $n \in \mathbb{N}$, define the kernel mixing coefficient as

$$\rho_k(n) = \left|\mathbb{E}\langle k(X_n,\cdot) - \mathbb{E}_{\mu}k(X,\cdot), k(X_0,\cdot) - \mathbb{E}_{\mu}k(X,\cdot)\rangle_{\mathcal{H}}\right|. \quad (1)$$

We denote the cumulative sum of the kernel mixing coefficient as $\Sigma_{\mu} := \sum_{n=0}^{\infty}\rho_k(n)$.

If we treat $\{k(X_i,\cdot)\}_{i\in\mathbb{Z}}$ as a sequence of Hilbert space valued stochastic process, then as shown by [44, Lemma 2.2] $\rho_k(n)$ can bounded by a constant multiple of the α-mixing coefficient, i.e., $\rho_k(n) \le 10\alpha(n)\bar{k}^2$. Thus, we get $\Sigma_{\mu} < \infty$ under the assumptions of exponential α-mixing, exponential β-mixing, and fast ϕ-mixing.

C. Examples of mixing processes

One notable example of ϕ-mixing processes is the uniformly ergodic Markov chain. A Markov chain is said to be uniformly ergodic if it is aperiodic and satisfies Doeblin's condition [37]. Thus, it is also called the Doeblin chain. A q-th order autoregressive (AR) process is ϕ-mixing if the Markov chain generated by stacking q consecutive states is a Doeblin chain. The ϕ-mixing coefficient decays exponentially for uniformly ergodic Markov chains, therefore satisfying the fast ϕ-mixing condition in Definition 5.

Examples of exponential β-mixing processes include V geometrically ergodic Markov chains. The Markov transition kernel $P : \mathrm{X} \times \mathrm{X} \to [0,1]$ with stationary distribution π is said to be V-geometrically ergodic if it satisfies

$\mathrm{TV}(P^n(x,\cdot),\pi) \le V(x)\rho^{\lfloor n/m \rfloor}$, for all n, (2) where $V : \mathrm{X} \to [1,\infty)$ is a measurable function, m is an constant integer, and $\rho \in [0,1)$. When V is bounded on X, it becomes the uniform ergodicity condition. From a dynamic system perspective, V-geometrically ergodic Markov chains subsume stable nonlinear systems with finite variance additive noise [see 45, Section 3.5]. The aforementioned examples all work as examples of exponential α-mixing processes. Additionally, measurable functionals of α, β, and ϕ-mixing processes are also α, β, and ϕ-mixing processes. The mixing coefficients are bounded by those of the original processes [45, Lemma 3.6].

III. PROBLEM FORMULATION

In this section, we first introduce the online change point detection problem and the commonly used performance metrics [see 19, 4]. Later, we discuss the proposed MMDCUSUM test and its properties.

In the sequel, we make the following assumption and restrict our attention to stochastic processes satisfying the exponential α/β-mixing and fast ϕ-mixing conditions in Definition 4, 5.

Assumption 1. The stochastic processes considered in what follows satisfy one of the three mixing conditions in Definition 4 and 5.

A. Online change point detection

The online change point detection problem is often formulated as a sequential two-sample test which has been widely considered in the past [6, 26, 21, 24]. Given a sequence of samples $\{X_i\}$ from a stationary mixing process X with distribution μ, at each time step, the following null and alternative hypotheses are proposed

$H_0 : \mu$ remains the same, $\quad H_1 : \mu$ has changed.

Test statistics are calculated using the samples collected up to the current time step. To detect the change quickly and accurately, one attempts to reject the null hypothesis H_0 via a threshold rule at every time step.

More formally, consider a stationary stochastic process $X = \{X_i\}_{i\in\mathrm{Z}} \in \mathrm{X}^\infty$ adapted to its natural filtration with unknown distribution μ. At some unknown but deterministic time index $\tau \in \mathrm{Z}$, we have $X_i \sim \mu$ for $0 \le i \le \tau$ and $X_i \sim \nu$ for $i \ge \tau + 1$, where $\mu,\nu \in \mathrm{P}(\mathrm{X})$ and $\mu \neq \nu$.

This can be conceptually thought of as having a separate and independent stochastic process $X' \in \mathrm{X}^\infty$ following unknown distribution ν running alongside X. From the outside, one can only observe X up to time τ, and at time τ, the observation is immediately switched to X'.

Suppose the null hypothesis is rejected at time $T(b)$, which is a stopping time adapted to the filtration $\{\mathrm{X}_{-i}{}_\infty\}_{i\in\mathrm{Z}}$ and a function of the threshold b. If we use E_∞ and E_0 to denote the expectation under H_0 and H_1 respectively, then the average-run-length ARL and the average-detection-delay ADD can be written in terms of the stopping time T as follows

$$\mathrm{ARL} = \mathrm{E}_\infty[T(b)] \qquad \text{and ADD} = \mathrm{E}_0[T(b)].$$

Unlike the Bayesian formulation, we assume the change point τ is unknown and deterministic, and thus we set $\tau = 0$ without loss of generality. ARL measures the robustness of the test against false alarms. Whereas ADD measures the quickness of the test in response to an abrupt change. The overall goal of online change point detection is to have a ARL that grows with b as fast as possible and a ADD that grows with b as slowly as possible.

B. MMD-CUSUM test

The MMD-CUSUM test is a sequential adaptation of the kernel two-sample test. Consider a bounded, measurable, characteristic, reproducing kernel $k : \mathrm{X} \times \mathrm{X} \to \mathrm{R}$. Denote the reference dataset as $\mathcal{D}_h = \{X'_i\}_{i=1}^h$ of size h. The detection algorithm processes the incoming data in blocks of size r, which is denoted as $\mathcal{B}_r(t) = \{X_i\}_{i=(t-1)r}^{tr-1}$ for an integer $t \ge 1$. Let $\hat{\nu}_h$ and $\hat{\mu}_r$ denote the empirical measure constructed using the dataset D_h and $\mathrm{B}_r(t)$. Define the MMD between these two empirical measures as

$$\begin{aligned}
&\mathrm{MMD}[\hat{\mu}_r,\hat{\nu}_h] \\
&= \Bigg(\frac{1}{r^2} \sum_{1\le n,m\le r} k(X_{(t-1)r+n-1}, X_{(t-1)r+m-1}) \\
&\quad + \frac{1}{h^2} \sum_{1\le n,m\le h} k(X'_n, X'_m) \\
&\quad - \frac{2}{rh} \cdot \sum_{\substack{1\le n\le r,\\ 1\le m\le h}} k(X_{(t-1)r+n-1}, X'_m) \Bigg)^{\frac{1}{2}},
\end{aligned} \tag{3}$$

At time step $i = t \cdot r$, the algorithm computes the following test statistic; otherwise, it collects the new observations and remains idle. Let integer $M \ge r$ be the minimum number of samples required to perform the test. We write the test statistics at time step i as

$$\hat{s}_{\lfloor i/r \rfloor} = \max_{1\le n\le t} s_{n:\lfloor i/r \rfloor}, \tag{4}$$

$$s_{n:\lfloor i/r \rfloor} = \sum_{t=n}^{\lfloor i/r \rfloor} \left\{ \mathrm{MMD}[\hat{\mu}_r, \hat{\nu}_h] - \Delta \right\},$$

where $\Delta > 0$ is a tunable parameter that keeps the summand slightly blew 0 under the null hypothesis. The corresponding stopping rule with threshold b and M minimum samples is written as

$$T(b,M) = r \cdot \inf\left\{ t \ge M/r : \hat{s}_t > b \right\}. \tag{5}$$

We make the following remarks regarding the above MMD-CUSUM statistics.

a: Convergence of Empirical MMD

To correctly configure the offset parameter Δ, we need to determine the envelope of the deviation of the empirical MMD from the true one. The result collected in the following lemma shows that the estimation error is bounded by a term diminishing in the sample size plus a small margin, almost surely for all three mixing conditions. Note that the empirical MMD can be equivalently written as the MMD between empirical measures. For probability measures μ and ν, we write $\mathrm{MMD}_d(\mu,\nu)$ as $\mathrm{MMD}(\mu,\nu)$ where $\hat{\mu}_r,\hat{\nu}_h$ are empirical measures of μ and ν with r and h samples, respectively.

Lemma 1. *Let X and X' be two independent processes with stationary distribution μ and ν satisfying the mixing conditions introduced before. Given $\delta > 0$, there exist constant $C(r,h)$ such that the following holds almost surely for sufficiently large h,*

$$\left|\mathbb{E}_{[\mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h)|\mathrm{D}_h]} - \mathrm{MMD}(\mu,\nu)\right| \le C(r,h) + \delta,$$

where $C(r,h) = O\left(\sqrt{\frac{1}{r}} + \sqrt{\frac{\log\log h}{h}}\right)$ and $\mathrm{E}[\cdot|\mathrm{D}_h]$ denotes the expectation taken over the randomness in $\hat{\mu}_r$ conditioned on the reference dataset D_h.

Proof:

Applying triangle inequality, we get the following two expressions:

$$\mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h) - \mathrm{MMD}(\mu,\nu) \le \mathrm{MMD}(\hat{\mu}_r,\mu) + \mathrm{MMD}(\nu,\hat{\nu}_h),$$

$$\mathrm{MMD}(\mu,\nu) - \mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h) \ge -\mathrm{MMD}(\hat{\mu}_r,\mu) - \mathrm{MMD}(\nu,\hat{\nu}_h).$$

Let us consider the first inequality above, and the other one follows similarly. Suppose we take expectation over the randomness of $\hat{\mu}_r$, and due to independence we have,

$$\mathrm{E}[\mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h)|\mathrm{D}_h] - \mathrm{MMD}(\mu,\nu) \le \mathrm{E}[\mathrm{MMD}(\hat{\mu}_r,\mu)] + \mathrm{MMD}(\nu,\hat{\nu}_h).$$

On the right hand side, the term $\mathrm{E}_X[\mathrm{MMD}(\hat{\mu}_r,\mu)] \le \sqrt{\frac{1+2\Sigma_\mu}{r}}$ by Lemma 7.1 of [43] for all $r > 0$ and X which satisfies $\Sigma_\mu < \infty$. It remains to bound $\mathrm{MMD}(\nu,\hat{\nu}_h)$ for a particular $\hat{\nu}_h$. Observe that

$$\mathrm{MMD}(\nu,\hat{\nu}_h) = h^{-1}\left\|\sum_i^r H_i\right\|_{\mathcal{H}_k},$$

where $\{H_i = k(X_i,\cdot) - \mathrm{E}_\nu k\}$ is a Hilbert space valued stochastic process and $\{H_i\}$ enjoys the same mixing property as X, since H_i is a measurable function of X_i'. Thus, we can apply the law of iterated logarithm for Hilbert space valued α-mixing processes [44, Theorem 6] or [46, Theorem 2] to conclude there exists constant $c_0 > 0$ such that almost surely

$$\limsup_{r\to\infty}\left\|\sum_{i=1}^r H_i\right\|_{\mathcal{H}_k} \le c_0\sqrt{h\log\log h}.$$

Note that the hypothesis of [44, Theorem 6] holds in our case under the assumption of bounded kernel k and exponential α/β-mixing and fast ϕ−mixing. Thus, there exists a constant $C(r,h) = O\left(\sqrt{\frac{1}{r}} + \sqrt{\frac{\log\log h}{h}}\right)$ such that $\mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h) - \mathrm{MMD}(\mu,\nu) \le C(r,h) + \delta$ for sufficiently large h. Similar, $\mathrm{MMD}(\mu,\nu) - \mathrm{MMD}(\hat{\mu}_r,\hat{\nu}_h)$ can bounded from below with $-C(r,h) - \delta$, and the proof is complete. ■

Lemma 1 indicates that under the null hypothesis (no change), the bias of empirical MMD is bounded by a positive quantity decaying at rate $o(r^{-1/2} + h^{-1/2}\log\log h)$ plus a small margin for sufficiently large reference data. To maintain a low value of the MMD-CUSUM statistics under the null hypothesis, it is necessary to apply a certain negative offset to the empirical MMD so that the cumulative sum in (4) does not blow up when change is absent which leads to the second remark regarding the parameter Δ.

b: Offset parameter Δ

We shall determine the appropriate range for the offset parameter Δ in (4) using Lemma 1. Note that Δ needs to be sufficiently large under the null hypothesis such that the MMD-CUSUM statistic does not blow up due to the estimation error of the empirical MMD. As suggested by Lemma 1, if Δ is strictly larger than $C(r,h)$, i.e., $\Delta \ge C(r,h) + \delta$ for some margin $\delta > 0$, then the empirical MMD is bounded by Δ almost surely for sufficiently large sample size. On the other hand, the upper bound for Δ appears under the alternative hypothesis (with post-change distribution ν). As we shall see in Theorem 3, Δ should be strictly less than $\mathrm{MMD}_k(\mu,\nu) - C(r,h) - \delta$ otherwise the ADD can be unbounded. To tune Δ in practice, one can simulate the pre-change scenario with different values of $\Delta \ge C(r,h) + \delta$ using the reference dataset. For each value of Δ, the ARL can be estimated with multiple runs of the experiment. Then, choose the smallest Δ that yields the acceptable ARL performance. Keeping the Δ small allows the MMD-CUSUM to achieve better ADD.

IV. MAIN RESULTS

In this section, we establish the detection performance of the MMD-CUSUM test using the metrics introduced in the

[1] Linear or sublinear dependency of sample size means a tail bound of $O(\exp(-g(n)\epsilon^2))$ where $g(n)$ grows linearly or sublinearly. By writing it this way, we assume the tail probability measures the event $\{\sum_{i=1}^n X_n - n\mathrm{E}X \ge n\epsilon\}$ where the deviation scales with sample size n.

previous section. The average-run-length ARL characterizes the average interval between false alarms, which is lowerbounded in Theorem 2. The average detection delay measures the quickness of the detection, and an upper bound is given in Theorem 3. The proofs are omitted due to the page limit, and they can be found in the Supplementary Material's Appendix.

Before we state the results, let us briefly summarize the technique we employed. Recall E_∞ denotes the expectation under H_0. We can expand the $\mathrm{E}_\infty[T(b,M)]$ as follows

$$\begin{aligned}
&\mathbb{E}_\infty[T(b,M)] = \sum_{t=1}^{\infty} \mathbb{P}_\infty[T(b,M) \geq t] \\
=&M + \sum_{l=M+1}^{\infty} \left(1 - \mathbb{P}_\infty\left\{\bigcup_{t=M+1}^{l} \{T(b,M) = t\}\right\}\right) \\
\geq&M + \sum_{l=M+1}^{\infty} \left(1 - \mathbb{P}_\infty\left\{\bigcup_{t=M+1}^{l} \bigcup_{k=1}^{t-M} \{s_{k:t} \geq b\}\right\}\right) \\
\geq&M + \sum_{l=M+1}^{\infty} \left(1 - \sum_{t=M+1}^{l} \sum_{k=1}^{t-M} \mathbb{P}_\infty\{s_{k:t} \geq b\}\right),
\end{aligned}$$

where union bound is applied to the last inequality. At this point, it suffices to obtain an upper bound on the tail probability $\mathrm{P}_\infty\{s_{k:t} \geq b\}$ using Proposition 4, 5. The tail probability bounds in Proposition 4, 5 offers simple explicit subGaussian decay rates with linear or sublinear dependency[1] on the sample size n inside the exponential. This kind of decay rate is necessary for our analysis as it dictates the scaling of ARL in threshold b. As we shall see in the theorem below, the slower decay rate of Proposition 5 causes the difference in ARL between exponential α/βmixing and fast ϕ-mixing processes.

We note that the existing concentration inequalities obtained for generic purposes are not well-suited for the task at hand. For example, the classic concentration inequalities for α-mixing, such as [45, Theorem 3.5], have tail bound with an additive term in addition to the common exponential term seen in the usual Hoeffding's inequality. When combined with our technique, it leads to a prohibitively cumbersome derivation of the ARL. The α-mixing concentration inequality in [47, Theorem 2] gives the tail bound on the relative deviation (scaled by variance) instead of the absolute deviation. The β-mixing results in [48] and the α-mixing results in [49] provide a subexponential bound of $O(\exp(-\epsilon))$ which is a weaker dependency on ϵ than we desired. The detailed discussion of the concentration inequalities we derived is postponed until the main results are introduced.

We now state the main result on the upper bound of ARL under the mixing condition described in Definition 4.

Theorem 2. *The average-run-length for test statistics* (4) *and stopping rule* (5) *under the null hypothesis has the following lower bounds.*

1) *Suppose X is α/β-mixing satisfying Definition 4, then*

$$\mathrm{ARL}[T(b,M)] \geq M - 1 + \exp\left(b^{\frac{\gamma}{\gamma+1}} \delta^{\frac{\gamma+2}{\gamma+1}}\right)(1 + o(1)), \quad (6)$$

2) *Suppose X is ϕ-mixing satisfying Definition 5, then*

$$\mathrm{ARL}[T(b,M)] \geq M - 1 + \exp(b\delta)(1 + o(1)), \quad (7)$$

where γ is defined in Definition 4, and $\delta > 0$ is defined in Lemma 1 and depends on Δ, h.

Proof:

Appendix D ■

Theorem 2 establishes the first ARL bound for the MMDCUSUM test under $\alpha/\beta/\phi$-mixing processes. The performance of the MMD-CUSUM test under α/β-mixing case has not been considered in the literature before, and Equation 6 provides the first exponential lower bound on the ARL. In previous studies, ϕ-mixing processes are considered in certain specific cases, such as the uniformly ergodic Markov chains [5] and hidden Markov models (HMM) [22]. Equation 7 generalizes the ARL bound therein to the broader ϕ-mixing processes without loss of performance. It also indicates that Markovian or HMM structures are not necessary for the exponential lower bond of the ARL.

The ARL bound in Equation 6 has a dependency on γ, which controls the mixing speed (Definition 4). This dependency on γ also is the result of applying the concentration bound in Proposition 5. Suppose the α or β-mixing coefficient has a decay rate of $O(\exp(-n))$, i.e., $\gamma = 1$, the ARL then achieves a $\Omega(\exp(b^{1/2}\delta^{3/2}))$ lower bound which is slighted degraded in terms of the threshold b compared to Equation 7.

Surprisingly, the ARL under the fast ϕ-mixing condition (Equation 7) achieves the $\Omega(\exp(b))$ lower bound (same as Markovian samples) while only requiring a summable ϕmixing coefficient. In comparison, the ARL lower bounds in [5] and [22] are obtained under the Doeblin's condition [37, page 402], which corresponds to exponential ϕ-mixing conditions.

To measure the quickness of the MMD-CUSUM test, we estimate the expected value of the stopping time $T(b,M)$ under the alternative (H_0). Recall that E_0 denotes the expectation under H_1. We can write the $\mathrm{E}_0[T(b,M)]$ as follows

$$\begin{aligned}
\mathrm{E}_0[T(b,M)] &= \sum_{t=1}^{\infty} \mathrm{P}_0[\hat{s}_t \leq b] \leq \sum_{t=1}^{\infty} \mathrm{P}_0[s_{1:t} \leq b] \\
&= \sum_{t=1}^{t_0} \mathrm{P}_0[s_{1:t} \leq b] + \sum_{t=t_0+1}^{\infty} \mathrm{P}_0[s_{1:t} \leq b],
\end{aligned}$$

where the first inequality is due to $s_{1:t} \leq \hat{s}_t$. Splitting the summation at t_0 and trivially bound the first term with t_0. With a certain choice of t_0, the second term can be shown to be ultimately negligible or

$o(1)$ compared to t_0 using the concentration inequality in Proposition 4 and 5.

Theorem 3. *Suppose X is a mixing process satisfies Definition 4 or 5 and pre and post change stationary distribution μ and ν satisfy* $\text{MMD}_k(\mu,\nu) > C(r,h) + \Delta + \delta$ *for some* $\delta > 0$. *The average-detection-delay for test statistics* (4) *and stopping rule* (5) *under the alternative hypothesis has the following upper bounds.*

$$\text{ADD}[T(b,M)] \leq \max\left\{M, \frac{b}{D(\mu,\nu) - \Delta - \delta}\right\}(1 + o(1)), \quad (8)$$

where $D(\mu,\nu) = \text{MMD}_k(\mu,\nu) - C(r,h)$, *and* $C(r,h)$ *is defined in Lemma 1.*

Proof:

Appendix E ■

Theorem 3 gives the first $O(b)$ upper bound on ADD under all three types of mixing conditions. Similar to the ARL lower bound, it was previously considered only under uniformly ergodic Markov chains and HMM. Our result shows that the Markovian or HMM structure is also not necessary for $O(b)$ upper bound on ADD.

Intuitively, the realization of the MMD-CUSUM statistics should track its mean, which is just $n\text{MMD}(\mu,\nu)$ for time n. Therefore, the threshold b should affect the average detection delay in a linear fashion. We note that the sufficient separation between μ and ν is required due to the estimation error of the empirical MMD as indicated by Lemma 1. This can be satisfied by choosing r,h sufficiently large and δ sufficiently small to ensure $\text{MMD}_k(\mu,\nu) - 2C(r,h) - 2\delta > 0$.

We now establish the concentration inequalities for the sum of bounded functions under mixing conditions. Proposition 4 is a Hoeffding-type inequality for ϕ-mixing processes with summable mixing coefficients. We provide a proof based on the martingale decomposition. Concentration inequality for exponential ϕ-mixing processes is obtained in [50] using an information inequality-based argument. The martingale-based method was used in [51] to study the concentration inequality of dependent random variables on countable spaces. Proposition 5 compliments the results therein by considering stationary ϕ-mixing processes on completely separable metric spaces. The sum of bounded functions of ϕ-mixing processes has a tight concentration bound that resembles that of i.i.d. random variables, which can be recovered by setting $\Phi = 0$.

Proposition 4. *Let X be a stationary ϕ-mixing process with coefficient satisfying Definition 5. Assume that f : X → R has bounded span and let $S_n = \sum_{i=0}^{i} f(X_i)$. Then for $\epsilon \geq 0$, it holds*

$$\mathbb{P}\left[S_n - \sum_{i=0}^{n-1}\mathbb{E}[f(X_i)] \geq n\epsilon\right] \leq \exp\left(-\frac{2n\epsilon^2}{(2\Phi+1)^2\text{span}(f)^2}\right),$$

where Φ is defined in Definition 5.

Proof:

Appendix B ■

Compared to the $O(\exp(-n\epsilon^2))$ tail bound in Proposition 4, the following concentration inequality for β-mixing processes has an $O(\exp(-n\epsilon\hat{}^2))$ tail bound where $n\hat{}$ grows sublinearly with the sample size n. The proof follows [47, Theorem 2] with the modification of replacing Bernstein's inequality with Hoeffding's Lemma (Lemma 10) to yield the desired result for our purpose.

Proposition 5. *Let X be a stationary β-mixing sequence with the coefficient satisfying Definition 4. Assume that f : Y → R has bounded span, i.e.,* $\text{span}(f) < \infty$, *and let $S_n = \sum_{i=0}^{n-1} f(X_i)$. Then for all $\epsilon \in (0, \text{span}(f))$, it holds*

$$\mathbb{P}\left[S_n - \sum_{i=0}^{n-1}\mathbb{E}[f(X_i)] \geq n\epsilon\right] \leq (1 + \beta/e)\exp\left\{-\frac{2\hat{n}\epsilon^2}{\text{span}(f)^2}\right\}_{-2},$$

where $n\hat{} = \lfloor n[(10n/c)^{1/(\gamma+1)}]^{-1}\rfloor$ and c,γ are defined in Definition 4.

Proof:

Appendix C ■

To our knowledge, the tail bound in the above form has not been considered previously. As opposed to the classic two-term version in [45, Theorem 3.5] and the relative error version in [47, Theorem 2], which can be difficult to be applied in our analysis, Proposition 5 streamlines the calculation of ARL and ADD in Theorem 2 and 3.

Compared to regular Hoeffding's inequality for bounded i.i.d. random variables [52], the exponent of the tail bound has a sublinear dependence on sample size due to the presence of $n\hat{}$. $n\hat{}$ is close to n when γ is large corresponding to a faster decaying β-mixing coefficient (Definition 4). This sublinear relation with respect to n is also reported by [49]

[2] A subGaussian bound on ϵ refers to a tail bound that looks like $O(\exp(-\epsilon^2))$. A subGaussian bound on ϵ refers to a tail bound that looks like $O(\exp(-\epsilon))$.

and [48] as well under exponential α and β-mixing conditions with $\gamma = 1$. They provided an $O(\exp(-n\epsilon/(\log n \log\log n)))$ tail bound, which is a faster rate in n compared to Proposition 5 with $\gamma = 1$. It is tempting to think that this tail could improve the lower bound of ARL in Theorem 2. However, the subexponential, instead of subGaussian[2], dependency on ϵ makes it not applicable to our proof. A similar concentration type inequality for αmixing processes is obtained in Proposition 14 following an analogous proof.

V. NUMERICAL SIMULATIONS

In this section, we apply the MMD-CUSUM test to a simulated stochastic process and verify the theoretical results. The stochastic process is generated by simulating a stable linear system $A \in \mathrm{R}^{4\times4}$ with an observations matrix $C \in \mathrm{R}^{2\times4}$.

Let $Z = \{Z_i\}_{i\in\mathbb{N}}$ denote the state process and $Y = \{Y_i\}_{i\in\mathbb{N}}$ denote the observation process. The system update equations can written as follows

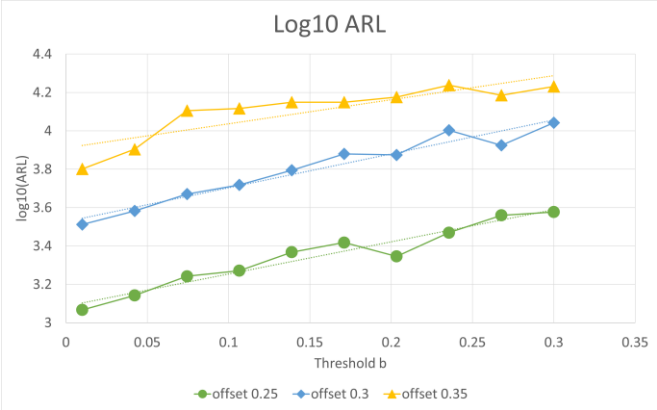
$$Z_{i+1} = AZ_i + W_i$$

$$Y_{i+1} = CZ_{i+1} + V_i,$$

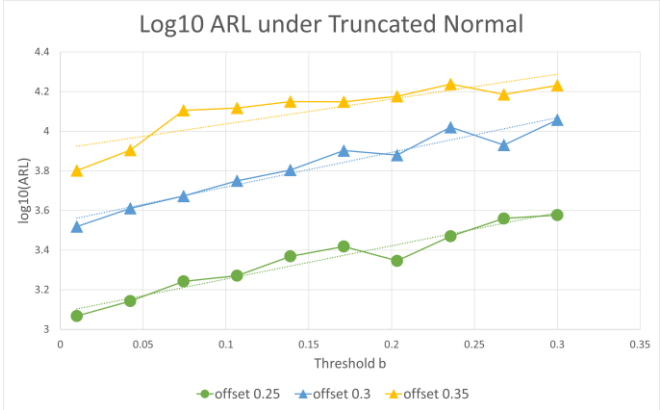
where A =[[0.96, 0.99, -0.88, 0.56],[0, 0.98, 0.75, -0.65],[0, 0, 0.97, 0.95], [0, 0, 0, 0.94]] and C =[[1, 0, 0, 0,], [0, 0, 0, 0], [0, 0, 1, 0], [0, 0, 0, 0]]. Randomness is introduced into the system through the actuation noise W_i and observation noise V_i where $W_i \overset{i.i.d.}{\sim} \mathrm{N}_1$ and $V_i \overset{i.i.d.}{\sim} \mathrm{N}_2$ for all i. In our experiments, N_1 and N_2 are two multivariate normal distributions. This is an example of a hidden Markov model (HMM). The state observation joint process (Z,Y) and the state process Z along are Markov chains; however, the observation process Y in general is not. The observation process of this system is exponential β-mixing. This can be deduced from the fact that the matrix A is stable and the noise has bounded variance [45, Section 3.5, page 100]. To obtain an exponential ϕ-mixing process from the observations, one can simulate the above system with truncated versions of N_1 and N_2.

The kernel chosen for the MMD-CUSUM test is the rational quadratic kernel $k_\sigma^{rq}(x,y) = (1+(2\sigma)^{-1}\|x-y\|^2)^{-\sigma}$ for $\sigma > 0$ instead the popular Gaussian RBF kernel $k_\sigma^{rbf}(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ for $\sigma > 0$. As demonstrated by [53], the rational quadratic kernel is favored over the Gaussian RBF kernel in GAN applications, which indicates its superior performance in separating probability distribution. We fix the parameter $\alpha = 1$ for all experiments. The reference dataset is obtained by recording Y for 10^4 steps under the pre-change configurations with an appropriate burn-in period applied to the samples to maintain stationarity. We estimate the ARL and ADD by taking the average of 50 independent experiments for each threshold. The experiments are performed under 3 different offsets to demonstrate the sensitivity of this parameter.

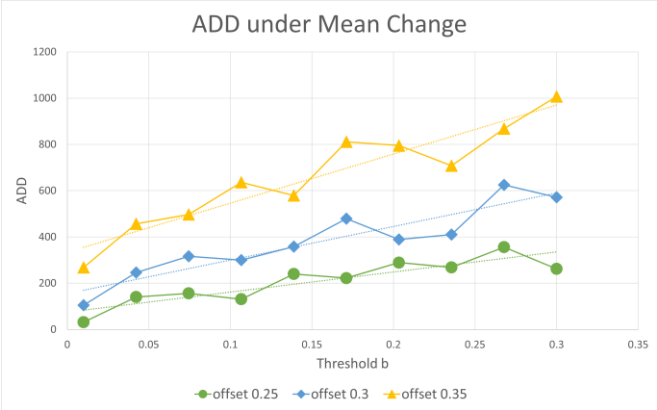
We apply abrupt changes to the noise distribution N_1 of the



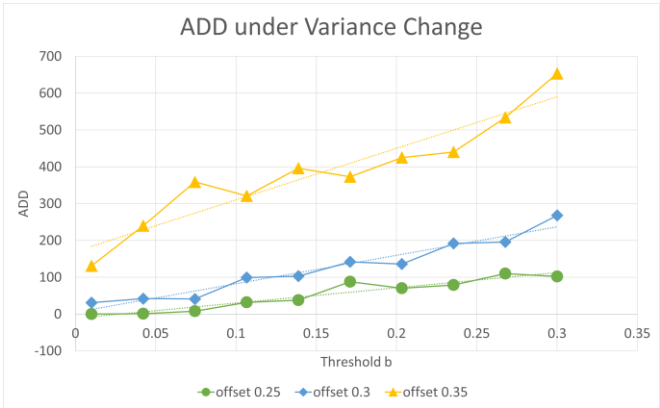
(a) $\log_{10}$(ARL) with Gaussian noise.



(b) $\log_{10}$(ARL) with truncated Gaussian noise.



(c) ADD under mean shift.



(d) ADD under variance change.

state process. The MMD-CUSUM test is applied to the observation process Y only. The noise observation poses an additional layer of challenge for the detector. To simulate an α/β-mixing process, it suffices to use the regular Gaussian noise. To simulate a ϕ-mixing, we sample from the same Gaussian distribution and reject the samples falling outside a $[-1,1]^4$ box. The random seeds are kept the same across the regular and the truncated cases to ensure comparability. The log(ARL) under both cases are shown in Figure Figure 1a and Figure 1b. To our surprise, the ARL under the regular Gaussian case maintains an exponential relationship with the threshold, which suggests the ARL bound for α/β-mixing process can be improved. We discuss the difficulty associated with this improvement in Section VI.

The ADD are estimated under regular Gaussian noise. We present the ADD under two cases: (i) mean shift $\mathcal{N}_1(0, 0.1I) \rightarrow \mathcal{N}_1'(0.01\mathbb{1}, 0.1I)$ (Figure Figure 1c) and (2) variance change $\mathcal{N}_1(0, 0.1I) \rightarrow \mathcal{N}_1'(0, 0.5I)$ (Figure Figure 1d). The ADD scales linearly with the threshold b, which corroborates our findings.

VI. DISCUSSION

A. Unbiased MMD estimator

The following unbiased estimator of the squared MMD, introduced in [7], can also be used to replace Equation 3. We write the unbiased estimator of the squared MMD between μ,ν using m samples from μ and n samples from ν as

$$\overline{\mathrm{MMD}}^2_k(\hat{\mu}_m, \hat{\nu}_n) = \frac{1}{m(m-1)} \sum_{1\le i,j\le N} k(X_i, X_j) + \frac{1}{n(n-1)} \sum_{1\le i,j\le M} k(X_i', X_j') - \frac{2}{nm} \sum_{i,j} k(X_i, X_j'),$$

where $X_i \sim \mu$ and $X_j' \sim \nu$ for $i = 1,\cdots,n$ and $j = 1,\cdots,m$. We abuse the notation here and write the empirical square MMD as the square MMD between empirical measures, although they are not equivalent to the unbiased estimator. Due to the unbiasedness, it is not always non-negative and thus should be directly plugged

into the partial sum with the square root. To adapt $\overline{\mathrm{MMD}}^2_k$ to the current framework, it suffices to obtain a consistency result such as Lemma 1, and the rest should follow. Consider two independent stochastic processes $X = \{X_i\}$ and $X' = \{X_i'\}$ with stationary distributions μ,ν and summable kernel mixing coefficients as in Definition 6. Suppose we use m consecutive

samples from X and n consecutive samples from Y. Then, we can bound the estimation bias caused by the dependency between samples as follows,

$$\begin{aligned}&\mathsf{E}[\overline{\mathrm{MMD}}_k^2(\hat{\mu}_m, \hat{\nu}_n)] - \mathrm{MMD}_k^2(\mu,\nu)\\ \le\ &\mathsf{E}[\overline{\mathrm{MMD}}_k^2(\hat{\mu}_m, \hat{\nu}_n)] - \mathsf{E}[\overline{\mathrm{MMD}}_k^2(\mu, \hat{\nu}_n)]\\ +\ &\mathsf{E}[\overline{\mathrm{MMD}}_k^2(\mu, \hat{\nu}_m)] - \mathrm{MMD}_k^2(\mu,\nu)\end{aligned}$$

where $\leq \frac{\Sigma_\mu}{m} + \frac{\Sigma_\nu}{n}$, the second inequality comes from [43, Lemma 7.1], Σ_μ, Σ_ν are defined in Definition 6, and the expectations are taken with respect to the randomness in the samples. After denoting $\overline{C}_{\mu,\nu}(m,n) := \frac{\Sigma_\mu}{m} + \frac{\Sigma_\nu}{n}$ and replacing $C_{\mu,\nu}(m,n)$ with $C_{\mu,\nu}(m,n)$ throughout the paper, the same set of results also holds for the CUSUM statistics defined with the unbiased estimator $\overline{\mathrm{MMD}}_k^2$.

B. Computation complexity

The time complexity at each time step is $O(rh)$ where r is the block size on the incoming data, and h is the size of the reference dataset. Compared to the overlapping block design in [5] with time complexity $O(r^2h)$, the non-overlapping block design here increases the speed at the expense of incurring a constant detection delay. The memory usage here is constant since only the current block and the reference data need to be stored. We present the implementation of the detection procedure in Algorithm 1.

C. Connection to HMM

Hidden Markov models (HMM) cover a wide array of real-world scenarios where the MMD-CUSUM test can be applied. For a comprehensive review of HMM, please refer to [54] and the references therein. Change point detection for HMM arises from the monitoring complex dynamic systems [55], such as communication networks [56], power plants [57], healthcare monitoring [58], manufacture process monitoring [59], distributed machine learning systems, etc.

For change detection, HMM can be treated as a mixing process. Consider a Markov chain $X := \{X_i\} \subset \mathsf{X}$ and its observation process $Y := \{Y_i\} \subset \mathsf{Y}$, where Y is a complete separable metric space with Borel σ-algebra Y. Define the observation kernel $Q_i : \mathsf{X} \times \mathsf{Y} \to [0,1]$ and $Q_i(X_i,A) = \mathbb{P}(Y_i \in A | \{Y_t\}_{t=-\infty}^{i-1}, \{X_t\}_{t=-\infty}^{i})$. Then, Y is $\alpha/\beta/\phi$-mixing as soon as X is $\alpha/\beta/\phi$-mixing [45, Theorem 3.12].

D. Asymptotic stationary processes

In practice, many mixing processes may not be strictly stationary but convergent towards a stationary distribution at a certain speed. For example, a Doeblin chain starts from an initial distribution that is different from its stationary distribution. Weak asymptotic stationarity was introduced in [60] to study the generalization bound of online algorithms. It combines the convergence to a stationary distribution and β-mixing into a single condition, which we choose not to include for the sake of simplicity. Instead, we provide a discussion on how to adapt Proposition 4 to asymptotic stationary processes in the supplementary materials. The adaption of Proposition 5 follows a similar argument. The intuition is that as long as the process converges sufficiently fast, the concentration of the partial sum will still hold. Thus, the same results on ARL and ADD can be extended to asymptotic stationary processes at no cost.

E. Obtain $\Omega(\exp(b))$ bound on ARL under α/β-mixing

As shown in Figure 1a and 1b, the difference in ARL between α/β-mixing and ϕ-mixing is minimal which might indicate a tighter $O(\exp(b))$ bound on ARL under α/β-mixing. This would be an improvement over the $\Omega(\exp(b^{1/\gamma}))$ in Theorem 2 where γ controls the mixing speed. However, the difficulties lie in the unavailability (to the best of our knowledge) of a subGaussian tail bound with linear dependency on the sample size n for stationary α/β-mixing processes. This bottleneck is also reported by a recent study [61] on the concentration of kernel density estimator with dependent data. Their findings are limited to ϕ-mixing processes due to the same issue. Circumventing this bottleneck might require significantly new techniques, which are left as future work.

VII. CONCLUSION

In this paper, we derive the ARL and ADD for the MMDCUSUM test under three stationary mixing conditions. Under the ϕ-mixing condition, the performance of the MMDCUSUM test is shown to match the i.i.d. case and the Markov chain case with uniform ergodicity. As a byproduct, we provide concentration inequalities of the partial sum of bounded functionals under α, β, and ϕ-mixing processes. To our knowledge, the concentration inequality in Proposition 5 and the proof of Proposition 4 are novel.

We note the limitations of this study and future directions as follows. MMD is known to have a poor separation between probability measures, with differences only in the high-frequency region [39]. The MMD-CUSUM test may experience performance degradation in such scenarios. A recent study [62] tackles this problem in the kernel twosample test setting via kernel spectral regularization. The spectral regularized kernel achieves the optimal minimax separation boundary, which results in an improved sample efficiency compared to the usual kernel two-sample test. Additionally, there have been several other exciting developments on kernel two-sample test [63, 64, 65]. It would be an interesting future direction to adapt those methods to the sequential test setting and analyze their performance.

Another limitation is that our technique does not exploit the finer structures produced by the max operator over the partial sum. The theory of extremes of random fields [66] provides handy tools to estimate the probability of events such as $\{\sup_{\theta\in\Theta} S_\theta > \epsilon\}$, where S_θ is the sum of n random variables in the random field and Θ is an index set, such as integers or real numbers. [12] has demonstrated the utility of this technique in the i.i.d. case and shown a sharp ARL bound of $O(\exp(b^2))$. However, the extension of this technique has yet to be explored in the non-i.i.d. cases. Additionally, leveraging the martingale property of the MMD-CUSUM statistics with an unbiased estimator and the non-asymptotic version of the law of logarithm for martingales [67] yields another possible route to establish the performance bounds. We plan to investigate these directions in the future.

APPENDIX

A. Auxiliary Facts

Definition 7 (Total variation metric). Let $\mathrm{B} := \{f : \|f\|_\infty \le 1, f : \mathrm{X} \to \mathrm{R}, f \text{ is X-measurable}\}$, the total variation metric between probability measures $\mu,\nu \in \mathrm{P}(\mathrm{X})$ is written as

$$\begin{aligned}\mathrm{TV}(\mu,\nu) &:= \frac{1}{2}\sup_{f\in\mathbb{B}}\left|\int f d\mu - \int f d\nu\right| \\ &= \sup_{A\in\mathcal{X}} |\mu(A) - \nu(A)|.\end{aligned}$$

Lemma 6 (Corollary D.2.5 in [68]). *Let $f : \mathrm{X} \to \mathrm{R}$ be an essentially bounded measurable function. For $\mu,\nu \in \mathrm{P}(\mathrm{X})$, we have*

$$|\mu(f) - \nu(f)| \le \mathrm{TV}(\mu,\nu)\mathrm{span}(f), \tag{9}$$

where $\mu(f)$, $\nu(f)$ denotes the expectation of f under μ, ν.

Lemma 7. *Suppose $\{X_i\}$ is a stationary ϕ-mixing process. Let $g : \mathrm{X}^\infty \to \mathrm{R}$ be an essentially bounded function and is measurable with respect to the σ-algebra $\mathcal{X}_{t+n}^\infty$. Then*

$$\left|\mathbb{E}[g(X_{t+n}^\infty)|x_{-\infty}^t] - \mathbb{E}[g(X_{t+n}^\infty)|y_{-\infty}^t]\right| \le 2\phi(n)\,\mathrm{span}(g),$$

where $x_{-\infty}^t y_{-\infty}^t$ are two realizations of the trajectory up to time t, and $\mathrm{span}(g) \le \|g\|_\infty$ when g is non-negative.

Proof:

$$\begin{aligned}&\left|\mathbb{E}[g(X_{t+n}^\infty)|x_{-\infty}^t] - \mathbb{E}[g(X_{t+n}^\infty)|y_{-\infty}^t]\right| \\ \le&\left|\mathbb{E}[g(X_{t+n}^\infty)|x_{-\infty}^t] - E[g(X_{t+n}^\infty)]\right| \\ +&\left|E[g(X_{t+n}^\infty) - \mathbb{E}[g(X_{t+n}^\infty)|y_{-\infty}^t]\right|\end{aligned}$$

$\le 2\phi(n)\mathrm{span}(g)$, where the first inequality is due to triangular inequality and the second is due to the Definition 3 of ϕ-mixing and Lemma 6. ∎

Lemma 8 (Corollary 2.2 in [45]). *Suppose $\{X_i\}$ is a stationary α-mixing process. Suppose $g_0,\ldots,g_l$ are essentially bounded functions, where g_i depends only on X_{ik}. Then*

$$\left|\mathbb{E}\left[\prod_{i=1}^l g_i\right] - \prod_{i=1}^l \mathbb{E}(g_i)\right| \le 4l\alpha(k)\prod_{i=1}^l \mathrm{span}(g_i),$$

where $\mathrm{span}(g_i) \le \|g_i\|_\infty$ when g_i is non-negative.

Lemma 9 (Theorem 2.1 in [45]). *Suppose $\{X_i\}$ is a stationary β-mixing process. Suppose $g_0,\ldots,g_l$ are essentially bounded functions, where g_i depends only on X_{ik}. Then*

$$\left|\mathbb{E}\left[\prod_{i=1}^l g_i\right] - \prod_{i=1}^l \mathbb{E}(g_i)\right| \le l\beta(k)\prod_{i=1}^l \mathrm{span}(g_i),$$

where $\mathrm{span}(g_i) \le \|g_i\|_\infty$ when g_i is non-negative.

Lemma 10 (Lemma 8.1 in [69]). *Let X be a random variable such that $a \le X \le b$ almost surely. Then, for $r > 0$,*

$$\mathrm{E}[\exp(r(X - \mathrm{E}X))] \le \exp[r^2(b-a)^2/8].$$

B. Proof of Proposition 4

We show a generalized version of Proposition 4 with timedependent functions in Proposition 11, that is, the partial sum S_n of interest is replaced by $\tilde{S}_n = \sum_{i=0}^{n-1} f_i(X_i)$ where f_i are potentially different. The key technique employed here is the martingale decomposition of the partial sum process generated by any stochastic process. In Lemma 12, we demonstrate the martingale decomposition. In Lemma 13, we establish that the martingale difference is bounded under the ϕ-mixing condition in Definition 5. Finally, we give the proof of Proposition 11 using the two supporting lemmas.

Proposition 11. *Let X be a stationary ϕ-mixing process with coefficient satisfying Definition 5. Assume that $f_i : \mathrm{X} \to \mathrm{R}$ has bounded span for $i = 0,\cdots,n-1$ and let $\tilde{S}_n = \sum_{i=0}^{n-1} f_i(X_i)$. Then for $\epsilon \ge 0$, it holds*

$$\mathbb{P}\left[\tilde{S}_n - \sum_{i=0}^{n-1}\mathbb{E}[f_i(X_i)] > n\epsilon\right] \le \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=-1}^{n-1} A_i^2}\right), \tag{10}$$

where Φ is defined in Definition 5 and $\{A_0,\cdots,A_{n-1}\}$ is defined in Equation 13.

First, we give the martingale decomposition of the partial sum process generated by any stochastic process.

Lemma 12. *Let $X = \{X_i\}$ be a stationary stochastic process on the probability space $(\mathrm{X}^\infty,\mathrm{X}^\infty,\tilde{\mathrm{P}})$. For $i \in \{0,\cdots,n-1\}$ and $n \in \mathbb{Z}_+$, let $f_i : \mathrm{X} \to \mathrm{R}$ be a essentially bounded function. Let $\tilde{S}_n := \sum_{i=1}^{n-1} f_i(X_i)$ be the partial sum. Let $\{\mathcal{X}_{-\infty}^i\}_{i=0}^{n-1}$ be the filtration generated by $\mathbf{X}$, i.e., X_{-i}*

$_{\infty} = \sigma(X_{\infty},\cdots,X_0,\cdots,X_i)$. *Then, there exists a martingale difference sequences* $\{D_i\}_{i=1}^{n-1}$ *adapted to* $\{\mathcal{X}_{-\infty}^i\}_{i=0}^{n-1}$ *such that*

$$\tilde{S}_n - \mathrm{XE}[f_i(X_i)]$$

$$= \mathrm{X} D_i + \mathrm{XE}[f_i(X_i)|\mathrm{X}_{--\infty 1}] - \mathrm{XE}[f_i(X_i)], \quad (11)$$

where the expectations are taken w.r.t. the stationary distribution.

Proof:

We employ the martingale decomposition technique introduced in Chapter 23 of [68]. For the following development, we use these short notations for tuples: $X_m^n :=$ $(X_m,\cdots,X_n)$ and $x_m^n := (x_m,\cdots,x_n)$, for $0 \le m \le n$.

Without loss of generality, we assume $\mathrm{E}[f_i(X_i)] = 0$ for $i = 0,\cdots,n-1$ to simplify the notation. For $i \in \{-1,0,\cdots,n-1\}$, we define

$$g_i(x_{-\infty}^i) := \mathrm{X} f_l(x_l) + \mathrm{X} \mathrm{E}[f_l(X_l)|x_{-\infty}^i], \quad (12)$$

where $g_{n-1}(x_{-\infty}^{n-1}) = \sum_{l=0}^{n-1} f_l(x_l)$ and $g_{-1}(x_{-\infty}^{-1}) = \sum_{l=0}^{n-1} \mathbb{E}[f_l(X_l)|x_{-\infty}^{-1}]$. With this definition, we observe that

$$g_{n-1}(x_{-\infty}^{n-1}) = \sum_{i=1}^{n-1} \left[g_i(x_{-\infty}^i) - g_{i-1}(x_{-\infty}^{i-1}) \right] + g_0(x_0),$$

and for $i \in \{0,\cdots,n-1\}$ and $x_{-\infty}^{i-1} \in \mathrm{X}_{-\infty}^i$,

$$g_{i-1}(x_{-\infty}^{i-1}) = \int_{\mathbb{X}^i} g_i(x_{-\infty}^{i-1}, x_i) d\tilde{\mathbb{P}}_i(\cdot|x_{-\infty}^i).$$

Recall that $\mathcal{X}_{-\infty}^i = \sigma(X_{-\infty},\cdots,X_0,\cdots,X_i)$, the above equation shows that $g_{i-1}(X_{-\infty}^{i-1}) = \mathbb{E}[g_i(X_{-\infty}^i)|\mathcal{X}_{-\infty}^{i-1}]$ $\tilde{\mathbb{P}}$ a.s. for $i \ge 0$. Thus, $\{g_i(X_{-\infty}^i)\}_{i=0}^{n-1}$ is a $\tilde{\mathrm{P}}$-martingale adapted to filtration $\{\mathcal{X}_{-\infty}^i\}_{i=-1}^{n-1}$. It follows the martingale decomposition of $\tilde{S}_n$ centered at $\sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]$,

$$\tilde{S}_n - \mathrm{XE}[f_i(X_i)] = g_{n-1}(X_{-n-\infty 1}) - \mathrm{XE}[f_i(X_i)]$$

$$= \mathrm{X} D_i + g_{-1}(X_{--\infty 1}) - \mathrm{XE}[f_i(X_i)],$$

where $\{D_i\}_{i=0}^{n-1} = \{g_i(X_{-\infty}^i) - g_{i-1}(X_{-\infty}^{i-1})\}_{i=1}^{n-1}$ is a martingale difference sequence. We arrive at the Equation 11. ■

Next, we show D_i is bounded by showing $x \mapsto g_i(x_{-\infty}^{i-1}, x)$ has bounded span for any $x_{-\infty}^{i-1} \in \mathrm{X}_{-\infty}^{i-1}$ and $i = 0,\cdots,n-1$.

Lemma 13. *Suppose X is a stationary ϕ-mixing process. For each $i \in \{-1,\cdots,n-1\}$, if we define*

$$A_i := 2\Phi\max\{\mathrm{span}(f_l) : i+1 \le l \le n-1\} + \mathrm{span}(f_i), \quad (13)$$

where $A_{-1} := 2\Phi\max\{\mathrm{span}(f_l) : 0 \le l \le n-1\}$. Then it holds that for $x_{-\infty}^i \in \mathrm{X}_{-\infty}^i$ and $i = -1,\cdots,n-1$

$$\inf_{x \in \mathrm{X}} g_i(x_{-\infty}^{i-1}, x) \le g_i(x_{-\infty}^i) \le \inf_{x \in \mathrm{X}} g_i(x_0^{i-1}, x) + A_i, \quad (14)$$

Proof:

It suffices to show Equation 14 holds. The first inequality is obvious. To show the second inequality, we pick arbitrary $x^* \in \mathrm{X}$ and we have

$$g_i(x_{i-\infty}) = \mathrm{X} f_l(x_l) + \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{i-\infty}]$$

$$\le \mathrm{X} f_l(x_l) + f_i(x_*) + \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{i-\infty}]$$

$$- \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{i--\infty 1}, x_*] + \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{i--\infty 1}, x_*]$$

$$+ \mathrm{span}(f_i).$$

Due to Lemma 6 (first inequality) and triangular inequality (second inequality), we can see that

$$\mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{-\infty}^{i}] - \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{i--\infty 1}, x_*]$$

$$\le \mathrm{X}\, \mathrm{span}(f_l)\, \mathrm{TV}(\tilde{\mathbb{P}}_l(\cdot|x_{-\infty}^i), \tilde{\mathbb{P}}_l(\cdot|x_{-\infty}^{i-1}, x^*))$$

$$\le \mathrm{X}\, \mathrm{span}(f_l) \left[\mathrm{TV}(\tilde{\mathbb{P}}_l(\cdot|x_{-\infty}^i), \mu) + \mathrm{TV}(\tilde{\mathbb{P}}_l(\cdot|x_{-\infty}^{i-1}, x^*), \mu) \right]$$

$$\le 2\Phi\max\{\mathrm{span}(f_l) : i+1 \le l \le n-1\},$$

where the last inequality follows from Lemma 7 and Definition 5. Thus, we have

$$g_i(x_{i0}) \le \mathrm{X} f_l(x_l) + f_i(x_*) + \mathrm{X}\, \mathrm{E}[f_l(X_l)|x_{-i-\infty 1}, x_*]$$

$$+ 2\Phi\max\{\mathrm{span}(f_l) : i+1 \le l \le n-1\} + \mathrm{span}(f_i)$$

$$\le g_i(x_0^{i-1}, x^*) + 2\Phi \max\{\mathrm{span}(f_l) : i+1 \le l \le n-1\}$$

+span(f_i)
$= g_i(x_0^{i-1}, x^*) + A_i$.

Since x^* is arbitrary, we obtain $g_i(x_0^i) \le \inf_{x^* \in \mathbb{X}} g_i(x_0^{i-1}, x^*) + A_i$, which completes the proof of Equation 14. ■

Finally, we estimate the tail bound using the classic Chernoff's bounding method [70]. In Remark 1, we note that a similar tail bound can be obtained for asymptotically stationary processes with a certain convergence (to stationary distribution) rate.

Proof:

(Proposition 11) By Lemma 12, we have the martingale decomposition of X as in Equation 11,

$$\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]$$

$$= \sum_{i=0}^{n-1} D_i + \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)|X_{-\infty}^{-1}] - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)],$$

where we abuse the notation and denote $D_{-1} = \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)|\mathcal{X}_{-\infty}^{-1}] - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]$ on the desired quantity. Taking the moment generating function on both sides of Equation 11 and applying the chain rule for conditional expectation recursively yield for $\theta \ge 0$

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\theta\left(\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right)\right)\right] \\
=&\mathbb{E}\left[\exp\left(\theta \sum_{i=-1}^{n-1} D_i\right)\right] \\
=&\mathbb{E}[\exp(\theta D_{-1})] \prod_{i=0}^{n-1} \mathbb{E}[\exp(\theta D_i)|\mathcal{F}_{i-1}] \qquad (15)
\end{aligned}$$

From lemma 13, we know that D_i lies in an interval of length A_i for all $i \in \{-1, \cdots, n-1\}$. By Hoeffding's Lemma [68, Lemma 23.1.4] for bounded martingale difference sequences, we have for $\theta \ge 0$,

$$\mathrm{E}[\exp(\theta D_{-1})] \le \exp(\theta^2 A_{-1}^2/8)$$

$$\mathrm{E}[\exp(\theta D_i)|\mathrm{F}_{i-1}] \le \exp(\theta^2 A_i^2/8), \text{ and}$$

plugging above into Equation 15 yields

$$\mathrm{E}\left[\exp\left(\theta\left(\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right)\right)\right] \le \exp\left(\frac{\theta^2}{8} \sum_{i=-1}^{n-1} A_i^2\right).$$

Applying Markov's inequality to the left-hand side, we have

$$\tilde{\mathbb{P}}\left[\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)] > n\epsilon\right]$$

$$\le \exp(-n\epsilon\theta)\mathbb{E}\left[\exp\left(\theta\left(\tilde{S}_n - \sum_{i=-1}^{n-1} \mathbb{E}[f_i(X_i)]\right)\right)\right]$$

$$\le \exp\left(-n\epsilon\theta + \frac{\theta^2}{8} \sum_{i=-1}^{n-1} A_i^2\right).$$

Picking $\theta = 4n\epsilon / \sum_{i=-1}^{n-1} A_i^2$ minimizes the right-hand side and yields

$$\tilde{\mathbb{P}}\left[\left(\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right) > n\epsilon\right] \le \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=-1}^{n-1} A_i^2}\right).$$

The tail probability of the other side can be bounded analogously. Therefore, we have

$$\tilde{\mathbb{P}}\left[\left|\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right| > n\epsilon\right] \le 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=-1}^{n-1} A_i^2}\right). \qquad (16)$$

This completes the proof after noticing that X is stationary

. One can use Chernoff's bounding method to obtain an exponential bound and hence $\mathrm{E}[f_i(X_i)] = \mu(f_i)$ for $i = 0, \cdots, n-1$. ■

Remark 1. *For asymptotically stationary processes, the marginal distribution of X_i differs from the stationary distribution μ but converges to μ as $i \to \infty$. To consider the tail probability of $\tilde{S}_n$ centered around $\sum_{i=0}^{n-1} \mu(f_i)$, we apply triangular inequality to (a) and Equation 16 to (b) and obtain*

$$\begin{aligned}
&\tilde{\mathbb{P}}\left[\left|\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right| > n\epsilon\right] \\
&\overset{(a)}{\le} \tilde{\mathbb{P}}\left[\left|\tilde{S}_n - \sum_{i=0}^{n-1} \mathbb{E}[f_i(X_i)]\right| + \sum_{i=0}^{n-1} \mathrm{span}(f_i) 2\mathrm{TV}(\tilde{\mathbb{P}}_i, \mu) > n\epsilon\right] \\
&\overset{(b)}{\le} 2\exp\left(-\frac{2[n\epsilon - \sum_{i=0}^{n-1} \mathrm{span}(f_i) 2\mathrm{TV}(\tilde{\mathbb{P}}_i, \mu)]^2}{\sum_{i=0}^{n-1} A_i^2}\right),
\end{aligned}$$

for $\epsilon \ge n^{-1} \sum_{i=0}^{n-1} \mathrm{span}(f_i) 2\mathrm{TV}(\tilde{\mathbb{P}}_i, \mu)$. Thus a similar tail bound can be obtained after assuming there exists a constant upperbound for $\sum_{i=0}^{n-1} \mathrm{span}(f_i) 2\mathrm{TV}(\tilde{\mathbb{P}}_i, \mu)$ for all n.

C. Proof of Proposition 5

We modify the proof of [47, Theorem 2] by replacing Bernstein's inequality with Hoeffding's Lemma (Lemma 10) in bounding the Φ_1 term to yield the desired result for our purpose.

Proof:

Given integer n, choose any integer $k \le n$ and define $l = \lfloor n/k \rfloor$. Let $p = n - kl$ and define the index sets I_i for $i = 1, 2, \ldots, k$ as follows

$$I_i = \begin{cases} \{i, i+k, \ldots, i+lk\} & 1 \le i \le p \\ \{i, i+k, \ldots, i+(l-1)k\} & p+1 \le i \le k. \end{cases}$$

Note that $\cup_i I_i = \{1, \ldots, n\}$ and within each set I_i the elements are pairwise separated by at least k. Let $G_i = f(X_i) - \mathbb{E}[f(X_1)]$, $T(i) = \sum_{j \in I_i} G_j$, and $p_i = |I_i|/n$ then

$$S_n - \mathbb{E}S_n = \sum_{i=1}^{n} G_i = \sum_{i=1}^{k} \sum_{j \in I_i} T(i)$$

$$\sum_{i=1}^{k} p_i = \frac{1}{n} \sum_{i=1}^{k} |I_i| = 1.$$

Now, we write the moment generating function of $\sum_{i=1}^{n} G_i/n$ for $r \ge 0$, which can be bounded as follows using the convexity of the exponential function,

$$\mathbb{E}\left[\exp\left(r \frac{\sum_{i=1}^{n} G_i}{n}\right)\right] \le \sum_{i=1}^{k} p_i \mathbb{E}\left[\exp\left(r \frac{T(i)}{|I_i|}\right)\right]. \quad (17)$$

We now bound the right-hand side in the following fashion. For $i = 1, 2, \ldots, k$, we have

$$\mathbb{E}\exp\left(r\frac{T(i)}{|I_i|}\right) = \mathbb{E}\prod_{j=1}^{|I_i|}\exp\left(r\frac{G_j}{|I_i|}\right)$$

$$\le \underbrace{\prod_{j=1}^{|I_i|}\mathbb{E}\exp\left(r\frac{G_j}{|I_i|}\right)}_{\Phi_1} + \underbrace{\mathbb{E}\prod_{j=1}^{|I_i|}\exp\left(r\frac{G_j}{|I_i|}\right) - \prod_{j=1}^{|I_i|}\mathbb{E}\exp\left(r\frac{G_j}{|I_i|}\right)}_{\Phi_2}. \quad (18)$$

For convenience, we denote the first term on the right-hand side of the above as Φ_1 and the second term as Φ_2. We bound them separately. Φ_1 can be estimated with Hoeffding's Lemma (Lemma 10) for bounded random variables. For $r > 0$,

$$\Phi_1 = \prod_{j=1}^{|I_i|}\mathbb{E}\left[\exp\left(r\frac{G_j}{|I_i|}\right)\right] \overset{(a)}{=} \left\{\mathbb{E}\left[\exp\left(r\frac{G_j}{|I_i|}\right)\right]\right\}^{|I_i|}$$

$$\overset{(b)}{\le} \exp\left[\frac{r^2 \mathrm{span}(f)^2}{8|I_i|}\right], \quad (19)$$

where (a) is due to stationarity and (b) comes from Lemma 10. Note that $|I_i| \ge l$ for $i = 1, 2, \ldots, k$, thus we have

$$\Phi_1 \le \exp\left[\frac{r^2 \mathrm{span}(f)^2}{8l}\right].$$

Φ_2 can be bounded by the β-mixing inequality in Lemma 9 and the exponential β-mixing condition (Definition 4).

$$\Phi_2 = \left|\mathbb{E}\left[\prod_{j=1}^{|I_i|}\exp\left(r\frac{G_j}{|I_i|}\right)\right] - \prod_{j=1}^{|I_i|}\mathbb{E}\left[\exp\left(r\frac{G_j}{|I_i|}\right)\right]\right|$$

$$\overset{(a)}{\le} \beta(k)(|I_i|-1)\prod_{j=1}^{|I_i|}\left\|\exp\left[\frac{rG_j}{|I_j|}\right]\right\|_\infty$$

$$\le \beta(k)(|I_i|-1)e^{r\mathrm{span}(f)}$$

$$\overset{(b)}{\le} e^{|I_i|-2}\beta(k)e^{-ck^\gamma}e^{r\mathrm{span}(f)}$$

$$\le \frac{\beta}{e^2}\exp\{|I_i| + r\mathrm{span}(f) - ck^\gamma\}, \quad (20)$$

where (a) is due to Lemma 9 and (b) is due to Definition 4 and the fact $\|I_i\| - 1 \le \exp(|I_i| - 2)$.

Now, we plug Equation (19) and (20) into Equation (18),

$$\mathbb{E}\left[\exp\left(r\frac{T(i)}{|I_i|}\right)\right] \le \exp\left[\frac{r^2\mathrm{span}(f)^2}{8l}\right] + \frac{\bar{\beta}}{e^2}\exp\{|I_i| + r\mathrm{span}(f) - ck^\gamma\}.$$

Since $|I_i|$ and k are free variables, we add some structure to simplify the right-hand side of the above. First, we require $4|I_i| \ge r\mathrm{span}(f)$ which leads to $\exp\{|I_i| + r\mathrm{span}(f) - ck^\gamma\} \le \exp\{5|I_i| - ck^\gamma\}$. Next, we require $\exp\{5|I_i| - ck^\gamma\} \le 1$, which holds if $5|I_i| \le ck^\gamma$. Since $|I_i| \le (n/k+1)$ and $n + k \le 2n$, it suffices to let $k = \lceil (10n/c)^{1/(\gamma+1)} \rceil$.

Then, we have

$$\mathbb{E}\exp\left(r\frac{T(i)}{|I_i|}\right) \le \exp\left[\frac{r^2\mathrm{span}(f)^2}{8l}\right] + \frac{\bar{\beta}}{e^2},$$

which holds for $0 < r \le \frac{4l}{\mathrm{span}(f)} \le \frac{4|I_i|}{\mathrm{span}(f)}$ for $i = 1, \ldots, k$. Plugging the above back to Equation 17 and using the fact that $\exp\left[\frac{r^2\mathrm{span}(f)^2}{8l}\right] \ge 1$, we have

$$\mathbb{E}\exp\left(r\frac{\sum_{i=1}^{n} G_i}{n}\right) \le (1 + \bar{\beta}/e^2)\exp\left[\frac{r^2\mathrm{span}(f)^2}{8l}\right].$$

Applying Markov's inequality, we have for $\epsilon > 0$

$$\mathbb{P}\left[S_n - \mathbb{E}S_n \ge n\epsilon\right] = \mathbb{P}\left[\exp\frac{r}{n}(S_n - \mathbb{E}S_n) \ge e^{r\epsilon}\right]$$

$$\le \frac{\mathbb{E}[\exp\frac{r}{n}(S_n - \mathbb{E}S_n)]}{e^{r\epsilon}}$$

$$\le (1 + \bar{\beta}/e^2)\exp\left[-r\epsilon + \frac{r^2\mathrm{span}(f)^2}{8l}\right]. \quad (21)$$

The right-hand side achieves minima w.r.t. r when $r = \frac{4\epsilon l}{\mathrm{span}(f)^2}$ which clearly satisfies when $\epsilon < \mathrm{span}(f)$.

Plugging the minimizer into Equation (21) yields

$$\mathbb{P}\left[S_n - \mathbb{E}S_n \geq n\epsilon\right] \leq (1+\bar{\beta}/e^2)\exp\left[-\frac{2l\epsilon^2}{\mathrm{span}(f)^2}\right].$$

Replacing l by $\hat{n} = \lfloor n/k \rfloor = \lfloor n\lceil(10n/c)^{1/(\gamma+1)}\rceil^{-1}\rfloor$ gives the

$$\mathsf{P}\left[S_n - \mathsf{E}S_n \geq n\epsilon\right] \leq (1+\bar{\beta}/e^2)\exp\left[-\frac{2\hat{n}\epsilon^2}{\mathrm{span}(f)^2}\right]$$

desired result ■

for $0 < \epsilon < \mathrm{span}(f)$.

A similar proof gives an analogous Hoeffding-type inequality for exponential α-mixing processes which is of independent interest. We document it here for completeness.

Proposition 14. *Let X be a stationary α-mixing sequence with the coefficient satisfying Definition 4. Assume that $f: \mathsf{Y} \to \mathsf{R}$ has bounded span, i.e., $\mathrm{span}(f) < \infty$. Let $S_n = \sum_{i=0}^{n-1} f(X_i)$. Then, for all $\epsilon \in (0, \mathrm{span}(f))$, it holds*

$$\mathsf{P}\left[S_n - \mathbb{E}S_n \geq n\epsilon\right] \leq (1+4e^{-2}\bar{\alpha})\exp\left\{-\frac{2n\hat{\epsilon}^2}{\mathrm{span}(f)^2}\right\}, \quad (22)$$

where $\hat{n} = \lfloor n\lceil(10n/c)^{1/(\gamma+1)}\rceil^{-1}\rfloor$ and c,γ are defined in Definition 4.

Proof:

The proof is the same as that of Proposition 5 except for the part where Φ_2 is estimated. In the α-mixing case, Φ_2 can be bounded by Lemma 8 and exponential α-mixing condition (Definition 4).

$$\begin{aligned}\Phi_2 &= \left|\mathbb{E}\left[\prod_{j=1}^{|I_i|}\exp\left(r\frac{G_j}{|I_i|}\right)\right] - \prod_{j=1}^{|I_i|}\mathbb{E}\left[\exp\left(r\frac{G_j}{|I_i|}\right)\right]\right| \\ &\overset{(a)}{\leq} 4\alpha(k)(|I_i|-1)\prod_{j=1}^{|I_i|}\left\|\exp\left[\frac{rG_j}{|I_j|}\right]\right\| \\ &\leq 4\alpha(k)(|I_i|-1)e^{r\,\mathrm{span}(f)} \\ &\overset{(b)}{\leq} e^{|I_i|-2}4\bar{\alpha}(k)e^{-ck^\gamma}e^{r\,\mathrm{span}(f)}\end{aligned}$$

$\leq 4e^{-2}\alpha^{-}\exp\{|I_i| + r\,\mathrm{span}(f) - ck^\gamma\}$, (23) where (a) is due to Lemma 8 and (b) is due to Definition 4 and the fact $\|I_i\| - 1 \leq \exp(|I_i| - 2)$. The rest of the proof follows that of Proposition 5. ■

D. Proof of Theorem 2

1) Case 1: β-mixing

When X is exponential β-mixing satisfying Definition 4, we show the lower bound of ARL as follows. For exponential α-mixing processes satisfying Definition 4, the proof follows the same procedure after replacing Proposition 5 with Proposition 14.

Proof:

To determine the upper bound for ARL, we condition on the fact that the change point τ is ∞. We use E_∞ and P_∞ to denote the expectation and the probability under $\tau = \infty$. For threshold $b > 0$, minimum burn-in period M, and stopping rule $T(b,M)$ in Equation (5), the ARL reads

$$\begin{aligned}\mathsf{E}_\infty[T(b,M)] &= \sum_{t=1}^{\infty}\mathsf{P}_\infty[T(b,M) \geq t] \\ &= M + \sum_{t=M+1}^{\infty}\mathsf{P}_\infty[T(b,M) \geq t] \\ &= M + \sum_{l=M+1}^{\infty}\mathsf{P}_\infty\Big[\bigcap_{t=M+1}^{l}\{T(b,M) = t\}\Big] \\ &\overset{(a)}{\geq} M + \sum_{l=M+1}^{\infty}\mathsf{P}_\infty\Big[\bigcap_{t=M+1}^{l}\{\hat{s}_t > b\}\Big] \\ &\geq M + \sum_{l=M+1}^{L}\mathsf{P}_\infty\Big[\bigcap_{t=M+1}^{l}\{s_{k:t} \geq b\}\Big] \\ &\overset{(b)}{\geq} M + \sum_{l=M+1}^{L}\Big[1 - \sum_{t=M+1}^{l}\mathsf{P}_\infty\{s_{k:t} \geq b\}\Big] \\ &= M + \sum_{l=M+1}^{L}\Big[1 - \sum_{t=1}^{l-M}\mathsf{P}_\infty\{s_{k:t+M} \geq b\}\Big], \quad (24)\end{aligned}$$

where (a) is due to the majorization of the event $\{T(b,M) = t\}$ to $\{\hat{s}_t > b\}$, (b) is due to the application of the union bound, and $M < L < \infty$ is an integer constant.

To further lower-bound the right-hand side of the above, we consider the tail probability in (24). Due to stationarity, we can study $\mathsf{P}_\infty\{s_{1:t} \geq b\}$ for some $t \geq 1$ without loss of generality. Suppose we pick the offset parameter $\Delta = C(r,h) + \delta$ for some $\delta > 0$. By Lemma 1, we known that $\mathsf{E}_\infty[s_{1:t}] = t\mathsf{E}_\infty[s(\mathrm{B}_r(1)] \in [-tC_{\mu,\mu}(r,h) - t\delta, -t\delta]$ almost surely for sufficiently large h.

Additionally, one can verify that $\{B_r(t)\}_{t=1}^{\infty}$ is β-mixing with coefficient $\tilde{\beta}(t) = \beta(tr)$.

By assumption, $\beta(tr)$ satisfies Definition 4 and $\tilde{\beta}(t) = \beta(tr) \le \bar{\beta}\exp(-cr^{\gamma}t^{\gamma})$, for $\beta,\gamma,c > 0$. Thus, we can apply Proposition 5 to obtain a tail probability bound on $s_{1:t}$. For $t \ge 1$,

$$\begin{aligned}
&\mathbb{P}_{\infty}[s_{1:t} > b] \\
&= \mathbb{P}_{\infty}\{s_{1:t} - \mathbb{E}_{\infty}[s_{1:t}] > b - \mathbb{E}_{\infty}[s_{1:t}]\} \\
&\overset{(a)}{\le} \frac{e^2}{e^2+\bar{\beta}} \exp\left\{-\frac{2\hat{t}(b - t\mathbb{E}_{\infty}[s(\mathcal{B}_r(1))])^2}{t^2\,\mathrm{span}(\widehat{\mathrm{MMD}}[\mathcal{B}_r(1),\mathcal{D}_h])^2}\right\} \\
&\overset{(b)}{\le} \frac{e^2+\bar{\beta}}{e^2} \exp\left[-\frac{\hat{t}(b+t\delta)^2}{t^2\bar{k}}\right] \\
&\overset{(c)}{\le} \frac{e^2+\bar{\beta}}{e^2} \exp\left\{\left[1-\left(\frac{r^{\gamma}c}{10}\right)^{\frac{1}{1+\gamma}}\right]\frac{(\frac{b}{\sqrt{t}}+\sqrt{t}\delta)^2}{t^{\frac{1}{1+\gamma}}\bar{k}}\right\}, \qquad (25)
\end{aligned}$$

where in (a) we apply Proposition 5 for the conditional probability under event A, in (b) we apply $\mathrm{span}(\mathrm{MMD}_d[B_r(1),D_h]) \le 2\bar{k}$ which can deduced from (3), and in (c) $\hat{t}$ is replace with its lower-bound,

$$\begin{aligned}
\hat{t} &= \left\lfloor \frac{t}{\lceil (10tr^{-\gamma}/c)^{\frac{1}{1+\gamma}} \rceil} \right\rfloor \ge \left\lfloor \frac{t}{(10tr^{-\gamma}/c)^{\frac{1}{1+\gamma}}+1} \right\rfloor \\
&\ge \left\lfloor \frac{t}{(10tr^{-\gamma}/c)^{\frac{1}{1+\gamma}}+t^{\frac{1}{1+\gamma}}} \right\rfloor \\
&\ge \left\lfloor \frac{t^{\frac{\gamma}{1+\gamma}}}{(10r^{-\gamma}/c)^{\frac{1}{1+\gamma}}} \right\rfloor \ge \frac{t^{\frac{\gamma}{1+\gamma}}}{(10r^{-\gamma}/c)^{\frac{1}{1+\gamma}}} - 1 \\
&\ge [(r^{\gamma}c/10)^{\frac{1}{1+\gamma}} - 1]t^{\frac{\gamma}{1+\gamma}}, \qquad (26)
\end{aligned}$$

assuming c,r,γ are sufficiently large such that $(r^{\gamma}c/10)^{\frac{1}{1+\gamma}} > 1$.

The right-hand side of (25) achieves its maximum when $t = t_* := (\frac{\gamma}{\gamma+2})\gamma\delta b$, which yields

$$\mathbb{P}_{\infty}[s_{1:t^*} > b] \le D, \qquad (27)$$

where

$$D := (1+\bar{\beta}e^{-2})\exp\left\{\frac{\xi(\gamma,c,r)}{\bar{k}}b^{\frac{\gamma}{\gamma+1}}\delta^{\frac{\gamma+2}{\gamma+1}}\right\}$$

$$\xi(\gamma,c,r) := \left[1-\left(\frac{r^{\gamma}c}{10}\right)^{\frac{1}{1+\gamma}}\right]\left(\frac{2}{\gamma}+2\right)^2\left(\frac{2}{\gamma}+1\right)^{-\frac{\gamma+2}{\gamma+1}}.$$

Note that $\xi(\gamma,c,r) < 0$ for sufficiently large c,r,γ.

Using (27) and stationarity, each of $\mathbb{P}_{\infty}[s_{k:t+M} > b]$ in (24) can be upper-bounded by $\mathbb{P}_{\infty}[s_{1:t^*} > b]$, which yields

$$\mathbb{E}_{\infty}[T(b,M)] \ge M + \sum_{l=1}^{L-M}\left\{1 - \sum_{t=1}^{l}\sum_{k=1}^{t}\mathbb{P}_{\infty}[s_{1:t^*} > b]\right\}$$

$$= L - D\sum_{l=1}^{L-M} l(l+1). \qquad (28)$$

The right-hand side achieves maxima when $L = L^*$, where L^* is obtained as the largest solution of

$$(L^*-M)(L^*-M+1) = 2/D.$$

Some simple calculation shows that

$$L^* = M - \frac{1}{2} + \frac{1}{2}\sqrt{1+8/D}.$$

Returning to (28), we have

$$\begin{aligned}
&\mathbb{E}_{\infty}[T(b,M)] \\
&\ge L^* - D\sum_{l=1}^{L^*-M} l^2 + l \\
&\ge L^* - D(L^*-M)(L^*-M+1)(L^*-M+2)/6 \\
&\ge M + \frac{1}{3}(L^*-M) - \frac{4}{3} \ge M - \frac{9}{6} + \frac{1}{6}\sqrt{1+8/D} \\
&\ge M + \frac{\sqrt{2}-9}{6} + \frac{2}{3}\sqrt{1/D},
\end{aligned}$$

where the second to last inequality uses the fact that $\sqrt{a+b} \ge \sqrt{a}/2 + \sqrt{b}/2$ for $a,b \ge 0$. Plugging the value of D yields the lower bound in (6). ■

2) Case 2: ϕ-mixing

When X is ϕ-mixing satisfying Definition 5, we show the lower bound of ARL as follows. For exponential α-mixing processes satisfying Definition 4, the proof follows the same procedure after replacing Proposition 5 with Proposition 14.

Proof:

Follow the same argument in Case 1 until the application of Proposition 5. Note that $\{\mathcal{B}_r(t)\}_{t=1}^{\infty}$ is ϕ-mixing with coefficient $\tilde{\phi}(t) = \phi(tr)$. Thus, $\{\mathcal{B}_r(t)\}_{t=1}^{\infty}$ satisfies Definition 5 with constant Φ as soon as X does. Then, we can apply Proposition 4 to obtain a tail bound on $s_{1:t}$. For $t \ge 1$, $\mathbb{P}_{\infty}[s_{1:t} > b] = \mathbb{P}_{\infty}\{s_{1:t} - \mathbb{E}_{\infty}[s_{1:t}] > b - \mathbb{E}_{\infty}[s_{1:t}]\}$

$$\le \exp\left\{-\frac{2(b - t\mathbb{E}_{\infty}[s(\mathcal{B}_r(1))])^2}{t(2\Phi+1)^2\,\mathrm{span}(\widehat{\mathrm{MMD}}[\mathcal{B}_r(1),\mathcal{D}_h])^2}\right\}$$

$$\leq \exp\left\{-\frac{(b+t\delta)^2}{t\bar{k}}\right\}.$$

The right-hand side of the above achieves its maximum when $t = t^* := b/\delta$, which yields

$$\mathbb{P}_\infty[s_{1:t^*} > b] \leq \exp\left[-\frac{4b\delta}{\bar{k}}\right], \tag{29}$$

The rest of the proof follows the same procedure in Case 1. Then, we have

$$\mathbb{E}_\infty[T(b,M)] \geq M + \frac{\sqrt{2}-9}{6} + \frac{2}{3}\exp\left[\frac{2b\delta}{\bar{k}}\right]$$

We arrive at the ARL lower-bound in (7). ■

E. Proof of Theorem 3

Case 1: β-mixing

When X is exponential β-mixing satisfying Definition 4, we show the upper bound of ADD as follows. Note the offset parameter $\Delta = C_{\mu,\mu}(r,h) + \delta$ as defined in the proof of Theorem 2 in Appendix D.

Proof:

Assuming the change point $\tau = 0$, we denote the probability and expectation under the alternative as P_0 and E_0, respectively. For $t \geq 1'$ $\mathrm{E}_0\{\mathrm{MMD}_{\mathrm{d}}[\mathrm{B}_r(t),\mathrm{D}_h]\} \geq \mathrm{MMD}_k(\mu,\nu) - C(r,h) - \delta$ almost surely for sufficiently large h according to Lemma 1. Let $D(\mu,\nu) := \mathrm{MMD}_k(\mu,\nu) - C(r,h)$. Now, we can write the average detection delay as follows

$$\begin{aligned}\mathbb{E}_0[T(b,M)] &= \sum_{t=1}^{\infty}\mathbb{P}_0[\hat{s}_t \leq b] \leq \sum_{t=1}^{\infty}\mathbb{P}_0[s_{1:t} \leq b] \\ &\leq \sum_{t=1}^{t_0}\mathbb{P}_0[s_{1:t} \leq b] + \sum_{t=t_0+1}^{\infty}\mathbb{P}_0[s_{1:t} \leq b] \\ &\leq \max\left\{M, \frac{b}{D(\mu,\nu)-\Delta-\delta}\right\} + \sum_{t=t_0+1}^{\infty}\mathbb{P}_0[s_{1:t} < b],\end{aligned} \tag{30}$$

where $t_0 = \max\left\{M, \frac{b}{D(\mu,\nu)-\Delta-\delta}\right\}$. In the rest of the proof, we aim to show that the second term on the right-hand side is ultimately bounded by a constant multiple of the first term, and the desired result is reached.

To bound the second term, we apply Proposition 5 to get the tail probability bound of $s_{1:t}$, similarly to the proof in Appendix D. For $t \geq t_0 + 1$, we have

$$\begin{aligned}&\mathrm{P}_0[s_{1:t} \leq b] \leq \mathrm{P}_0\{s_{1:t} - \mathrm{E}_0[s_{1:t}] \leq b - \mathrm{E}_0[s_{1:t}]\} \\ &\leq (1+\bar{\beta}/e^2)\exp\left\{-\frac{t^2\,\mathrm{span}(\mathrm{MMD}_r(\mathcal{B}_t,\mathcal{D}_h))^2}{2\hat{t}(\mathrm{E}_0[s_{1:t}]-b)^2}\right\} \overset{(a)}{\leq} (1+\bar{\beta}/e^2)\exp\left\{-\frac{\hat{t}[t(D(\mu,\nu)-\Delta-\delta)-b]^2}{t^2\bar{k}}\right\} \\ &\overset{(b)}{\leq} (1+\bar{\beta}/e^2)\exp\left\{-\frac{\hat{t}[t(D(\mu,\nu)-\Delta-\delta)-b]^2}{t^2\bar{k}}\right\} \\ &\overset{(c)}{\leq} (1+\bar{\beta}/e^2)\exp\left\{-\psi(c,\gamma)\frac{[t(D(\mu,\nu)-\Delta-\delta)-b]^2}{t^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\},\end{aligned}$$

where $\mathrm{span}(\mathrm{MMD}_{\mathrm{d}}[\mathrm{B}_r(t),\mathrm{D}_h]) \leq 2\sqrt{\bar{k}}$ can be deduced from Equation (3), (a) follows from Proposition 5 and $\hat{t} = \lfloor t\lceil(10t/c)^{1/(\gamma+1)}\rceil^{-1}\rfloor$, (b) uses stationarity of the post-change process and the conditioning on A', and (c) follows from the relation in Equation (26) and $\psi(c,\gamma) := (10/c+1)^{-\frac{1}{1+\gamma}}-1$. After magnifying the exponential term by setting b to $t_0(D(\mu,\nu)-\Delta-\delta)$ and splitting the summation at $\bar{t} := \lceil t_0^{\sigma}\rceil$ for some $\sigma \in (\frac{2+\gamma}{2(1+\gamma)}, 1)$, the second term on the right-hand side of Equation (30) becomes

$$\begin{aligned}&\frac{1}{1+\bar{\beta}/e^2}\sum_{t=t_0+1}^{\infty}\mathbb{P}_0 \\ &\leq \sum_{t=t_0+1}^{\infty}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2(t-t_0)^2}{t^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} \\ &= \sum_{t=1}^{\bar{t}-1}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2 t^2}{(t+t_0)^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} \\ &+ \sum_{t=\bar{t}}^{\infty}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2 t^2}{(t+t_0)^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} \\ &\leq (\bar{t}-1)\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2}{(1+t_0)^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} \\ &+ \sum_{t=\bar{t}}^{\infty}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2 t^{\frac{\gamma}{1+\gamma}}}{(1+t^{\frac{1}{\sigma}-1})^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\},\end{aligned}$$

where the first term is magnified by fixing $t = 1$ for all summands, and the second term is magnified by majorizing the denominator inside the exponential for each summand via $(1+t_0/t)^{\frac{2+\gamma}{1+\gamma}}\bar{k} \leq (1+\bar{t}^{\frac{1}{\sigma}}/t)^{\frac{2+\gamma}{1+\gamma}}\bar{k} \leq (1+t^{\frac{1}{\sigma}-1})^{\frac{2+\gamma}{1+\gamma}}\bar{k}$.

At this point, we can compare the growth rate of the two terms above and t_0, which yields

$$\begin{aligned}\lim_{t_0\to\infty}\frac{(\bar{t}-1)}{t_0}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2}{(1+t_0)^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} &= 0, \\ \lim_{t_0\to\infty}\sum_{t=\bar{t}}^{\infty}\exp\left\{-\psi(c,\gamma)\frac{(D(\mu,\nu)-\Delta-\delta)^2 t^{\frac{\gamma}{1+\gamma}}}{(1+t^{\frac{1}{\sigma}-1})^{\frac{2+\gamma}{1+\gamma}}\bar{k}}\right\} &= 0,\end{aligned}$$

where both equations above follow from $\sigma \in (\frac{2+\gamma}{2(1+\gamma)}, 1)$. Thus, we have

$$\lim_{t_0\to\infty}\frac{1}{t_0}\sum_{t=t_0+1}^{\infty}\mathbb{P}_0[s_{1:t} \leq b] = 0. \tag{31}$$

We have reached the desired result in Equation (8) after combining Equation (30) and (31). ■

$$\leq \sum_{t=t_0+1}^{\infty} \exp\left\{ - \frac{2(\mathbb{E}_0[s_{0:t}] - b)^2}{t(2\Phi+1)^2 \mathrm{span}(\widehat{\mathrm{MMD}}[\mathcal{B}_r(t), \mathcal{D}_h])^2} \right\}$$

Case 2: ϕ-mixing

When X is ϕ-mixing satisfying Definition 5, the upper bound of ADD is shown to follow Equation 8 using the same recipe as in Appendix E. Using Proposition 4, the second term on the right-hand side of Equation 30 can also be proven ultimately negligible compared to the first term.

Proof:

We shall directly start bounding the second term on the right-hand side of Equation 30 using Proposition 4, which is written as

$$\begin{aligned}
&\sum_{t=t_0+1}^{\infty} \mathbb{P}_0[s_{1:t} \leq b] \\
&= \sum_{t=t_0+1}^{\infty} \mathbb{P}_0\left\{ s_{1:t} - \mathbb{E}_0[s_{1:t}] \leq b - \mathbb{E}_0[s_{1:t}] | A' \right\} \\
&\leq \sum_{t=t_0+1}^{\infty} \exp\left\{ - \frac{[t(D(\mu,\nu) - \Delta - \delta) - b]^2}{t\bar{k}} \right\}.
\end{aligned}$$

If we magnify the exponential term by setting b to $t_0(D(\mu,\nu)-\Delta-\delta)$ and split the summation at $\bar{t} := \lceil t_0^{2/3} \rceil$, then it becomes

$$\begin{aligned}
&\sum_{t=t_0+1}^{\infty} \left\{ \mathbb{P}_0[s_{1:t} \leq b] \right\} \\
&\leq \sum_{t=t_0+1}^{\infty} \exp\left\{ - \frac{(D(\mu,\nu) - \Delta - \delta)^2 (t - t_0)^2}{t\bar{k}} \right\} \\
&= \sum_{t=1}^{\bar{t}-1} \exp\left\{ - \frac{(D(\mu,\nu) - \Delta - \delta)^2 t^2}{(t+t_0)\bar{k}} \right\} \\
&+ \sum_{t=\bar{t}}^{\infty} \exp\left\{ - \frac{(D(\mu,\nu) - \Delta - \delta)^2 t^2}{(t+t_0)\bar{k}} \right\} \\
&\leq (\bar{t}-1) \exp\left\{ - \frac{(D(\mu,\nu) - \Delta - \delta)^2}{(1+t_0)\bar{k}} \right\} \\
&+ \sum_{t=\bar{t}}^{\infty} \exp\left\{ - \frac{(D(\mu,\nu) - \Delta - \delta)^2 t}{(1+t^{1/2})\bar{k}} \right\}.
\end{aligned}$$

At this point, we can easily verify that both terms on the right-hand side are ultimately negligible compared to t_0, and the proof is complete. ■

F. MMD-CUSUM Test Pseudocode

ACKNOWLEDGMENT

The authors gratefully acknowledge the insightful discussions with our collaborators from Cisco Systems on potential applications of this work, including Ashish Kundu, Jayanth Srinivasa, and Hugo Latapie. Hao Chen and Abhishek Gupta's work was supported by Cisco Systems grant GR127553. Yin Sun's work was supported in part by the NSF under grant No. CNS-2239677, and by the ARO under grant No. W911NF-21-1-0244. Ness Shroff's work has been supported in part by NSF grants: CNS-2312836, CNS2223452, CNS-2225561, CNS-2112471, CNS-2106933, a grant from the Army Research Office: W911NF-21-1-0244, and was sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0225. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints

Algorithm 1: MMD-CUSUM test

Data: Data stream $\{X_i\}$, reference data D_h of size h; empty buffer B_r of size r; $s_{\min}$ stores the min of the partial sum; pick Δ during calibration; pick threshold $b > 0$. $i,t \leftarrow 0$;

$\hat{s}_0, s_0, s_{\min} \leftarrow 0$;

$\mathrm{B}_r \leftarrow \emptyset$;

while $\hat{s}_t \leq b$ do $\mathrm{B}_r \leftarrow \mathrm{B}_r \cup X_i$;

  if $(i \bmod r) = 0$ then

    $s_{t+1} \leftarrow s_t + \mathrm{MMD}[\hat{\mu}_r, \hat{\nu}_h] - \Delta$; $s_{\min} \leftarrow \min\{s_{t+1}, s_{\min}\}$;

    $\hat{s}_{t+1} \leftarrow s_{t+1} - s_{\min}$;

    $\mathrm{B}_r \leftarrow \emptyset$; $t \leftarrow t+1$; end

  $i \leftarrow i+1$; end

References

[1] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 4, pp. 613–644, 1995.

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

[3] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.

[4] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 494–514, 2021.

[5] H. Chen, J. Tang, and A. Gupta, "Change detection of Markov kernels with unknown pre and post change kernel," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 4814–4820, IEEE, 2022. [6] E. S. Page,

"Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. Smola, "A kernel two-sample¨ test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[8] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos, "MMD gan: Towards deeper understanding´ of moment matching network," *Advances in neural information processing systems*, vol. 30, 2017.

[9] T. Gartner, "A survey of kernels for structured data,"¨ *ACM SIGKDD explorations newsletter*, vol. 5, no. 1, pp. 49–58, 2003.

[10] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Scholkopf,¨ *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[11] S. Li, Y. Xie, H. Dai, and L. Song, "M-statistic for kernel change-point detection," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[12] S. Li, Y. Xie, H. Dai, and L. Song, "Scan B-statistic for kernel change-point detection," *Sequential Analysis*, vol. 38, no. 4, pp. 503–544, 2019.

[13] W.-C. Chang, C.-L. Li, Y. Yang, and B. Poczos, "Kernel´ change-point detection with auxiliary deep generative models," in *International Conference on Learning Representations*, 2018.

[14] T. Flynn and S. Yoo, "Change detection with the kernel cumulative sum algorithm," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 6092–6099, IEEE, 2019.

[15] R. C. Bradley, "Basic properties of strong mixing conditions. A survey and some open questions," *Probability surveys*, vol. 2, pp. 107–144, 2005.

[16] A. N. Kolmogorov and Y. A. Rozanov, "On strong mixing conditions for stationary Gaussian processes," *Theory of Probability & Its Applications*, vol. 5, no. 2, pp. 204–208, 1960.

[17] R. L. Adler, "Ergodic and mixing properties of infinite memory channels," *Proceedings of the American Mathematical Society*, vol. 12, no. 6, pp. 924–930, 1961.

[18] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory of Probability & Its Applications*, vol. 7, no. 4, pp. 349–382, 1962.

[19] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," in *2013 Information Theory and Applications Workshop (ITA)*, pp. 1–20, IEEE, 2013.

[20] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[21] W. A. Shewhart and W. E. Deming, *Statistical method from the viewpoint of quality control*. Courier Corporation, 1986.

[22] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou, "Datadriven quickest change detection in hidden Markov models," in *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 2643–2648, IEEE, 2023.

[23] V. V. Veeravalli and T. Banerjee, "Quickest change detection," in *Academic press library in signal processing*, vol. 3, pp. 209–255, Elsevier, 2014.

[24] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[25] A. N. Shiryaev, *Optimal stopping rules*, vol. 8. Springer Science & Business Media, 2007.

[26] G. Lorden, "Procedures for reacting to a change in distribution," *The annals of mathematical statistics*, pp. 1897–1908, 1971.

[27] M. Pollak and A. G. Tartakovsky, "Optimality properties of the shiryaev-roberts procedure," *Statistica Sinica*, pp. 1729–1739, 2009.

[28] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *the Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[29] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *The Annals of Statistics*, pp. 1464–1469, 1990.

[30] C.-D. Fuh, "SPRT and CUSUM in hidden Markov models," *The Annals of Statistics*, vol. 31, no. 3, pp. 942–977, 2003.

[31] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic Bayesian theory of quickest change detection," *Theory of Probability & Its Applications*, vol. 49, no. 3, pp. 458–497, 2005.

[32] C.-D. Fuh and Y. Mei, "Quickest change detection and kullback-leibler divergence for two-state hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4866–4878, 2015.

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

[34] C.-D. Fuh and A. G. Tartakovsky, "Asymptotic Bayesian theory of quickest change detection for hidden Markov models," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 511–529, 2018.

[35] G. V. Moustakides, "Detecting changes in hidden Markov models," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2394–2398, IEEE, 2019.

[36] J. J. Ford, J. James, and T. L. Molloy, "Exactly optimal Bayesian quickest change detection for hidden Markov models," *Automatica*, vol. 157, p. 111232, 2023.

[37] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. USA: Cambridge University Press, 2nd ed., 2009.

[38] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[39] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Scholkopf, and G. R. Lanckriet, "Hilbert space¨ embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[41] P. Doukhan, *Mixing: properties and examples*, vol. 85. Springer Science & Business Media, 2012.

[42] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proceedings of the national Academy of Sciences*, vol. 42, no. 1, pp. 43–47, 1956.

[43] B.-E. Cherief-Abdellatif and P. Alquier, "Finite sample´ properties of parametric MMD estimation: robustness to misspecification and dependence," *Bernoulli*, vol. 28, no. 1, pp. 181–213, 2022.

[44] H. Dehling and W. Philipp, "Almost sure invariance principles for weakly dependent vector-valued random variables," *The Annals of Probability*, pp. 689–701, 1982.

[45] M. Vidyasagar, *Learning and Generalisation: With Applications to Neural Networks*. Springer Science & Business Media, 2002.

[46] F. Merlevede, "On a maximal inequality for strongly` mixing random variables in Hilbert spaces. Application to the compact law of the iterated logarithm.," *Annales de l'ISUP*, vol. LII, pp. 47–60, May 2008.

[47] B. Zou, L. Li, and Z. Xu, "The generalization performance of erm algorithm with strongly mixing observations," *Machine learning*, vol. 75, no. 3, pp. 275–295, 2009.

[48] J. T. Krebs, "A large deviation inequality for β-mixing time series and its applications to the functional kernel regression model," *Statistics & Probability Letters*, vol. 133, pp. 50–58, 2018.

[49] F. Merlevede, M. Peligrad, and E. Rio, "Bernstein` inequality and moderate deviations under strong mixing conditions," *arXiv preprint arXiv:1202.4777*, 2012.

[50] P.-M. Samson, "Concentration of measure inequalities for Markov chains and ϕ-mixing processes," *The Annals of Probability*, vol. 28, no. 1, pp. 416–461, 2000.

[51] L. A. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *The Annals of Probability*, vol. 36, no. 6, pp. 2126 – 2158, 2008.

[52] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *The collected works of Wassily Hoeffding*, pp. 409–426, 1994.

[53] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gret-´ ton, "Demystifying MMD GANs," *arXiv preprint arXiv:1801.01401*, 2018.

[54] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Transactions on information theory*, vol. 48, no. 6, pp. 1518–1569, 2002.

[55] P. Smyth, "Hidden Markov models for fault detection in dynamic systems," *Pattern recognition*, vol. 27, no. 1, pp. 149–164, 1994.

[56] K. Salamatian and S. Vaton, "Hidden Markov modeling for network communication channels," *ACM SIGMETRICS Performance Evaluation Review*, vol. 29, no. 1, pp. 92–101, 2001.

[57] K.-C. Kwon and J.-H. Kim, "Accident identification in nuclear power plants using hidden Markov models," *Engineering Applications of Artificial Intelligence*, vol. 12, no. 4, pp. 491–501, 1999.

[58] T. Al-ani, C. K. Karmakar, A. H. Khandoker, and M. Palaniswami, "Automatic recognition of obstructive sleep apnoea syndrome using power spectral analysis of electrocardiogram and hidden Markov models," in *2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 285–290, IEEE, 2008.

[59] Y. Li, H. Li, Z. Chen, and Y. Zhu, "An improved hidden Markov model for monitoring the process with autocorrelated observations," *Energies*, vol. 15, no. 5, p. 1685, 2022.

[60] A. Agarwal and J. C. Duchi, "The generalization ability of online algorithms for dependent data," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 573– 587, 2012.

[61] S. Arvanitis, "Concentration inequalities for kernel density estimators under uniform mixing," *Journal of the Korean Statistical Society*, pp. 1–10, 2023.

[62] O. Hagrass, B. K. Sriperumbudur, and B. Li, "Spectral regularized kernel two-sample tests," *arXiv preprint arXiv:2212.09201*, 2022.

[63] F. Biggs, A. Schrab, and A. Gretton, "MMDFUSE: Learning and combining kernels for twosample testing without data splitting," *arXiv preprint arXiv:2306.08777*, June 2023. arXiv:2306.08777 [cs, math, stat].

[64] A. Schrab, I. Kim, B. Guedj, and A. Gretton, "Efficient aggregated kernel tests using incomplete *u*statistics," *arXiv preprint arXiv:2206.09194*, Jan. 2023. arXiv:2206.09194 [cs, math, stat].

[65] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton, "MMD aggregated two-sample test," *arXiv preprint arXiv:2110.15073*, Aug. 2023. arXiv:2110.15073 [cs, math, stat].

[66] B. Yakir, *Extremes in random fields: a theory and its applications*. John Wiley & Sons, 2013.

[67] A. Balsubramani, "Sharp finite-time iteratedlogarithm martingale concentration," *arXiv preprint arXiv:1405.2639*, 2014.

[68] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov chains*. Springer, 2018.

[69] L. Devroye, L. Gyorfi, and G. Lugosi,¨ *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media, 2013.

[70] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.