Data-Driven Quickest Change Detection in Hidden Markov Models

Qi Zhang, Zhongchang Sun, Luis C. Herrera and Shaofeng Zou
Department of Electrical Engineering
University at Buffalo, the State University of New York
Email: qzhang48@buffalo.edu, zhongcha@buffalo.edu, lcherrer@buffalo.edu, szou3@buffalo.edu

***Abstract*—The problem of quickest change detection in hidden Markov models (HMMs) is investigated. A sequence of samples are generated from a HMM, and at some unknown time, the transition kernel and/or the emission probability of the HMM changes. The goal is to detect the change as soon as possible subject to false alarm constraints. The data-driven setting is investigated, where none of the pre-, post-change Markov transition kernels or the emission probabilities are known. In this paper, a kernel based data-driven algorithm is developed. Performance bounds on its average running length (ARL) to false alarm and worst-case average detection delay (WADD) are theoretically characterized, where the WADD is at most of the order of the logarithm of the ARL. Numerical results are provided to validate the performance of the proposed algorithm.**

***Index Terms*—Sequential Change Detection, Maximum Mean Discrepancy, Kernel Method, Non-i.i.d..**

I. Introduction

In the quickest change detection (QCD) [1]–[5] problem, a sequence of observations is taken from a stochastic process. At some unknown time (change-point), an event occurs and changes the data-generating distribution. The goal is to detect the change as quickly as possible subject to false alarm constraints. QCD finds a wide range of applications, e.g., fault detection in DC microgrids [6], early detection of epidemics [7] and signal processing in genetic area [8].

Existing studies are mostly limited to the setting where the observations are independent and identically distributed (i.i.d.) before and after the change, respectively. However, in a wide range of applications such as speech recognition [9] and molecular biology [10], samples are not independent. The general theory for the non-i.i.d. setting were developed in [11]–[15], where the normalized log likelihood ratio is assumed to be asymptotically stable, and the approaches are model-based. For the non-i.i.d. setting, theoretical contributions on the cases where samples are generated according to a Markov model were also developed in, e.g., [16]–[19]. Extensions of QCD from Markov models to hidden Markov models (HMMs) where the observed samples are generated from HMMs have also been investigated [19]–[22]. The above studies focused on the model-based setting, where the pre- and post-change distributions were assumed to be known exactly. However, the information about the pre- and post-change distributions may not be available in practice, especially when the change is of an unknown type.

In this paper, we focus on the data-driven setting of QCD in HMMs. Specifically, the samples are generated according to a HMM and at some unknown time, the transition kernel and/or the emission probability of the HMM change. The goal is to detect this change as quickly as possible subject to false alarm constraints. None of the pre-, post-change transition kernels or the emission probabilities are known, and only a sequence of observations following the pre-change HMM is available.

The problem of QCD in HMMs with a finite state space and known pre- and post-change distributions was studied in [20]. A CuSum procedure was proposed and the basic idea is to use the L_1-norm of products of Markov random matrices to update the log-likelihood function. Then the ratio of L_1-norm of products of Markov random matrices was shown to converge and the asymptotic optimality of the proposed CuSum algorithm can be proved. This approach however requires the state space to be finite so that the theory of Markov random matrices can be applied. The problem with the same setting was studied in [21], where a Shiryayev–Roberts–Pollak (SRP) procedure was proposed and the log-likelihood was written as the ratio of the L_1-norms of products of Markov random matrices. The asymptotic optimality of the proposed SRP algorithm was also proved. In [22], a computationally efficient algorithm was proposed for the same problem. The basic idea is to design a quasi-generalized likelihood ratio that admits a recursive update to avoid the infeasible computation of the likelihood ratio. The Kullback-Leibler divergence between the pre- and post-change stationary distributions was shown to converge and the average detection delay (ADD) and the average run length (ARL) to false alarm were theoretically characterized. However, the proposed recursive score scheme only works with HMMs with two states. In [19], the problem in the Bayesian setting was studied, where the change point is assumed to follow some prior distribution. The Shiryaev algorithm was shown to be asymptotically optimal.

All of the above studies focus on finite-state HMMs and require the knowledge of pre- and post-change transition kernels and emission probability, which cannot be directly applied to the data-driven setting in this paper. For the QCD problem in HMMs under the data-driven setting, the auto-regressive model with unknown parameters was studied in [23], where a data-driven method was developed to approximate the likelihood ratio. However, no theoretical performance analysis was provided.

In this paper, we develop a data-driven algorithm for QCD in HMMs, where none of the pre-, post-change transition kernels or the pre-, post-change emission probabilities are known. To the best of the authors' knowledge, this is the first data-driven algorithm for QCD in HMMs with theoretical performance guarantees. More specifically, We theoretically show that the lower bound on ARL is exponential in the threshold and the upper bound on WADD is linear in the threshold. Therefore, the WADD is at most in the logarithm of the ARL. This matches with the results under the model-based setting, and is of great practical importance as the number of samples taken to make a false alarm shall be exponential in the number of samples taken to detect the change. The major challenge in our analysis lies in quantifying the correlation overtime due to the HMM structure. The overall computational complexity of our algorithm at time t is $\mathcal{O}(mt)$, where m is the block size, and thus our algorithm is computationally efficient. Our simulation results also validate our theoretical finding that the WADD is in the logarithm of the ARL.

II. Problem Statement

Consider a Markov chain $\{X_t\}_{t=1}^{\infty}$ defined on a probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P})$. Denote by $P: \mathcal{X} \to \mathcal{P}(\mathcal{X})$ its transition kernel, where $\mathcal{P}(\mathcal{X})$ denotes the probability simplex on $\mathcal{X}$. The state X_t is not directly observable, and instead we observe a sequence $\{X'_t\}_{t=1}^{\infty}$ which is adjoined to the Markov chain $\{X_t\}_{t=1}^{\infty}$ such that $\{X_t, X'_t\}_{t=1}^{\infty}$ is a Markov chain and for any measurable set $A \subseteq \mathcal{X}$,

$$\mathbb{P}(X_t \in A|X_{t-1},...,X_1,X'_{t-1},...,X'_1) = P(X_t \in A|X_{t-1}),$$
$$\mathbb{P}(X'_t \in A|X_t,...,X_1,X'_{t-1},...,X'_1) = P'(X'_t \in A|X_t).$$

Here $P': \mathcal{X} \to \mathcal{P}(\mathcal{X})$ denotes an emission probability.

At some unknown time τ, the transition kernel P changes into $Q: \mathcal{X} \to \mathcal{P}(\mathcal{X})$ and the emission probability P' changes into $Q': \mathcal{X} \to \mathcal{P}(\mathcal{X})$. The goal is to detect the change as quickly as possible subject to false alarm constraints.

Let $\hat{T}$ be a stopping time and denote by $\mathbb{P}_\tau$ ($\mathbb{E}_\tau$) the probability measure (expectation) when the change happens at time τ and $\mathbb{P}_\infty$ ($\mathbb{E}_\infty$) the probability measure (expectation) when there is no change. In this paper, we focus on the data-driven setting, where P, P', Q, Q' are *unknown*. We assume that we have a reference sequence of observations $\{Y'_t\}_{t=1}^{\infty}$ generated from a HMM with transition kernel P and emission probability P'. Let Y_t be the hidden state. Denote by $\mathcal{F}_t$ the σ-field generated by $\{X_i, X'_i, Y_i, Y'_i\}_{i=1}^{t}$. The average running length (ARL) to false alarm and the worst-case average detection delay (WADD) of $\hat{T}$ are defined as follows [24]:

$$\mathrm{ARL}(\hat{T}) = \mathbb{E}_\infty[\hat{T}], \tag{1}$$
$$\mathrm{WADD}(\hat{T}) = \sup_{\tau \geq 1} \operatorname{esssup} \mathbb{E}_\tau[(\hat{T}-\tau)^+|\mathcal{F}_{\tau-1}]. \tag{2}$$

The goal is to minimize the WADD subject to a constraint on the ARL:

$$\min_{\hat{T}:\ \mathrm{ARL}(\hat{T}) \geq \psi} \mathrm{WADD}(\hat{T}), \tag{3}$$

where $\psi > 0$ is some pre-specified constant.

In this paper, we assume that the Markov chains with transition kernels P and Q are uniformly ergodic.

Assumption 1. *The Markov chains with transition kernels P and Q are uniformly ergodic: for any measurable set $A \subseteq \mathcal{X}$*

$$|\mathbb{P}_\infty(X_{t+i} \in A|X_i = x) - \pi_P(A)| \leq R_P \lambda_P^t, \tag{4}$$
$$|\mathbb{P}_1(X_{t+i} \in A|X_i = x) - \pi_Q(A)| \leq R_Q \lambda_Q^t, \tag{5}$$

where π_P and π_Q denote the stationary distributions for P and Q, and $0 < R_P, R_Q < \infty$, $0 < \lambda_P, \lambda_Q < 1$.

It can then be easily shown that under $\mathbb{P}_\infty$, the stationary distribution of $\{X_t, X'_t\}_{t=1}^{\infty}$ is $\pi_P(x)P'(x'|x)$, and under $\mathbb{P}_1$, its stationary distribution is $\pi_Q(x)Q'(x'|x)$. We then denote the marginal distribution of the observation under the stationary distribution by π'_P and π'_Q, respectively:

$$\pi'_P(x') = \int_{\mathcal{X}} \pi_P(dx)P'(x'|x),$$
$$\pi'_Q(x') = \int_{\mathcal{X}} \pi_Q(dx)Q'(x'|x). \tag{6}$$

A. Maximum Mean Discrepancy

In this section, we briefly introduce the kernel mean embedding and the maximum mean discrepancy (MMD) [25], [26]. Let k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the positive definite kernel function of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$. We use $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ to denote the inner product in the RKHS. For any $x, y \in \mathcal{X}$, we have $k(x, \cdot) \in \mathcal{H}_k$ and $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_k}$. According to the reproducing property, given any function $f \in \mathcal{H}_k$, we have $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_k}$. In this paper, we consider a bounded kernel function k, i.e., $0 \leq k(x,y) \leq 1$. For a probability distribution F, denote by $\mu_F = \mathbb{E}_{X\sim F}[k(X,\cdot)]$ its kernel mean embedding. For a characteristic kernel k, we have $\mu_F = \mu_G$ if and only if $F = G$ [26]. The MMD between F and G are defined as:

$$D(F,G) = \sup_{f:\|f\|_{\mathcal{H}_k} \leq 1} \Big| \mathbb{E}_{X\sim F}[f(X)] - \mathbb{E}_{Y\sim G}[f(Y)] \Big|. \tag{7}$$

The squared MMD can be equivalently written as [27]:

$$D^2(F,G) = \mathbb{E}_{X,\bar{X}\sim F}[k(X,\bar{X})] + \mathbb{E}_{Y,\bar{Y}\sim G}[k(Y,\bar{Y})] - 2\mathbb{E}_{X\sim F, Y\sim G}[k(X,Y)]. \tag{8}$$

III. Higher-Order Markov Chain

For Markov models, even if $P \neq Q$, π_P can still be the same as π_Q, making the detection difficult if only using the first-order statistic [28], [29]. The same issue also appears in HMMs. Even if $P \neq Q$ and/or $P' \neq Q'$, the stationary distributions and/or the induced marginal distributions of the observations may still be the same.

For Markov models, a solution is to consider the 2nd-order Markov chain, defined as $\{X_t, X_{t+1}\}_{t=1}^{\infty}$. As long as $P \neq Q$, then the stationary distribution of the 2nd-order Markov chain $\pi_P \cdot P$ is not the same as $\pi_Q \cdot Q$.

We generalize this idea to the HMMs and consider the higher-order HMM. Define the ith-order HMM:

$$\Big\{\big((X_t,X_t'),(X_{t+1},X_{t+1}'),...,(X_{t+i-1},X_{t+i-1}')\big)\Big\}_{t=1}^{\infty},$$

and the corresponding ith-order observation sequence:

$$\{(X_t',X_{t+1}',...,X_{t+i-1}')\}_{t=1}^{\infty}.$$

It can be easily shown that the ith-order HMM is an HMM, and is also uniformly ergodic. With a slight abuse of notation, we denote the marginal distribution of the i-th order observation under the stationary distribution by π_P' and π_Q' (similar to (6) and also see (9) and (10)).

For notational convenience, we focus on the case $i=2$, and make the following assumption.

Assumption 2. *There exist $A,B\subseteq\mathcal{X}$ such that*

$$\pi_P'(X_{t+1}'\in A|X_t'\in B)\neq\pi_Q'(X_{t+1}'\in A|X_t'\in B).$$

This assumption guarantees that $\pi_P'(X_t',X_{t+1}')$ and $\pi_Q'(X_t',X_{t+1}')$ are different, and thus the change is detectable.

We note that using an order higher than 2 will result in a relaxed assumption than Assumption 2, but will introduce a higher computational cost. To keep the presentation clean, we use 2nd-order in this paper. Generalization of our algorithm and analysis to a higher-order is straightforward.

IV. Main Results

Denote by $\widetilde{\mathcal{X}}=\mathcal{X}\times\mathcal{X}$ the state space of the 2nd-order HMM and denote the product σ-algebra on $\widetilde{\mathcal{X}}$ by $\mathcal{F}\otimes\mathcal{F}=\sigma\{A\times B: A\in\mathcal{F},B\in\mathcal{F}\}$. Denote by $\widetilde{X}_t'=\big(X_t',X_{t+1}'\big)$ the second-order observation and $\widetilde{X}_t=\big(X_t,X_{t+1}\big)$ the second-order hidden state. For the Markov chain $\big\{\big((X_t,X_t'),(X_{t+1},X_{t+1}')\big)\big\}_{t=1}^{\infty}$, the observation's marginal probability of the stationary distribution can be written as

$$\pi_P'(A,B)=\int_{\mathcal{X}}\pi_P(dx_t)P'\big(X_t'\in A|X_t=x_t\big)$$
$$\cdot\int_{\mathcal{X}}P(dx_{t+1}|X_t=x_t)P'\big(X_{t+1}'\in B|X_{t+1}=x_{t+1}\big)\quad(9)$$

under $\mathbb{P}_\infty$, and

$$\pi_Q'(A,B)=\int_{\mathcal{X}}\pi_Q(dx_t)Q'\big(X_t'\in A|X_t=x_t\big)$$
$$\cdot\int_{\mathcal{X}}Q(dx_{t+1}|X_t=x_t)Q'\big(X_{t+1}'\in B|X_{t+1}=x_{t+1}\big)\quad(10)$$

under $\mathbb{P}_1$, for any $A,B\subseteq\mathcal{X}$. From Assumption 2, we have that $\pi_P'(\widetilde{X}_t')\neq\pi_Q'(\widetilde{X}_t')$.

Denote by $\widetilde{Y}_t'=\big(Y_t',Y_{t+1}'\big)$ the observation at time t of the second-order reference sequence. We divide the samples into non-overlapping blocks and the size of each block is m. For the t-th observation block, $t=0,1,2,...$, we have $m-1$ second-order samples and the empirical distribution for this block can be written as $F_{\widetilde{X}'}^t=\frac{1}{m-1}\sum_{i=1}^{m-1}\delta_{\widetilde{X}'_{mt+i}}$, where $\delta_{\widetilde{X}_i'}$ is the Dirac measure. Similarly, for the t-th reference block, we have $F_{\widetilde{Y}'}^t=\frac{1}{m-1}\sum_{i=1}^{m-1}\delta_{\widetilde{Y}'_{mt+i}}$. We then can calculate the MMD between these two empirical distributions as follows

$$D\Big(F_{\widetilde{X}'}^t,F_{\widetilde{Y}'}^t\Big)=\frac{1}{(m-1)}\Big(\sum_{mt<i,j<m(t+1)}k\big(\widetilde{X}_i',\widetilde{X}_j'\big)$$
$$+\sum_{mt<i,j<m(t+1)}k\big(\widetilde{Y}_i',\widetilde{Y}_j'\big)-2\sum_{mt<i,j<m(t+1)}k\big(\widetilde{X}_i',\widetilde{Y}_j'\big)\Big)^{\frac{1}{2}}.$$

Here the kernel function is defined to be $\mathcal{X}^2\times\mathcal{X}^2\to\mathbb{R}$. Since the MMD is always non-negative, we define our test statistic as

$$S_t'=D\Big(F_{\widetilde{X}'}^t,F_{\widetilde{Y}'}^t\Big)-\sigma',$$

where the offset $\sigma'>0$ is some positive constant to be specified later. We introduce the offset σ' to guarantee that the expectation of the test statistic S_t' is negative before the change and is positive after the change. Then, our stopping time is defined as

$$T(c)=\inf\Bigg\{mt+t:\max_{0\le i\le t}\sum_{j=i}^{t}S_j'>c\Bigg\},\qquad(11)$$

where $c>0$ is a the threshold. This algorithm can be updated recursively:

$$\max_{0\le i\le t}\sum_{j=i}^{t}S_j'=\max\Big\{0,\max_{0\le i\le t-1}\sum_{j=i}^{t-1}S_j'+S_t'\Big\}.$$

In addition, the computational complexity for MMD is $\mathcal{O}(m^2)$ for every m samples. For n samples, there are $\lfloor\frac{n}{m}\rfloor$ non-overlapping blocks in total. Hence the overall computational complexity at time n is $\mathcal{O}(mn)$.

As will be shown later, for large m,

$$D\Big(F_{\widetilde{X}'}^t,F_{\widetilde{Y}'}^t\Big)\approx D\left(\pi_P'(\widetilde{X}'),\pi_P'(\widetilde{Y}')\right)=0$$

before the change and

$$D\Big(F_{\widetilde{X}'}^t,F_{\widetilde{Y}'}^t\Big)\approx D\left(\pi_Q'(\widetilde{X}'),\pi_P'(\widetilde{Y}')\right)>0$$

after the change. Therefore, we choose

$$0<\sigma'<D\left(\pi_P'(\widetilde{X}'),\pi_Q'(\widetilde{Y}')\right).$$

In this way, before the change the statistic has a negative drift and after the change it has a positive drift.

Define

$$a_P=\sqrt{\frac{2-2\lambda_P+4R_P}{(m-1)(1-\lambda_P)}},\quad a_Q=\sqrt{\frac{2-2\lambda_Q+4R_Q}{(m-1)(1-\lambda_Q)}},$$
$$a=a_P+a_Q,\qquad\qquad d=D(\pi_P',\pi_Q')-\sigma'.\quad(12)$$

We then derive the upper bound on the WADD.

Theorem 1. *The WADD for the stopping time in* (11) *can be bounded as follows:*

$$\mathrm{WADD}(T(c))$$
$$\le\frac{2\sqrt{ad}mc}{(a-d)^2}+\frac{(a+d)mc}{(a-d)^2}+2m+\frac{a+\sqrt{ad}}{d-a}m$$

$$=\mathcal{O}(mc). \tag{13}$$

Note that $a=\mathcal{O}\Big(\frac{1}{\sqrt{m}}\Big)$. Therefore, $a\approx 0$ when m is large. Then the WADD grows linearly with m.

Proof sketch of Theorem 1. Let τ' be the index of the last sample within the block where the change happens. Then, samples after τ' follow the transition kernel Q. Let

$$\xi'=\frac{a_P+a_Q+\sqrt{(D(\pi'_P,\pi'_Q)-\sigma')(a_P+a_Q)}}{D(\pi'_P,\pi'_Q)-\sigma'-a_P-a_Q}$$

such that

$$D(\pi'_P,\pi'_Q)-\sigma'-\frac{\xi'+1}{\xi'}(a_P+a_Q)>0.$$

Denote by $n'_c=\Big\lceil\frac{(\xi'+1)c}{D(\pi'_P,\pi'_Q)-\sigma'}\Big\rceil$. Firstly, we have that

$$\mathbb{E}_\tau\Big[\frac{(T(c)-\tau')^+}{mn'_c}\Big|\mathcal{F}_{\tau-1}\Big]\leq\sum_{j=0}^{\infty}\mathbb{P}_\tau\Big(\frac{T(c)-\tau'}{mn'_c}\geq j\Big|\mathcal{F}_{\tau-1}\Big).$$

By the definition of $T(c)$, we have that

$$\begin{aligned}&\mathbb{P}_\tau\Big(\frac{T(c)-\tau'}{mn'_c}\geq j\Big|\mathcal{F}_{\tau-1}\Big)\\ &\leq\mathbb{P}_\tau\Big(\sum_{i=(j-1)n'_c+1}^{jn'_c}S'_i<c\Big|\mathcal{F}_{\tau-1},\frac{T(c)-\tau'}{mn'_c}\geq j-1\Big)\\ &\times\mathbb{P}_\tau\Big(\frac{T(c)-\tau'}{mn'_c}\geq j-1\Big|\mathcal{F}_{\tau-1}\Big).\end{aligned} \tag{14}$$

It then suffices to bound $\mathbb{P}_\tau\Big(\sum_{i=(j-1)n'_c+1}^{jn'_c}S'_i<c\Big|\mathcal{F}_{\tau-1},\frac{T(c)-\tau'}{mn'_c}\geq j-1\Big)$ and to apply (14) recursively. By the triangle inequality and Markov inequality, we have that

$$\begin{aligned}&\mathbb{P}_\tau\left(\sum_{i=(j-1)n'_c+1}^{jn'_c}S'_i<c\middle|\mathcal{F}_{\tau-1},\frac{T(c)-\tau'}{mn'_c}\geq j-1\right)\\ \leq&\mathbb{P}_\tau\left(\sum_{i=(j-1)n'_c+1}^{jn'_c}\Big(D(\pi'_P,\pi'_Q)-\sigma'-D\Big(F^i_{\widetilde{X}'},\pi'_Q\Big)\right.\\ &\left.-D\Big(F^i_{\widetilde{Y}'},\pi'_P\Big)\Big)<c\middle|\mathcal{F}_{\tau-1},\frac{T(c)-\tau'}{mn'_c}\geq j-1\right)\\ \leq&\sum_{i=(j-1)n'_c+1}^{jn'_c}\\ &\frac{\mathbb{E}_\tau\Big[D\Big(F^i_{\widetilde{X}'},\pi'_Q\Big)+D\Big(F^i_{\widetilde{Y}'},\pi'_P\Big)\Big|\mathcal{F}_{\tau-1},\frac{T(c)-\tau'}{mn'_c}\geq j-1\Big]}{n'_c(D(\pi'_P,\pi'_Q))-\sigma')-c}\\ \leq&\frac{a_P+a_Q}{D(\pi'_P,\pi'_Q)-\sigma'}\left(1+\frac{1}{\xi'}\right).\end{aligned} \tag{15}$$

The last inequality in (15) follows from Proposition 1, which will be given later. Let $\delta'=\frac{a_P+a_Q}{D(\pi'_P,\pi'_Q)-\sigma'}\Big(1+\frac{1}{\xi'}\Big)$, we can show that

$$\mathbb{P}_\tau\Big(\frac{T(c)-\tau'}{mn'_c}\geq j|\mathcal{F}_{\tau-1}\Big)\leq\delta'^j.$$

Then it follows that

$$\mathbb{E}_\tau\Big[\frac{(T(c)-\tau')^+}{mn'_c}\Big|\mathcal{F}_{\tau-1}\Big]<(1-\delta')^{-1}.$$

Then we plug in the values of δ' and ξ' and this concludes the proof. ☐

One key step in the proof of Theorem 1 is to prove the upper bound on the expectation of the MMD between the marginal distribution of the stationary distribution π'_P and the empirical probability $F^t_{\widetilde{X}'}$ under $\mathbb{P}_\infty$ and the MMD between π'_Q and $F^t_{\widetilde{X}'}$ under $\mathbb{P}_1$ (Proposition 1). Before that, we first provide a useful lemma to measure the correlation between samples. For HMMs, we define the coefficient

$$\begin{aligned}&\rho'_\infty(x_0,i,j)\\ &=\Big|\mathbb{E}_\infty\Big[\Big\langle k\Big(\widetilde{X}'_i,\cdot\Big)-\mu_{\pi'_P},k\Big(\widetilde{X}'_j,\cdot\Big)-\mu_{\pi'_P}\Big\rangle_{\mathcal{H}_k}|X_0=x_0\Big]\Big|\end{aligned}$$

under $\mathbb{P}_\infty$ and

$$\begin{aligned}&\rho'_1(x_0,i,j)\\ &=\Big|\mathbb{E}_1\Big[\Big\langle k\Big(\widetilde{X}'_i,\cdot\Big)-\mu_{\pi'_Q},k\Big(\widetilde{X}'_j,\cdot\Big)-\mu_{\pi'_Q}\Big\rangle_{\mathcal{H}_k}|X_0=x_0\Big]\Big|\end{aligned}$$

under $\mathbb{P}_1$ to measure the correlation between the i-th and j-th samples conditioned on $X_0=x_0$. We then have the following lemma.

Lemma 1. *Under $\mathbb{P}_\infty$ and $\mathbb{P}_1$, $\forall x_0\in\mathcal{X},j>i>0$, we have*

$$\begin{aligned}\rho'_\infty(x_0,i,j)&\leq 2R_P\lambda_P^{j-i-1},\\ \rho'_1(x_0,i,j)&\leq 2R_Q\lambda_Q^{j-i-1}.\end{aligned} \tag{16}$$

Based on Lemma 1, the expectation of the MMD between $\pi'_P(\pi'_Q)$ and $F^t_{\widetilde{X}'}$ under $\mathbb{P}_\infty(\mathbb{P}_1)$ can be bounded in the following proposition.

Proposition 1. $\forall x_{mt}\in\mathcal{X},\mathbb{E}_\infty\Big[D\Big(F^t_{\widetilde{X}'},\pi'_P\Big)|X_{mt}=x_{mt}\Big]\leq a_P$ *and* $\mathbb{E}_1\Big[D\Big(F^t_{\widetilde{X}'},\pi'_Q\Big)|X_{mt}=x_{mt}\Big]\leq a_Q$.

We then will derive a lower bound on the ARL, which is exponential in the threshold. Before that we provide a high-probability bound on the sum of two MMDs between $\pi'_P(\pi'_Q)$ and $F^t_{\widetilde{X}'}$. The main challenge here lies in that the generalization of the McDiarmid's inequality [30] cannot be applied to HMMs directly.

Proposition 2. *For any x_{mt}, $y_{mt}\in\mathcal{X}$ and $\delta>0$,*

$$\begin{aligned}\mathbb{P}_\infty\Big(D(F^t_{\widetilde{X}'},\pi'_P)+D(F^t_{\widetilde{Y}'},\pi'_P)\leq\sqrt{\frac{\log(\frac{1}{\delta})}{\Gamma'_P}}\\ +2a_P\Big|X_{mt}=x_{mt},Y_{mt}=y_{mt}\Big)\geq 1-2\delta,\end{aligned} \tag{17}$$

where $\Gamma'_P=\frac{\Gamma_P}{(1+\sqrt{2m\Gamma_P})^2}>0$ and $\Gamma_P>0$ is a constant.

We then theoretically develop the lower bound on the ARL for our test in (11). Denote by $h=\sigma-2a_P$. Since the value of a_P decreases with m, we can always find a large m so that $h > 0$. Define a function

$$\phi(q) = \frac{\sqrt{\pi}}{\sqrt{\Gamma'_P}} q \exp\Big(-qh + \frac{q^2}{4\Gamma'_P}\Big).$$

For this continuous function, let $q > 0$ be a constant such that $\phi(q) \leq \frac{1}{2}$. Such a q always exists since $\phi(0) = 0$ and $\phi(q) \to \infty$ as $q \to \infty$.

Theorem 2. *The lower bound on the ARL in* (11) *is exponential in the threshold c:*

$$\text{ARL}(T(c)) \geq m \exp(qc). \tag{18}$$

Proof sketch. Define the following stopping times

$$\Delta'_1 = \inf\Big\{t : \sum_{i=0}^{t} S'_i < 0\Big\}$$

and

$$\Delta'_{r+1} = \inf\Big\{t > \Delta'_r : \sum_{i=\Delta'_r+1}^{t} S'_i < 0\Big\}.$$

Let $R' = \inf\{r \geq 0 : \Delta'_r < \infty$ and $\sum_{i=\Delta'_r+1}^{t} S'_i \geq c$ for some $t > \Delta'_r\}$. By the definition of $T(c)$, we then have that

$$\mathbb{E}_\infty[T(c)] \geq m\mathbb{E}_\infty[R'] \geq m \sum_{r=0}^{\infty} \mathbb{P}_\infty(R' > r). \tag{19}$$

Further, we can prove

$$\mathbb{P}_\infty(R' > r) \geq \Big(1 - \mathbb{P}_\infty\big(\max_{n \geq \Delta'_r+1} \sum_{i=\Delta'_r+1}^{n} S'_i < c \big| \mathcal{F}_{\Delta'_r m}\big)\Big) \times \mathbb{P}_\infty(R' > r-1).$$

We then show that

$$\mathbb{P}_\infty\Bigg(\max_{n \geq \Delta'_r+1} \sum_{i=\Delta'_r+1}^{n} S'_i \geq c \Bigg| \mathcal{F}_{\Delta'_r m}\Bigg) \leq \mathbb{E}_\infty\big[\exp\big(qS'_{\Delta'_r+1}\big)\big|\mathcal{F}_{\Delta'_r m}\big]/\exp(qc).$$

It then suffices to bound

$$\mathbb{E}_\infty\big[\exp\big(qS'_{\Delta'_r+1}\big)\big|\mathcal{F}_{\Delta'_r m}\big]/\exp(qc).$$

By Proposition 2, we show that

$$\mathbb{E}_\infty\big[\exp\big(qS'_{\Delta'_r+1}\big)\big|\mathcal{F}_{\Delta'_r m}\big] \leq 1.$$

It then follows that

$$1 - \mathbb{P}_\infty\Bigg(\max_{n \geq \Delta'_r+1} \sum_{i=\Delta'_r+1}^{n} S'_i < c \Bigg| \mathcal{F}_{\Delta'_r m}\Bigg) \geq 1 - \exp(-qc).$$

This further suggests that

$$\mathbb{E}_\infty[R'] \geq \sum_{r=0}^{\infty} \mathbb{P}_\infty(R' > r) \geq \sum_{r=0}^{\infty} \big(1 - \exp(-qc)\big)^r = \exp(qc).$$

This concludes the proof. □

Note that q and m are independent of c. Therefore, the lower bound on ARL is exponentially in the threshold c.

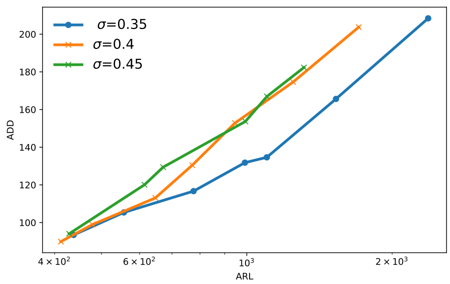

Fig. 1. ADD v.s. ARL: $m = 0.15$.

Recall that in [12] a universal lower bound on WADD was developed, which shows that for any stopping time with ARL $\geq \psi$, the detection delay is at least $\mathcal{O}(\log(\psi))$. For our test to satisfy the ARL constraint in (3), the threshold c should be set to $\frac{\log(\psi) - \log(m)}{q}$. With this threshold, by Theorem 1, our algorithm achieves a detection delay of $\mathcal{O}(\log(\psi))$ while satisfying the false alarm constraint. This matches with (order-level) the universal lower bound in [12] for the general non-i.i.d. setting.

V. Numerical Results

In this section, we provide some simulation results. For the Markov chain, let the transition kernel be $P = [0.2, 0.7, 0.1; 0.9, 0.0, 0.1; 0.2, 0.8, 0.0]^\top$ before the change and $Q = [0.5, 0.5, 0.0; 0.0, 0.5, 0.5; 0.2, 0.3, 0.5]^\top$ after the change. The emission probability density matrix is $[0.8, 0.1, 0.1; 0.2, 0.6, 0.2; 0.3, 0.3, 0.4]^\top$ and doesn't change. We use the Gaussian kernel

$$k(x, y) = \exp\big(-\beta(x - y)^2\big),$$

where β is the bandwidth parameter. We pick $\beta = \frac{1}{14}$ that achieves the best ADD and ARL tradeoff. We set $m = 15$, and use the offset $\sigma = 0.35, 0.4$ and 0.45 respectively. To compare the ADD and the ARL, in Fig 1, we plot the ADD as a function of the log of ARL by varying the threshold. From Fig 1, we can see that for all σ, the ADD grows with the log of ARL linearly, which matches with our theoretical results.

VI. Conclusion

In this paper, we developed a data-driven approach to detect a change in the transition kernel and/or emission probability in HMMs. We theoretically characterize its ARL and WADD, and show that the ADD is at most in the logarithm of the ARL, which matches with (order-level) the universal lower bound for the general non-i.i.d. problem in [12]. We also provide simulation results to validate the performance of our algorithm.

VII. Acknowledgment

The work of Q. Zhang, Z. Sun and S. Zou is partially supported by National Science Foundation (NSF) under grants CCF-1948165, CCF-2106560, and ECCS-2112693. The work of L.C.Herrera is supported by NSF under grant ECCS-1855888.

References

[1] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2009.

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

[3] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 2, pp. 494–514, 2021.

[4] V. V. Veeravalli and T. Banerjee, "Quickest change detection," *Academic press library in signal processing: Array and statistical signal processing*, vol. 3, pp. 209–256, 2013.

[5] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: theory and application*, vol. 104. Prentice Hall Englewood Cliffs, 1993.

[6] K. Gajula, V. Le, X. Yao, S. Zou, and L. Herrera, "Quickest Detection of Series Arc Faults on DC Microgrids," in *Proc. IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 796–801, IEEE, 2021.

[7] M. Baron, V. Antonov, C. Huber, M. Nikulin, and V. Polischook, "Early detection of epidemics as a sequential change-point problem," *Longevity, Aging and Degradation Models in Reliability, Public Health, Medicine and Biology, LAD*, pp. 7–9, 2004.

[8] J. J. Shen and N. R. Zhang, "Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing," *The Annals of Applied Statistics*, vol. 6, no. 2, pp. 476–496, 2012.

[9] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[10] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *Journal of Molecular Biology*, vol. 235, no. 5, pp. 1501–1531, 1994.

[11] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 613–658, 1995.

[12] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[13] 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.

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

[15] 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.

[16] B. Yakir, "Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space," *Lecture Notes-Monograph Series*, pp. 346–358, 1994.

[17] A. M. Polansky, "Detecting change-points in Markov chains," *Computational statistics & data analysis*, vol. 51, no. 12, pp. 6013–6026, 2007.

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

[19] 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.

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

[21] C.-D. Fuh, "Asymptotic operating characteristics of an optimal change point detection in hidden Markov models," *The Annals of Statistics*, vol. 32, no. 5, pp. 2305–2339, 2004.

[22] 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.

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

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

[25] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

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

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

[28] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou, "Data-Driven Quickest Change Detection in Markov Models," *arXiv preprint arXiv:2210.11988*, 2022.

[29] H. Chen, J. Tang, and A. Gupta, "Change Detection of Markov Kernels with Unknown Post Change Kernel using Maximum Mean Discrepancy," *arXiv preprint arXiv:2201.11722*, 2022.

[30] A. Havet, M. Lerasle, E. Moulines, and E. Vernet, "A quantitative McDiarmid's inequality for geometrically ergodic Markov chains," *Electronic Communications in Probability*, vol. 25, pp. 1–11, 2020.