A Change-Detection based Framework for Piecewise-stationary Multi-Armed Bandit Problem

Fang Liu and JooHyun Lee and Ness Shroff
The Ohio State University
Columbus, Ohio 43210
{liu.3977, lee.7119, shroff.11} @osu.edu

Abstract
The multi-armed bandit problem has been extensively studied under the stationary assumption. However in reality, this assumption often does not hold because the distributions of rewards themselves may change over time. In this paper, we propose a change-detection (CD) based framework for multi-armed bandit problems under the piecewise-stationary setting, and study a class of change-detection based UCB (Upper Confidence Bound) policies, CD-UCB, that actively detects change points and restarts the UCB indices. We then develop CUSUM-UCB and PHT-UCB, that belong to the CD-UCB class and use cumulative sum (CUSUM) and Page-Hinkley Test (PHT) to detect changes. We show that CUSUM-UCB obtains the best known regret upper bound under mild assumptions. We also demonstrate the regret reduction of the CD-UCB policies over arbitrary Bernoulli rewards and Yahoo! datasets of webpage click-through rates.

1 Introduction
The multi-armed bandit problem, introduced by Thompson (1933), models sequential allocation in the presence of uncertainty and partial feedback on rewards. It has been extensively studied and has turned out to be fundamental to many problems in artificial intelligence, such as reinforcement learning (Sutton and Barto 1998), online recommendation systems (Li, Karatzoglou, and Gentile 2016) and computational advertisement (Buccapatnam et al. 2017). In reinforcement learning (Sutton and Barto 1998), online recommendation systems (Li, Karatzoglou, and Gentile 2016) and computational advertisement (Buccapatnam et al. 2017). In the classical multi-armed bandit problem (Lai and Robbins 1985), a decision maker needs to choose one of \( K \) independent arms and obtains the associated reward in a sequence of time slots (rounds). Each arm is characterized by an unknown reward distribution and the rewards are independent and identically distributed (i.i.d.).

The goal of a bandit algorithm, implemented by the decision maker, is to minimize the regret over \( T \) time slots, which is defined as the expectation of the difference between the total rewards collected by playing the arm with the highest expected reward and the total rewards obtained by the algorithm. To achieve this goal, the decision maker is faced with an exploration versus exploitation dilemma, which is the trade-off between exploring the environment to find the most profitable arms and exploiting the current empirically best arm as often as possible. A problem-dependent regret lower bound, \( \Omega(\log T) \), of any algorithm for the classical bandit problem has been shown in Lai and Robbins (1985). Several algorithms have been proposed and proven to achieve \( O(\log T) \) regret, such as Thompson Sampling (Agrawal and Goyal 2012), \( \epsilon \)-greedy and Upper Confidence Bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002). Variants of these bandit policies can be found in Bubeck and Cesa-Bianchi (2012).

Although the stationary (classical) multi-armed bandit problem is well-studied, it is unclear whether it can achieve \( O(\log T) \) regret in a non-stationary environment, where the distributions of rewards change over time. This setting often occurs in practical problems. For example, consider the dynamic spectrum access problem (Alaya-Feki, Moulines, and LeCornec 2008) in communication systems. Here, the decision maker wants to exploit the empty channel, thus improving the spectrum usage. The availability of a channel is dependent on the number of users in the coverage area. The number of users, however, can change dramatically with time of day and, therefore, is itself a non-stationary stochastic process. Hence, the availability of the channel also follows a distribution that is not only unknown, but varies over time. To address the changing environment challenge, a non-stationary multi-armed bandit problem has been proposed in the literature. There are two main approaches to deal with the non-stationary environment: passively adaptive policies (Garivier and Moulines 2008; Besbes, Gur, and Zeevi 2014; Wei, Hong, and Lu 2016) and actively adaptive policies (Hartland et al. 2007; Mellor and Shapiro 2013; Allesiardo and Féraud 2015).

First, passively adaptive policies are unaware of when changes happen but update their decisions based on the most recent observations in order to keep track of the current best arm. Discounted UCB (D-UCB), introduced by Kocsis and Szepesvári (2006), where geometric moving average over the samples is applied to the UCB index of each arm, has been shown to achieve the regret upper-bounded by \( O(\sqrt{\gamma T\log T}) \), where \( \gamma T \) is the number of change points up to time \( T \) (Garivier and Moulines 2008). Based on the analysis of D-UCB, they also proposed and analyzed Sliding-Window UCB (SW-UCB), where the algorithm updates the UCB index based on the observations within a
moving window of a fixed length. The regret of SW-UCB is at most \(O(\sqrt{\gamma T \log T})\). Exp3.S (Auer et al. 2002) also achieves the same regret bound, where a uniform exploration is mixed with the standard Exp3 (Cesa-Bianchi and Lugosi 2006) algorithm. Similarly, Besbes, Gur, and Zeevi (2014) proposed a Rexp3 algorithm, which restarts the Exp3 algorithm periodically. It is shown that the regret is upper-bounded by \(O(V_T^{1/3} T^{2/3})\), where \(V_T\) denotes the total reward variation budget up to time \(T\). The increased regret of Rexp3 comes from the adversarial nature of the algorithm, which assumes that the environment changes every time slot. It is shown that the regret is upper-bounded by \(O(\sqrt{\gamma T \log T})\). To the best of our knowledge, this is the first regret bound for actively adaptive UCB policies in the bandit feedback setting.

3. The performance of the proposed and existing policies is validated by both synthetic and real world datasets, and we show that our proposed algorithms are superior to other existing policies in terms of regret.

We present the problem setting in Section 2 and introduce our framework in Section 3. We propose our algorithms in Section 4. We then present performance guarantees in Section 5. In Section 6, we compare our algorithms with other existing algorithms via simulation. Finally, we conclude the paper.

2 Problem Formulation

2.1 Basic Setting

Let \(K = \{1, \ldots, K\}\) be a set of arms. Let \(\{1, 2, \ldots, T\}\) denote the decision slots faced by a decision maker and \(T\) is the time horizon. At each time slot \(t\), the decision maker chooses an arm \(I_t \in K\) and obtains a reward \(X_t(I_t) \in [0, 1]\). Note that the results can be generalized to any bounded interval. The rewards \(\{X_t(i)\}_{t \geq 1}\) for arm \(i\) are modeled by a sequence of independent random variables from potentially different distributions, which are unknown to the decision maker. Let \(\mu_i(t)\) denote the expectation of reward \(X_t(i)\) at time slot \(t\), i.e., \(\mu_i(t) = \mathbb{E}[X_t(i)]\). Let \(i^*_t\) be the arm with highest expected reward at time slot \(t\), denoted by \(\mu_i(t) \triangleq \mu_i(i^*_t) = \max_{i \in K} \mu_i(i)\). Let \(\Delta_{i^*_t}(t) \triangleq \min \{\mu_i(t) - \mu_i(t) : t \leq T, i \neq i^*_t\}\), be the minimum difference over all time slots between the expected rewards of the best arm \(i^*_t\) and the arm \(i\) while the arm \(i\) is not the best arm.

A policy \(\pi\) is an algorithm that chooses the next arm to play based on the sequence of past plays and obtained rewards. The performance of a policy \(\pi\) is measured in terms of the regret. The regret of \(\pi\) after \(T\) plays is defined as the expected total loss of playing suboptimal arms. Let \(R_\pi(T)\) denote the regret of policy \(\pi\) after \(T\) plays and let \(\tilde{N}_T(i)\) be the number of times arm \(i\) has been played when it is not the best arm by \(\pi\) during the first \(T\) plays.

\[
R_\pi(T) = \mathbb{E} \left[ \sum_{t=1}^{T} (X_t(i^*_t) - X_t(I_t)) \right],
\]

\[
\tilde{N}_T(i) = \sum_{t=1}^{T} \mathbb{I}_{\{I_t = i, \mu_i(t) \neq \mu_{i^*_t}(t)\}}.
\]

Note that the regret \(R_\pi(T)\) of policy \(\pi\) is upper-bounded by \(\sum_{i=1}^{K} \mathbb{E}[\tilde{N}_T(i)]\) since the rewards are bounded in (1). In
Section 5, we provide an upper bound on $E[\tilde{N}_T(i)]$ to obtain a regret upper bound.

2.2 Piecewise-stationary Environment

We consider the notion of a piecewise-stationary environment in Yu and Mannor (2009), where the distributions of rewards remain constant for a certain period and abruptly change at some unknown time slots, called breakpoints.

Let $\gamma_T$ be the number of breakpoints up to time $T$, $\gamma_T = \sum_{t=1}^{T-1} 1\{\exists i \in K : \mu_t(i) \neq \mu_{t+1}(i)\}$. In addition, we make three mild assumptions for tractability.

**Assumption 1.** (Piecewise Stationarity) The shortest interval between two successive breakpoints is greater than $KM$, for some integer $M$.

Assumption 1 ensures that the shortest interval between two successive breakpoints is greater than $KM$, so that we have enough samples to estimate the mean of each arm before the change happens. Note that this assumption is equivalent to the notions of an abruptly changing environment used in Garivier and Moulines (2008) and a switching environment in Mellor and Shapiro (2013). However, it is different from the adversarial environment assumption, where the environment changes all the time. We make a similar assumption as Assumption 4.2 in Yu and Mannor (2009) about the detectability in this paper.

**Assumption 2.** (Detectability) There exists a known parameter $\epsilon > 0$, such that for all $i \in K$ and for all $t \leq T - 1$, if $\mu_t(i) \neq \mu_{t+1}(i)$, then $|\mu_t(i) - \mu_{t+1}(i)| \geq 3\epsilon$.

Assumption 2 excludes infinitesimal mean shift, which is reasonable in practice when detecting abrupt changes bounded from below by a certain threshold.

**Assumption 3.** (Bernoulli Reward) The distributions of all the arms are Bernoulli distributions.

Assumption 3 has also been used in the literature (Besbes, Gur, and Zeevi 2014; Mellor and Shapiro 2013; Kaufmann, Korda, and Munos 2012; Agrawal and Goyal 2012). By Assumption 3, the empirical average of $M$ Bernoulli random variables must be one of the grid points $\{0, 1/M, \ldots, 1\}$. Let $\lambda_T(i) = \min\{(\mu_t(i) - \epsilon) - (\mu_{t+1}(i) + \epsilon) : i \in K, t \leq T\}$ be the minimal non-trivial gap between expectation and closest grid point of arm $i$. We define the minimal gap of all arms as $\lambda = \min_{i \in K} \lambda_T(i)$.

3 Change-Detection based Framework

Our change-detection based framework consists of two components: a change detection algorithm and a bandit algorithm, as shown in Figure 1. At each time $t$, the bandit algorithm outputs a decision $I_t \in K$ based on its past observations of the bandit environment. The environment generates the corresponding reward of arm $I_t$, which is observed by both the bandit algorithm and the change detection algorithm. The change detection algorithm monitors the distribution of each arm, and sends out a positive signal to restart the bandit algorithm once a breakpoint is detected. One can find that our framework is a generalization of the existing actively adaptive policies.

Since the bandit algorithms are well-studied in the bandit setting, what remains is to find a change detection algorithm, which works in the bandit environment. Change point detection problems have been well studied, see, e.g., the book (Basseville and Nikiforov 1993). However, the change detection algorithms are applied in a context that is quite different from the bandit setting. There are two key challenges in adapting the existing change detection algorithms in the bandit setting.

1. **Unknown priors:** In the context of the change detection problem, one usually assumes that the prior distributions before and after a change point are known. However, such information is unknown to the decision maker in the bandit setting. Even though there are some simple methods, such as estimating the priors and then applying the change detection algorithm like PHT, there are no analytical results in the literature.

2. **Insufficient samples:** Due to the bandit feedback setting, the decision maker can only observe one arm at each time. However, there are $K$ change detection algorithms running in parallel since each arm is associated with a change detection procedure to monitor the possible mean shift. So the change detection algorithms in most arms are hungry for samples at each time. If the decision maker does not feed these change detection algorithms intentionally, the change detection algorithm may miss detection opportunities because they do not have enough recent samples.

4 Application of the Framework

In this section, we introduce our Change-Detection based UCB (CD-UCB) policy, which addresses the issue of insufficient samples. Then we develop a tailoredCUSUM algorithm for the bandit setting to overcome the issue of unknown priors. Finally, we combine our CUSUM algorithm with the UCB algorithm as CUSUM-UCB policy, which is a specific instance of our change-detection based framework. Performance analysis is provided in Section 5.

4.1 CD-UCB policy

Suppose we have a change detection algorithm, CD$(\cdot, \cdot)$, which takes arm index $i$ and observation $X_t(i)$ as input at
time $t$, and it returns 1 if there is an alarm for a breakpoint. Given such a change detection algorithm, we can employ it to control the UCB algorithm, which is our CD-UCB policy as shown in Algorithm 1. We clarify some useful notations as follows. Let $\tau_i = \tau_i(t)$ be the last time that the CD($i$, ·) alarms and restarts for arm $i$ before time $t$. Then the number of valid observations (after the latest detection alarm) for arm $i$ up to time $t$ is denoted as $N_i(t)$. Let $n_i$ be the total number of valid observations for the decision maker. For each arm $i$, let $X_i(t)$ be the sample average and $C_i(t)$ be the confidence padding term. In particular,

$$N_i(t) = \sum_{s = \tau_i}^{t} 1_{(s = i)}, \quad n_t = \sum_{i = 1}^{K} N_i(t),$$

$$\bar{X}_i(t) = \frac{1}{N_i(t)} \sum_{s = \tau_i}^{t} X_i(s), \quad C_i(t) = \frac{\xi \log n_t}{N_i(t)},$$

where $\xi$ is some positive real number. Thus, the UCB index for each arm $i$ is $\bar{X}_i(t) + C_i(t)$. Parameter $\alpha$ is a tuning parameter we introduce in the CD-UCB policy. At each time $t$, the policy plays the arm

$$I_t = \arg \max_{i \in K} \left( \bar{X}_i(t) + C_i(t) \right), \quad \text{w.p. } 1 - \alpha \quad \forall i \in K, \text{ w.p. } \frac{\alpha}{K}.$$  \hfill (5)

Parameter $\alpha$ controls the fraction of plays we exploit to feed the change detection algorithm. A large $\alpha$ may drive the algorithm to a linear regret performance while a small $\alpha$ can limit the detectability of change detection algorithm. We will discuss the choice of $\alpha$ in Sections 5 and 6.

### 4.2 Tailored CUSUM algorithm

A change detection algorithm observes a sequence of independent random variables, $y_1, y_2, \ldots$, in an online manner, and outputs an alarm once a change point is detected. In the context of the traditional change detection problem, one assumes that the parameters $\theta_0$ and $\theta_1$ are known for the density function $p(y_0 | \theta_0)$ and $p(y_1 | \theta_1)$. In addition, $y_k$ is sampled from distribution $p(y_k | \theta_k)$ before (after) the breakpoint. Let $u_0$ ($u_1$) be the mean of $y_k$ before (after) the change point. The CUSUM algorithm, originally proposed by (Page 1954), has been proven to be optimal in detecting abrupt changes in the sense of worst mean detection delay (Lorden 1971). The basic idea of the CUSUM algorithm is to take a function of the observed sample (e.g., the logarithm of likelihood ratio $\log \frac{p(y_k | \theta_1)}{p(y_k | \theta_0)}$) as the step of a random walk. This random walk is designed to have a positive mean drift after a change point and have a negative mean drift without a change. Hence, CUSUM signals a change if this random walk crosses some positive threshold $h$.

We propose a tailored CUSUM algorithm that works in the bandit setting. To be specific, we use the first $M$ samples to calculate the average, $\bar{u}_0 \triangleq (\sum_{k=1}^{M} y_k) / M$. Then we construct two random walks, which have negative mean drifts before the change point and have positive mean drifts after the change. In particular, we design a two-sided CUSUM algorithm, described in Algorithm 2, with an upper (lower) random walk monitoring the possible positive (negative) mean shift. Let $s_{k}^+$ ($s_{k}^-$) be the step of the upper (lower) random walk. Then $s_{k}^+$ and $s_{k}^-$ are defined as

$$(s_{k}^+, s_{k}^-) = (y_k - \bar{u}_0 - \epsilon, \bar{u}_0 - y_k - \epsilon) 1_{(k > M)}.$$  \hfill (6)

Let $g_{k}^+$ ($g_{k}^-$) track the positive drift of upper (lower) random walk. In particular,

$$g_{k}^+ = \max(0, g_{k-1}^++s_{k}^+), \quad g_{k}^- = \max(0, g_{k-1}^-+s_{k}^-).$$  \hfill (7)

The change point is detected when either of them crosses the threshold $h$. The parameter $h$ is important in the detection delay and false alarm trade-off. We discuss the choice of $h$ in Section 5.

### 4.3 CUSUM-UCB policy

Now we are ready to introduce our CUSUM-UCB policy, which is a CD-UCB policy with CUSUM as a change detection algorithm. In particular, it takes $K$ parallel CUSUM algorithms as CD($i$, ·) in CD-UCB. Formal description of CUSUM-UCB can be found in Algorithm 3, provided in Section F in the supplementary material.

### 4.4 PHT-UCB policy

We introduce another instance of our CD-UCB with the PHT algorithm (Hinkley 1971) running as the change detection algorithm, named PHT-UCB. PHT can be viewed as a variant of Algorithm 2 by replacing (6) with $(s_{k}^+, s_{k}^-) = (y_k - \bar{y}_k - \epsilon, \bar{y}_k - y_k - \epsilon)$, where $\bar{y}_k = \frac{1}{k} \sum_{s=1}^{k} y_s$. 

---

**Algorithm 1 CD-UCB**

**Require:** $T$, $\alpha$ and an algorithm CD($\cdot$, ·)

**Initialize** $\tau_i = 1$, $\forall i$.

**for** $t$ from 1 to $T$ **do**

**Update** according to equations (3-5).

Play arm $I_t$, and observe $X_i(I_t)$.

**if** $\text{CD}(I_t, X(I_t)) == 1$ **then**

$\tau_{I_t} = t + 1$; reset $\text{CD}(I_t, \cdot)$.

**end if**

**end for**

---

**Algorithm 2 Two-sided CUSUM**

**Require:** parameters $\epsilon$, $M$, $h$ and $\{y_k\}_{k \geq 1}$

**Initialize** $g_0^+ = 0$ and $g_0^- = 0$.

**for** each $k$ **do**

Calculate $s_k^+$ and $s_k^-$ according to (6).

Update $g_k^+$ and $g_k^-$ according to (7).

**if** $g_k^+ \geq h$ or $g_k^- \geq h$ **then**

Return 1

**end if**

**end for**
5 Performance Analysis

In this section, we analyze the performance in each part of the proposed algorithm: (a) our bandit algorithm (i.e., CD-UCB), and (b) our change detection algorithm (i.e., two-sided CUSUM). First, we present the regret upper bound result of CD-UCB for a given change detection guarantee. This is of independent interest in understanding the challenges of the non-stationary environment. Second, we provide performance guarantees of our modified CUSUM algorithm in terms of the mean detection delay, $E[D]$, and the expected number of false alarms up to time $T$, $E[F]$. Then, we combine these two results to provide the regret upper bound of our CUSUM-UCB. The proofs are presented in our supplementary material.

**Theorem 1. (CD-UCB)** Let $\xi = 1$. Under Assumption 1, for any $\alpha \in [0, 1)$ and any arm $i \in \{1, \ldots, K\}$, the CD-UCB policy achieves,

$$E[\tilde{N}_T(i)] \leq (\gamma_T + E[F]) \cdot \left(\frac{4\log T}{\alpha^2 T} + \frac{\pi^2}{3}\right) + \frac{\pi^2}{3} + \gamma_T \cdot E[D] + \frac{\alpha T}{K}.$$  

Recall that the regret of the CD-UCB policy is upper-bounded by $\sum_{i=1}^K E[\tilde{N}_T(i)]$. Therefore, given the parameter values (e.g., $\alpha$) and the performance of a change detection algorithm (i.e., $E[F]$ and $E[D]$), we can obtain the regret upper bound of that change detection based bandit algorithm. By letting $\alpha = 0$, we obtain the following result.

**Corollary 1. (CD-UCB $\alpha = 0$)** If $\alpha = 0$ and $\xi = 1$, then the regret of CD-UCB is

$$R_{\tau,\text{CD-UCB}}(T) = O((\gamma_T + E[F]) \cdot \log T + \gamma_T \cdot E[D])).$$  

**Remark 1.** If one can find an oracle algorithm that detects the change point with the properties that $E[F] \leq O(\gamma_T)$ and $E[D] \leq O(\log T)$, then one can achieve $O(\gamma_T \cdot \log T)$ regret, which recovers the regret result in Yu and Mannor (2009). We note that the WMD (Windowed Mean-shift Detection) change detection algorithm proposed by Yu and Mannor (2009) achieves these properties when side observations are available.

In the next proposition, we introduce the result of Algorithm 2 about the conditional expected detection delay and the conditional expected number of false alarms given $\hat{u}_0$. Note that the expectations exclude the first $M$ slots for initial observations.

**Proposition 1. (CUSUM $\hat{u}_0$)** Recall that $h$ is the tuning parameter in Algorithm 2. Under Assumptions 1 and 2, the conditional expected detection delay $E[D|\hat{u}_0 - u_0| < \epsilon]$ and the conditional expected number of false alarms $E[F|\hat{u}_0 - u_0| < \epsilon]$ satisfy

$$E[D|\hat{u}_0 - u_0| < \epsilon] \leq \frac{\alpha + 1}{\hat{u}_0 - u_0} - \epsilon,$$

$$E[F|\hat{u}_0 - u_0| < \epsilon] \leq \frac{2T}{\exp(r(\hat{u}_0))h},$$

where $r(\hat{u}_0) = \min(r^-(\hat{u}_0), r^+(\hat{u}_0))$, $r^-(\hat{u}_0)$ is the non-zero root of $\log E_{\hat{u}_0} [e^{r^+ M} + 1]$ and $r^+(\hat{u}_0)$ is the non-zero root of $\log E_{\hat{u}_0} [e^{r^+ M+1}]$. In the case of $|\hat{u}_0 - u_0| > \epsilon$, the algorithm restarts in at most $\frac{h+1}{|\hat{u}_0 - u_0| - \epsilon}$ time slots.

In the next theorem, we show the result for $E[D]$ and $E[F]$ when CUSUM is used to detect the abrupt change. Note again that the expectations exclude the first $M$ time slots.

**Theorem 2. (CUSUM)** Under Assumptions 1, 2 and 3, the expected detection delay $E[D]$ and the expected number of false alarms $E[F]$ of the Algorithm 2 satisfy

$$E[D] \leq C_2(h + 1),$$  

$$E[F] \leq \frac{2T}{(1 - 2 \exp(-2\epsilon^2 M)) \exp(C_1 h)},$$

where $C_2 \triangleq \log(3) + 2 \exp(-2\epsilon^2 M)/\lambda$, $C_1^- \triangleq \log \left(\frac{4\epsilon}{1 + \gamma T}\left(\frac{M}{2\epsilon M}\right)^{M + 1}\right)$, and $C_1^+ \triangleq \log \left(\frac{4\epsilon}{1 + \gamma T}\left(\frac{M}{2\epsilon M}\right)^{M + 1}\right)$.

Summing the result of Theorems 1 and 2, we obtain the regret upper bound of the CUSUM-UCB policy. To the best of our knowledge, this is the first regret bound for an actively adaptive UCB policy in the bandit feedback setting.

**Theorem 3. (CUSUM-UCB)** Let $\xi = 1$. Under Assumptions 1, 2 and 3, for any $\alpha \in (0, 1)$ and any arm $i \in \{1, \ldots, K\}$, the CUSUM-UCB policy achieves,

$$E[\tilde{N}_T(i)] \leq R_1 \cdot R_2 + \frac{\pi^2}{3} + \frac{\alpha T}{K},$$

for $R_1 = \gamma_T + \frac{2T}{(1 - 2 \exp(-2\epsilon^2 M)) \exp(C_1 h)}$, $R_2 = \frac{4\log T}{\Delta_{\mu T}^2} + \frac{\pi^2}{3} + M + \frac{C_2(h + 1)K}{\alpha}$.

**Corollary 2.** Under the Assumptions 1, 2 and 3, if horizon $T$ and the number of breakpoints $\gamma_T$ are known in advance, then we can choose $h = \frac{1}{C_1} \log \frac{T}{\gamma_T}$ and $\alpha = K \sqrt{\frac{C_2 \gamma_T}{C_1^+}} \log \frac{T}{\gamma_T}$ so that

$$R_{\tau,\text{CUSUM-UCB}}(T) = O\left(\frac{\gamma_T \log T}{(\Delta_{\mu T}^2)^2} + \sqrt{T} \gamma_T \log \frac{T}{\gamma_T}\right).$$

**Remark 2.** The choices of parameters depend on the knowledge of $\gamma_T$. This is common in the non-stationary bandit literature. For example, the discounting factor of D-UCB and sliding window size of SW-UCB depend on the knowledge of $\gamma_T$. The batch size of $R_{\text{exp3}}$ depends on the knowledge of $V_T$, which denotes the total reward variation. It is practically viable when the reward change rate is regular such that one can accurately estimate $\gamma_T$ based on history.

**Remark 3.** As shown in Garivier and Moulines (2008), the lower bound of the problem is $\Omega(\sqrt{T})$. Our policy approaches the optimal regret rate in an order sense.

**Remark 4.** For the SW-UCB policy, the regret analysis result is $R_{\tau,\text{SW-UCB}}(T) = O\left(\frac{\gamma_T \log T}{(\Delta_{\mu T}^2)^2}\right)$ (Garivier and
Moulines 2008). If $\Delta_{\mu_T(i)}$ is a constant with respect to $T$, then $\sqrt{T^2 \log T}$ term dominates and our policy achieves the same regret rate as SW-UCB. If $\Delta_{\mu_T(i)}$ goes to 0 as $T$ increases, then the regret of CUSUM-UCB grows much slower than SW-UCB.

Table 1 summarizes the regret upper bounds of the existing and proposed algorithms in the non-stationary setting when $\Delta_{\mu_T(i)}$ is a constant in $T$. Our policy has a smaller regret term with respect to $\gamma_T$ compared to SW-UCB.

### 6 Simulation Results

We evaluate the existing and proposed policies in three non-stationary environments: two synthetic dataset (flipping and switching scenarios) and one real-world dataset from Yahoo! News articles. Our PHT-UCB is similar to Adapt-EvE, but they are different in that Adapt-EvE ignores the issue of insufficient samples and includes other heuristic methods dealing with the detection points.

In the simulation, the parameters $h$ and $\alpha$ are tuned around $h = \log(T/\gamma_T)$ and $\alpha = \sqrt{T^2 \log(T)}$ based on the flipping environment. We suggest the practitioners to take the same approach because the choices of $h$ and $\alpha$ in Corollary 2 are minimizing the regret upper bound rather than the regret. We use the same parameters $h$ and $\alpha$ for CUSUM-UCB and PHT-UCB to compare the performances of CUSUM and PHT. Parameters are listed in Table 2 in Section G of the appendix. Note that $\epsilon$ and $M$ are obtained based on the prior knowledge of the datasets. The baseline algorithms are tuned similarly with the knowledge of $\gamma_T$ and $T$. We take the average regret over 1000 trials for the synthetic dataset.

#### 6.1 Synthetic Datasets

**Flipping Environment.** We consider two arms (i.e., $K = 2$) in the flipping environment, where arm 1 is stationary and the expected reward of arm 2 flips between two values. All arms are associated with Bernoulli distributions. In particular, $\mu_t(1) = 0.5$ for any $t \leq T$ and

$$
\mu_t(2) = \begin{cases} 
0.5 - \Delta, & \frac{T}{3} \leq t \leq \frac{2T}{3} \\
0.8, & \text{otherwise}
\end{cases}
$$

(16)

The two change points are at $T/3$ and $2T/3$. Note that $\Delta$ is equivalent to $\Delta_{\mu_T(2)}$. We let $\Delta$ vary within the interval $[0.02, 0.3]$, and compare the regrets of D-UCB, SW-UCB and CUSUM-UCB to verify Remark 4. For this reason, results of other algorithms are omitted. As shown in Figure 2a, CUSUM-UCB outperforms D-UCB and SW-UCB. In addition, the gap between CUSUM-UCB and SW-UCB increases as $\Delta$ decreases.

**Switching Environment.** We consider the switching environment, introduced by Mellor and Shapiro (2013), which is defined by a hazard function, $\beta(t)$, such that,

$$
\mu_t(i) = \begin{cases} 
\mu_{t-1}(i), & \text{with probability } 1 - \beta(t) \\
\mu \sim U[0, 1], & \text{with probability } \beta(t)
\end{cases}
$$

(17)

Note that $U[0, 1]$ denotes the uniform distribution over the interval $[0, 1]$ and $\mu_0(i)$ are independent samples from $U[0, 1]$. In the experiments, we use the constant hazard function $\beta(t) = \gamma_T/T$. All the arms are associated with a Bernoulli distribution.

The regrets over the time horizon are shown in Figure 2b. Although Assumptions 1 and 2 are violated, CUSUM-UCB and PHT-UCB outperform the other policies. To find the polynomial order of the regret, we use the non-linear least squares method to fit the curves to the model $at^b + c$. The resulting exponents $b$ of Exp3.R, D-UCB, Exp3.S, SW-UCB, Exp3.S, CUSUM-UCB and PHT-UCB are 0.92, 0.89,
Figure 3: Rewards and regret over the Yahoo! dataset with $K = 5$.

The regret curves are shown in Figure 3b. We again fit the curves to the model $at^b + c$. The resulting exponents $b$ of D-UCB, Rexp3, SW-UCB, Exp3.R, Exp3.S, CUSUM-UCB and PHT-UCB are 1, 1, 1, 0.81, 0.85, 0.69 and 0.79, respectively. The passively adaptive policies, D-UCB, SW-UCB and Rexp3, receive a linear regret for most of the time. CUSUM-UCB and PHT-UCB achieve much better performance and show sublinear regret, because of their active adaptation to changes. Another observation is that CUSUM-UCB outperforms PHT-UCB. The reason behind is that the Yahoo! dataset has more frequent breakpoints than the switching environment (i.e., high $\gamma_T$). Thus, the estimation $\hat{y}_k$ in PHT test may drift away before PHT detects the change, which in turn results in more detection misses and the higher regret.

Figure 4: Regret over the Yahoo! dataset with $K = 100$.

The regret curves are shown in Figure 3b. We again fit the curves to the model $at^b + c$. The resulting exponents $b$ of D-UCB, Rexp3, SW-UCB, Exp3.R, Exp3.S, CUSUM-UCB and PHT-UCB are 1, 1, 1, 0.88, 0.9, 0.85 and 0.9, respectively. The passively adaptive policies, D-UCB, SW-UCB and Rexp3, receive a linear regret for most of the time. CUSUM-UCB and PHT-UCB show robust performance in this larger scale experiment.

### 6.2 Yahoo! Dataset

**Yahoo! Experiment 1 ($K = 5$).** Yahoo! has published a benchmark dataset for the evaluation of bandit algorithms (Yahoo!). The dataset is the user click log for news articles displayed on the Yahoo! Front Page (Li et al. 2011). Given the arrival of a user, the goal is to select an article to present to the user, in order to maximize the expected click-through rate, where the reward is a binary value for user click. For the purpose of our experiment, we randomly select the set of 5 articles (i.e., $K = 5$) from a list of 100 permutations of possible articles which overlapped in time the most. To recover the ground truth of the expected click-through rates of the articles, we take the same approach as in Mellor and Shapiro (2013), where the click-through rates were estimated from the dataset by taking the mean of an article’s click-through rate every 5000 time ticks (the length of a time tick is about one second), which is shown in Figure 3a.

The regret of CUSUM-UCB and PHT-UCB shows the better sublinear function of time compared to the other policies. Another observation is that PHT-UCB performs better than CUSUM-UCB, although we could not find a regret upper bound for PHT-UCB. The reason behind is that the PHT test is more stable and reliable (due to the updated estimation $\hat{y}_k$) in the switching environment.

**Yahoo! Experiment 2 ($K = 100$).** We repeat the above experiment with $K = 100$. The regret curves are shown in Figure 4. We again fit the curves to the model $at^b + c$. The resulting exponents $b$ of D-UCB, Rexp3, SW-UCB, Exp3.R, Exp3.S, CUSUM-UCB and PHT-UCB are 1, 1, 1, 0.88, 0.9, 0.85 and 0.9, respectively. The passively adaptive policies, D-UCB, SW-UCB and Rexp3, receive a linear regret for most of the time. CUSUM-UCB and PHT-UCB achieve much better performance and show sublinear regret, because of their active adaptation to changes. Another observation is that CUSUM-UCB outperforms PHT-UCB. The reason behind is that the Yahoo! dataset has more frequent breakpoints than the switching environment (i.e., high $\gamma_T$). Thus, the estimation $\hat{y}_k$ in PHT test may drift away before PHT detects the change, which in turn results in more detection misses and the higher regret.

### 7 Conclusion

We propose a change-detection based framework for multi-armed bandit problems in the non-stationary setting. We study a class of change-detection based policies, CD-UCB, and provide a general regret upper bound given the performance of change detection algorithms. We then develop CUSUM-UCB and PHT-UCB, that actively react to the environment by detecting breakpoints. We analytically show that the regret of CUSUM-UCB is $O(\sqrt{T\gamma_T \log \frac{T}{\gamma_T}})$, which is lower than the regret bound of existing policies for the non-stationary setting. To the best of our knowledge, this is the first regret bound for actively adaptive UCB policies. Finally, we demonstrate that CUSUM-UCB outperforms existing policies via extensive experiments over arbitrary Bernoulli rewards and the real world dataset of webpage click-through rates.
Acknowledgment

This work has been supported in part by grants from the Army Research Office W911NF-14-1-0368 W911NF-15-1-0277, and MURI W911NF-12-1-0385, DTRA grant HDTRA1-14-1-0058, and NSF grant CNS-1719371.

References

Agrawal, S., and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In COLT, 39.1–39.26.

Alaya-Feki, A. B. H.; Moulines, E.; and LeCorneac, A. 2008. Dynamic spectrum access with non-stationary multi-armed bandit. In 2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications, 416–420. IEEE.

Allesiardo, R., and Féraud, R. 2015. Exp3 with drift detection for the switching bandit problem. In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. 1–7. IEEE.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing 32(1):48–77.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. Machine learning 47(2-3):235–256.

Basseville, M., and Nikiforov, I. V. 1993. Detection of abrupt changes: theory and application, volume 104. Prentice Hall Englewood Cliffs.

Besbes, O.; Gur, Y.; and Zeevi, A. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In Advances in neural information processing systems, 199–207.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721.

Buccapatnam, S.; Liu, F.; Eryilmaz, A.; and Shroff, N. B. 2017. Reward maximization under uncertainty: Leveraging side-observations on networks. arXiv preprint arXiv:1704.07943.

Cesa-Bianchi, N., and Lugosi, G. 2006. Prediction, learning, and games. Cambridge university press.

Gallager, R. G. 2012. Discrete stochastic processes, volume 321. Springer Science & Business Media.

Garivier, A., and Moulines, E. 2008. On upper-confidence bound policies for non-stationary bandit problems. arXiv preprint arXiv:0805.3415.

Hartland, C.; Baskiotis, N.; Gelly, S.; Sebag, M.; and Teytaud, O. 2007. Change point detection and meta-bandits for online learning in dynamic environments. CAp 237–250.

Hinkley, D. V. 1971. Inference about the change-point from cumulative sum tests. Biometrika 509–523.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In International Conference on Algorithmic Learning Theory, 199–213. Springer.

Khan, R. A. 1981. A note on page’s two-sided cumulative sum procedure. Biometrika 717–719.

Kocsis, L., and Szepesvári, C. 2006. Discounted ucb. In 2nd PASCAL Challenges Workshop, 784–791.

Lai, T. L., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. Advances in applied mathematics 6(1):4–22.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In Proceedings of the fourth ACM international conference on Web search and data mining, 297–306. ACM.

Li, S.; Karatzoglou, A.; and Gentile, C. 2016. Collaborative filtering bandits. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 539–548. ACM.

Lorden, G. 1971. Procedures for reacting to a change in distribution. The Annals of Mathematical Statistics 1897–1908.

Mellor, J., and Shapiro, J. 2013. Thompson sampling in switching environments with bayesian online change detection. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, 442–450.

Page, E. S. 1954. Continuous inspection schemes. Biometrika 41(1/2):100–115.

Pollard, D. 1984. Convergence of Stochastic Processes. Springer.

Srivastava, V.; Reverdy, P.; and Leonard, N. E. 2014. Surveillance in an abruptly changing world via multiarmed bandits. In Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on, 692–697. IEEE.

Sutton, R. S., and Barto, A. G. 1998. Reinforcement learning: An introduction, volume 1. MIT press Cambridge.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 237–250.

Yahoo! Webscope program. http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=49 [Online; accessed 18-Oct-2016].

Yu, J. Y., and Mannor, S. 2009. Piecewise-stationary bandit problems with side observations. In Proceedings of the 26th Annual International Conference on Machine Learning, 1177–1184. ACM.
A Lemma List

Lemma 1. (Chernoff-Hoeffding bound) Let $Y_1, \ldots, Y_n$ be random variables with common range $[0, 1]$ and such that $E[Y_i | Y_1, \ldots, Y_{i-1}] = u$. Let $S_n = Y_1 + \cdots + Y_n$. Then for all $a \geq 0$

$$\Pr \{ S_n \geq nu + a \} \leq e^{-2a^2/n},$$  \hspace{1cm} (18)

$$\Pr \{ S_n \leq nu - a \} \leq e^{-2a^2/n}. \hspace{1cm} (19)$$

Lemma 1 states the well known Chernoff-Hoeffding inequalities, proof of which is referred to (Pollard 1984). The next two lemmas are used in the proof of Theorem 2.

Lemma 2. (Wald’s identity) Let $\{Y_k; k \geq 1\}$ be independent and identically distributed, and let $\Lambda(r) = \log \{E[e^r Y_1] \}$. Let $J(Y)$ be the interval of $r$ over which $\Lambda(r)$ exists. For each $n \geq 1$, let $S_n = Y_1 + \cdots + Y_n$. Let $a < 0$ and $b > 0$ be arbitrary, and let $L$ be the smallest $n$ for which either $S_n \geq b$ or $S_n \leq a$. Then for each $r \in J(Y)$,

$$E[\exp (r S_L - \Lambda(r))] = 1. \hspace{1cm} (20)$$

Proofof Lemma 2 is referred to (Gallager 2012).

Lemma 3. (Two-sided CUSUM) Let $\tau_1, \tau_2, \ldots$ be independent random variables. For $\epsilon \geq 0$, define $s_k^+ = \eta_k - \epsilon$ and $s_k^- = -\eta_k - \epsilon$. Let $g_k^+ = \max(0, g_{k-1}^+ + s_k^+)$ and $g_k^- = \max(0, g_{k-1}^- + s_k^-)$. For $h > 0$, define $H^+ = \inf \{ k : g_k^+ \geq h \}$ and $H^- = \inf \{ k : g_k^- \geq h \}$, $H = \min(H^+, H^-)$. Then, we have that

$$\frac{1}{E[H]} = \frac{1}{E[H^+]} + \frac{1}{E[H^-]} \hspace{1cm} (21)$$

Note that the two-sided CUSUM algorithm (that signals a change at $H$) defined in Lemma 3 is called symmetric. Proof of Lemma 3 is referred to (Khan 1981)

B Proof of Theorem 1

Proof. Recall that $i_t$ is the arm with the best UCB index at time $t$, i.e., $i_t = \arg \max_{a \in \mathcal{A}} (\bar{X}_t(i) + C_t(i))$. In addition $r_t$ is the random arm sampled from uniform distribution. At each time $t$, if the arm $i$ is played, then the CD-UCB algorithm is either sampling a random arm ($r_t = i$) or playing the arm with the best UCB index ($i_t = i$). So, the probability that arm $i$ is chosen at time $t$ and arm $i$ is not the best arm is

$$\Pr \{ I_t = i \neq i^*_t \} \leq \alpha \Pr \{ r_t = i \} + (1 - \alpha) \Pr \{ i_t = i \neq i^*_t \} \leq \frac{\alpha}{K} + (1 - \alpha) \Pr \{ i_t = i \neq i^*_t \} \hspace{1cm} (22)$$

By the definition of $\hat{N}_T(i)$ in equation (1), we have that

$$E[\hat{N}_T(i)] \leq \sum_{t=1}^{T} \Pr \{ I_t = i \neq i^*_t \} \hspace{1cm} (23)$$

$$\leq \sum_{t=1}^{T} \left( \frac{\alpha}{K} + (1 - \alpha) \Pr \{ i_t = i \neq i^*_t \} \right) \hspace{1cm} (24)$$

$$\leq \alpha T/K + \sum_{t=1}^{T} \Pr \{ i_t = i \neq i^*_t \} \hspace{1cm} (25)$$

Now, it remains to bound the second term of (25), denoted by (a). Let $A_i$ be a constant defined as $A_i \triangleq \frac{\log T}{(\gamma_T + F_i)^2}$. Then we can decompose the event $\{ i_t = i \neq i^*_t \}$ as

$$\{ i_t = i \neq i^*_t, N_t(i) < A_i \} \cup \{ i_t = i \neq i^*_t, N_t(i) \geq A_i \}$$

Consider an experiment of the CD-UCB over $T$ plays. Let $F_i$ be the number of false alarms up to time $T$ and $D_i^+$ be the detection delay of $i$-th breakpoint on arm $i$, where $j \leq T$. Then, the total number of detection points, when the change detection algorithm CD$(\cdot, \cdot)$ signals an alarm on arm $i$, is upper-bounded by $\gamma_T + F_i$. Let $\tau_i(t)$ be the latest detection points (including false arms) up to time $t$. For each arm $i$, we define $T_i$ as the set of time slots that no breakpoint occurs after $A_i$ time slots away from the detection points.

$$T_i \triangleq \{ t \in \{1, \ldots, T\} : \mu_a(i) = \mu_t(i) \mbox{ and } \tau_i(t) < s, t \geq \tau_i(t) + A_i \} \hspace{1cm} (26)$$

Then, we can bound the term (b) $\triangleq \sum_{i=1}^{T} \{ i_t = i \neq i^*_t \}$ by

$$b = \sum_{t=1}^{T} \{ \{ i_t = i \neq i^*_t, N_t(i) < A_i \} \cup \{ i_t = i \neq i^*_t, N_t(i) \geq A_i \} \} \hspace{1cm} (27)$$

$$\leq (\gamma_T + F_i) A_i + \sum_{t=1}^{T} \{ i_t = i \neq i^*_t, N_t(i) \geq A_i \} \hspace{1cm} (27)$$

$$\leq (\gamma_T + F_i) A_i + \sum_{i=1}^{T} \{ i_t = i \neq i^*_t, N_t(i) \geq A_i \} \hspace{1cm} (27)$$

For each $t \in T_i$, the event $\{ i_t = i \neq i^*_t, N_t(i) \geq A_i \}$ implies the following union of events

$$\{ \overline{X}_t(i) \geq \mu_t(i) + C_t(i) \} \cup \{ \overline{X}_t(i^*_t) \leq \mu_t(i^*_t) - C_t(i^*_t) \}$$

$$\cup \{ \mu_t(i^*_t) - \mu_t(i) < 2C_t(i), N_t(i) \geq A_i \} \hspace{1cm} (28)$$

On the event $\{ N_t(i) \geq A_i \}$, we have that

$$C_t(i) = \sqrt{\frac{\xi \log n_t}{N_t(i)}} \leq \sqrt{\frac{\xi \log T}{A_i}} = \frac{\Delta_{\tau_T}(i)}{2} \hspace{1cm} (29)$$

by the definition of $A_i$. Thus, $2C_t(i) \leq \mu_t(i^*_t) - \mu_t(i)$ holds when $N_t(i) \geq A_i$. This implies that the last event of (28) is impossible. So we have that $\{ i_t = i \neq i^*_t, N_t(i) \geq A_i \}$ is at most

$$\{ \overline{X}_t(i) \geq \mu_t(i) + C_t(i) \} \cup \{ \overline{X}_t(i^*_t) \leq \mu_t(i^*_t) - C_t(i^*_t) \} \hspace{1cm} (30)$$

By the Lemma 1, we have that

$$\Pr \{ \overline{X}_t(i) \geq \mu_t(i) + C_t(i) \} \leq \frac{1}{n_t^{2\xi}} \hspace{1cm} (31)$$

$$\Pr \{ \overline{X}_t(i^*_t) \leq \mu_t(i^*_t) - C_t(i^*_t) \} \leq \frac{1}{n_t^{2\xi}} \hspace{1cm} (32)$$

Let $I_0, I_1, \ldots$ be the length of intervals between successive detection points. Then, we have that $\sum_{t=1}^{\gamma_T + F_i} I_m $. Let $L = 1$, we have that

$$\sum_{t=1}^{\gamma_T + F_i} I_m \leq \sum_{m=0}^{t} 2n_t^{2} \hspace{1cm} (33)$$

$$\leq \sum_{m=0}^{t} 2n_t^{2} \leq \sum_{m=0}^{t} \sum_{s=1}^{\gamma_T + F_i} I_m \hspace{1cm} (34)$$

$$\leq 2 \sum_{m=0}^{\gamma_T + F_i} n_t^{2}/6 = (\gamma_T + F_i + 1)^2/3. \hspace{1cm} (35)$$
Let $\mathbb{E}[F]$ be the expected number of false alarms in time horizon $T$, and $\mathbb{E}[D]$ be the expected detection delay. Summing all the results, we have $\mathbb{E}[a] = \mathbb{E}[(b)]$ and

$$(a) \leq (\gamma_T + \mathbb{E}[F])A_1 + \gamma_T\mathbb{E}[D]$$

$$+ \mathbb{E} \left[ \sum_{i \in I_k} P\{i_t = i \neq i^*_k, N_t(i) \geq A_i \} \right].$$

(36)

Combining (36) and (25), we obtain (8).

\[ \square \]

**C Proof of Proposition 1**

**Proof.** Let $y_1, y_2, \ldots$ be a sequence of independent random variables with bounded support $[0, 1]$. Before (after) the change, the random variable $y_k$ follows a distribution with parameter $\theta_0$ ($\theta_1$). Let $u_0$ ($u_1$) denote the mean before (after) the change. Under the Assumption 2, we have that $|u_0 - u_1| \geq 2\epsilon$. Note that we use the first $M$ samples to estimate the mean before the change by $\hat{u}_0$. Then, $s_k^+$ and $s_k^-$ become a non-trivial function of observation $y_k$. We define the expected time excluding the first $M$ time slots until the alarm occurs as the average run length (ARL), which is consistent with the literature in change detection problems (Basseville and Nikiforov 1993). In particular, let $L(\theta)$ be the ARL function defined as

$$L(\theta) = \mathbb{E}_{\theta} \left[ \inf \{ t : g_{i+M}^+ > h \text{ or } g_{i+M}^- > h \} \right].$$

(37)

It is clear that Algorithm 2 is a symmetric version of two-sided CUSUM-type algorithm (Khan 1981). Let $s_k^+$ and $g_k^+$ ($s_k^-$ and $g_k^-$) be the upper (lower) side CUSUM random walk. Then, we define the ARL function for lower and upper side CUSUM as

$$L^-(\theta) = \mathbb{E}_{\theta} \left[ \inf \{ t : g_{i+M}^- > h \} \right],$$

(38)

$$L^+(\theta) = \mathbb{E}_{\theta} \left[ \inf \{ t : g_{i+M}^+ > h \} \right].$$

(39)

By Lemma 3, we have that

$$\frac{1}{L(\theta)} = \frac{1}{L^-(\theta)} + \frac{1}{L^+(\theta)}.$$  

(40)

The ARL function $L(\theta)$ characterizes the detection delay and false alarm performances of our CUSUM algorithm. In particular, $L(\theta_0)$ is the mean detection delay, and $L(\theta_1)$ is the mean time between false alarms. Thus, $\mathbb{E}[D] = L(\theta_0)$ and $\mathbb{E}[F] = T/L(\theta_1)$. By (40), it remains to calculate $L^-(\theta)$ and $L^+(\theta)$. In the following, we show the results for $L^-(\theta)$ and obtain the results for $L^+(\theta)$ by symmetry.

Let $s_k^- = s_{k+1}^- + \ldots + s_{n+M}^-$ be a random walk. Let $H^-$ be the smallest $n$ such that $s_k^-$ crosses the bounded interval $[0, h]$, i.e., $H^- = \inf\{S_k^- < 0 \text{ or } S_k^- > h\}$. The procedure that monitors the stopping time that $S_k^-$ crosses the boundary is also called a sequential probability ratio test (SPRT). Then the lower side CUSUM can be viewed as a repeated SPRT such that CUSUM restarts the SPRT once $S_k^-$ crosses the 0 boundary (i.e., $S_k^- < 0$) and outputs an alarm the first time that $S_k^-$ crosses the $h$ boundary (i.e., $S_k^- > h$) (Page 1954). Let $P(0) \triangleq P\{S_h^- < 0\}$ be the probability that SPRT ends up with $S_h^- < 0$. Let $G$ be the number of SPRT tests that CUSUM runs until an alarm. Then $G - 1$ is a geometrically distributed random variable with parameter $P_0(0)$. Hence, we have that

$$L^-(\theta) = \mathbb{E}_\theta[H^-|S_h^- < 0]\mathbb{E}_\theta[G - 1] + \mathbb{E}_\theta[H^-|S_h^- > h]$$

$$= \mathbb{E}_\theta[H^-|S_h^- < 0]P_0(0) + \mathbb{E}_\theta[H^-|S_h^- > h]$$

(41)

$$= \frac{\mathbb{E}_\theta[H^-]}{1 - P_0(0)}.$$  

(42)

Let $\Lambda_\theta(r) \triangleq \log\{\mathbb{E}_\theta[e^{(r-h)\gamma}]\}$. Then, by Lemma 2 we have

$$\mathbb{E}_\theta[\exp(rS_h^- - h\Lambda_\theta(r))] = 1.$$  

(43)

We use the two implications from the Wald’s identity (43). For simplicity, we assume that $k \geq M$ when we consider the expectation of $s_k^-$ and $s_k^-$. First, taking the derivative of both sides and letting $r = 0$, we have that

$$\mathbb{E}_\theta[S_h^-] = \mathbb{E}_\theta[H^-]\mathbb{E}_\theta[s_k^-].$$

(44)

Second, letting $r = r^-(\theta)$ such that $\Lambda_\theta(r^-(\theta)) = 0$ and $r^-(\theta) \neq 0$, we have that

$$\mathbb{E}_\theta[\exp(r^-(\theta)S_h^-)] = 1.$$  

(45)

By the Assumption 1, $\hat{u}_0$ is the average of $M$ samples from distribution under $\theta_0$. By Lemma 1, we have that

$$P\{\hat{u}_0 - u_0 > \epsilon\} \leq 2e^{-2\epsilon^2M}.$$  

(46)

Now, we classify the possible scenarios into four different cases depending on $\hat{u}_0, u_0$ and $u_1$, under which we can derive the upper bound of the mean detection delay or the mean time between false alarms.

**Case 1:** $|\hat{u}_0 - u_0| < \epsilon$ and $u_1 < u_0$

Under $\theta_1$, we show the upper bound of $L^-(\theta_1)$ (mean detection delay of lower side CUSUM). Note that $\mathbb{E}_{\theta_1}[s_k^-] = \hat{u}_0 - u_1 - \epsilon > 0$. Then, by (42) and (44) we have that

$$L^-(\theta_1) = \frac{\mathbb{E}_{\theta_1}[H^-]}{1 - P_0(0)} = \frac{\mathbb{E}_{\theta_1}[S_h^-]}{\mathbb{E}_{\theta_1}[S_h^-]}(1 - P_0(0))$$

$$= \frac{\mathbb{E}_{\theta_1}[S_h^- - S_h^- > h]}{\mathbb{E}_{\theta_1}[S_h^-]}(1 - P_0(0))$$

$$= \frac{\mathbb{E}_{\theta_1}[S_h^- - S_h^- < 0]}{\mathbb{E}_{\theta_1}[S_h^-]}(1 - P_0(0))$$

$$\leq \frac{\mathbb{E}_{\theta_1}[S_h^- - S_h^- > h]}{\mathbb{E}_{\theta_1}[S_h^-]}$$

$$\leq \frac{h + 1}{\mathbb{E}_{\theta_1}[s_k^-]}.$$  

(47)

(48)

(49)

(50)

Under $\theta_0$, we show the lower bound of $L^-(\theta_0)$ (mean time between false alarms of lower side CUSUM). Note that $\mathbb{E}_{\theta_0}[s_k^-] = \hat{u}_0 - u_0 - \epsilon < 0$, which implies that $r^-(\theta_0) > 0$. 

By (45), we have that
\[
1 = \mathbb{E}_{\theta_0}[\exp(-\theta_0 \cdot S_{H-})] = \mathbb{P}_{\theta_0}(\{0\}) \mathbb{E}_{\theta_0}[(\exp((-\theta_0 \cdot S_{H-})|S_{H-} < 0)] + (1 - \mathbb{P}_{\theta_0}(\{0\})) \mathbb{E}_{\theta_0}[(\exp((-\theta_0 \cdot S_{H-})|S_{H-} > 0)] \\
+ (1 - \mathbb{P}_{\theta_0}(\{0\})) \mathbb{E}_{\theta_0}\left[(\exp((-\theta_0 \cdot S_{H-})|S_{H-} > h) \geq (1 - \mathbb{P}_{\theta_0}(\{0\})) \mathbb{E}_{\theta_0}[(\exp((-\theta_0 \cdot S_{H-})|S_{H-} > h)] \geq (1 - \mathbb{P}_{\theta_0}(\{0\})) \exp(-\theta_0 \cdot h) \tag{54}
\right]
\]

Note that $\mathbb{E}_{\theta_0}[H^-] \geq 1$. Hence, we have that
\[
L^-(\theta_0) = \frac{\mathbb{E}_{\theta_0}[H^-]}{1 - \mathbb{P}_{\theta_0}(\{0\})} \geq \frac{1}{1 - \mathbb{P}_{\theta_0}(\{0\})} \geq \exp((-\theta_0 \cdot h)). \tag{55}
\]

Under $\theta_0$, we obtain the lower bound of $L^+(\theta_0)$ (mean time between false alarms of upper side CUSUM) with the similar arguments. In particular, $\mathbb{E}_{\theta_0}[s_k^+] = u_0 - \hat{u}_0 - \epsilon < 0$, which implies that $r^+(\theta_0) > 0$. Then we have that
\[
L^+(\theta_0) \geq \exp(r^+(\theta_0)h). \tag{56}
\]

Let $r^+(\theta_0) = \min(r^+(\theta_0), r^-(\theta_0))$. By (40), we have that
\[
L(\theta_1) \leq L^-(\theta_1) \leq \frac{h + 1}{\mathbb{E}_{\theta_1}[s_k^+]}, \tag{57}
\]
\[
L(\theta_0) \geq \exp(r^+(\theta_0)h)/2. \tag{58}
\]

**Case 2:** $|\hat{u}_0 - u_0| < \epsilon$ and $u_1 > u_0$

Similarly, we obtain the results for upper side CUSUM as
\[
L^+(\theta_1) \leq \frac{h + 1}{\mathbb{E}_{\theta_1}[s_k^+]}, \tag{59}
\]
\[
L^+(\theta_0) \geq \exp(r^+(\theta_0)h). \tag{60}
\]

Under $\theta_0$, we check that $\mathbb{E}_{\theta_0}[s_k^+] = \hat{u}_0 - u_0 - \epsilon < 0$. Hence, we have that
\[
L^-(\theta_0) \geq \exp(-\theta_0 \cdot h). \tag{61}
\]

By (40), we have that
\[
L(\theta_1) \leq L^+(\theta_1) \leq \frac{h + 1}{\mathbb{E}_{\theta_1}[s_k^+]}, \tag{62}
\]
\[
L(\theta_0) \geq \exp(r^+(\theta_0)h)/2. \tag{63}
\]

**Case 3:** $\hat{u}_0 - u_0 > \epsilon$

When the estimate $\hat{u}_0$ is large ($\hat{u}_0 - u_0 > \epsilon$), $\mathbb{E}_{\theta_0}[s_k^+] = \hat{u}_0 - u_0 - \epsilon > 0$. Then we have that
\[
L(\theta_0) \leq L^-(\theta_0) \leq \frac{h + 1}{\mathbb{E}_{\theta_0}[s_k^+]]. \tag{64}
\]

**Case 4:** $u_0 - \hat{u}_0 > \epsilon$

When the estimate $\hat{u}_0$ is small ($u_0 - \hat{u}_0 > \epsilon$), $\mathbb{E}_{\theta_0}[s_k^+] = u_0 - \hat{u}_0 - \epsilon > 0$. Then we have that
\[
L(\theta_0) \leq L^+(\theta_0) \leq h + 1/\mathbb{E}_{\theta_0}[s_k^+]]. \tag{65}
\]

### D Proof of Theorem 2

Note that the bounds for $L(\theta)$, $L^-(\theta)$ and $L^+(\theta)$ in the Proposition 1 are the conditional expectations under $\theta$ given $\hat{u}_0$. By the law of total expectation, we can obtain the expected bound results by taking expectation over $\hat{u}_0$. Let $\mathbb{E}_{\hat{u}_0}$ denote the expectation over $\hat{u}_0$. We classify the possible scenarios into two cases, under which we take the conditional expectations and sum the total expectation finally.

**Case 1:** $|\hat{u}_0 - u_0| < \epsilon$

By the Proposition 1, given the condition $\hat{u}_0$ and $|\hat{u}_0 - u_0| < \epsilon$, we have that
\[
L(\theta_1) \leq \frac{h + 1}{|u_1 - \hat{u}_0| - \epsilon} \tag{66}
\]
\[
L^-(\theta_0) \geq \exp(-\theta_0 \cdot h) \tag{67}
\]
\[
L^+(\theta_0) \geq \exp(r^+(\theta_0)h) \tag{68}
\]

First, we consider the upper bound of $L(\theta_1)$ when $u_1 < u_0$. Let $p_Z(\cdot)$ denote the probability density function (or probability mass function) of random variable $Z$. Then, we have that
\[
\mathbb{E}_{\hat{u}_0}\left[\frac{h + 1}{|u_1 - \hat{u}_0| - \epsilon} |\hat{u}_0 - u_0| < \epsilon \right] = \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{h + 1}{|u_1 - u_0| - \epsilon} p_{\hat{u}_0}(u)du \tag{69}
\]
\[
= \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{h + 1}{u_1 - u_0 - \epsilon} p_{\hat{u}_0}(u)du \tag{70}
\]

We define $\delta_T(i) = \min\{|\mu_t(i) - \mu_{t+1}(i)| - 2\epsilon : \mu_t(i) \neq \mu_{t+1}(i), t \leq T\}$. Note that $\delta_T(i)$ is the minimum over a finite set of positive real numbers (i.e., $\delta_T(i) > 0$). Then, $u_0 - u_1 - \epsilon > \delta_T(i) + \epsilon$. Therefore, we have
\[
\int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{1}{u - u_1 - \epsilon} p_{\hat{u}_0}(u)du \tag{71}
\]
\[
\leq \int_{-\epsilon}^{\epsilon} \frac{1}{u + u_0 - u_1 - \epsilon} du \tag{72}
\]
\[
\leq \frac{1}{u + \delta_T(i)} + \epsilon \tag{73}
\]
\[
= \log(1 + 2\epsilon/\delta_T(i)). \tag{74}
\]

Hence, we have that
\[
\mathbb{E}_{\hat{u}_0}\left[\frac{h + 1}{|u_1 - \hat{u}_0| - \epsilon} |\hat{u}_0 - u_0| < \epsilon \right] \leq \frac{(h + 1)\log(1 + 2\epsilon/\delta_T(i))}{\mathbb{P}_{\hat{u}_0}(|\hat{u}_0 - u_0| < \epsilon)} \tag{75}
\]

The same result holds for $u_1 > u_0$. Note that $\delta_T(i) \geq \epsilon$ by the Assumption 2. Then, we have that
\[
\mathbb{E}_{\hat{u}_0}\left[\frac{h + 1}{|u_1 - \hat{u}_0| - \epsilon} |\hat{u}_0 - u_0| < \epsilon \right] \leq \frac{(h + 1)\log(3)}{\mathbb{P}_{\hat{u}_0}(|\hat{u}_0 - u_0| < \epsilon)} \tag{76}
\]
Second, we consider the lower bound of $L^-(\theta_0)$ and $L^+(\theta_0)$. By the Assumption 3, the $y_k$ is a Bernoulli random variable. Then, we have that

\[
\Lambda_{\theta_0}^-(r) = \log(e^{-r}u_0 + 1 - u_0) \geq r(\hat{u}_0 - \epsilon) \quad (79)
\]

\[
\Lambda_{\theta_0}^+(r) = \log(e^r u_0 + 1 - u_0) \geq -r(\hat{u}_0 + \epsilon) \quad (80)
\]

Let $\hat{r}^-(\theta_0)$ ($\hat{r}^+(\theta_0)$) be the solution of $\frac{d}{dr}\Lambda_{\theta_0}^-(r) = 0$ ($\frac{d}{dr}\Lambda_{\theta_0}^+(r) = 0$). Then, the convexity of $\Lambda_{\theta_0}^-(r)$ ($\Lambda_{\theta_0}^+(r)$) implies that $r^-(\theta_0) > \hat{r}^-(\theta_0)$ ($r^-(\theta_0) > \hat{r}^-(\theta_0)$). In addition, $\hat{r}^-(\theta_0)$ and $\hat{r}^+(\theta_0)$ satisfy

\[
\exp(\hat{r}^-(\theta_0)) = \frac{u_0(\hat{u}_0 - \epsilon)^{-1} - u_0}{1 - u_0} \quad (81)
\]

\[
\exp(\hat{r}^+(\theta_0)) = \frac{1 - u_0}{u_0(\hat{u}_0 + \epsilon)^{-1} - u_0} \quad (82)
\]

Note that $\hat{r}^-(\theta_0)$ and $\hat{r}^+(\theta_0)$ always exist when $2\epsilon < u_0 < 1 - 2\epsilon$. One can scale the reward linearly without changing the problem by function $f(x) = (1 - 4\epsilon)x + 2\epsilon$ so that the expected reward is within the interval $(2\epsilon, 1 - 2\epsilon)$. Hence, we assume that $\hat{r}^-(\theta_0)$ and $\hat{r}^+(\theta_0)$ exist without loss of generality. We first derive a lower bound for (81).

\[
\mathbb{E}_{\hat{u}_0} \left[ (\hat{u}_0 - \epsilon)^{-1} | \hat{u}_0 - u_0 | < \epsilon \right] = \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{1 - u_0}{\hat{u}_0 - u_0} p_{\hat{u}_0}(u) du \quad (83)
\]

\[
\geq \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{1 - u_0}{\hat{u}_0 - u_0} p_{\hat{u}_0}(u) du \geq \frac{1}{u_0} \quad (84)
\]

\[
\geq \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{1 - u_0}{\hat{u}_0 - u_0} p_{\hat{u}_0}(u) du \geq \frac{u_0}{\hat{u}_0 - u_0} + 1 \quad (85)
\]

\[
\geq \int_{u_0 - \epsilon}^{u_0 + \epsilon} \frac{1 - u_0}{\hat{u}_0 - u_0} p_{\hat{u}_0}(u) du \geq \frac{u_0}{\hat{u}_0 - u_0} + 1 \quad (86)
\]

\[
\geq \frac{u_0}{\hat{u}_0 - u_0} + 1 \quad (87)
\]

\[
\geq \frac{u_0}{\hat{u}_0 - u_0} + 1 \quad (88)
\]

\[
\geq \frac{u_0}{\hat{u}_0 - u_0} \left( \frac{M}{|u_0 M|} \right) \left( \frac{u_0 M}{(|u_0 M|)} \right) (2\epsilon)^M + 1 \quad (89)
\]

Hence, we have that

\[
\mathbb{E}_{\hat{u}_0} \left[ \exp(\hat{r}^-(\theta_0)) | \hat{u}_0 - u_0 | < \epsilon \right] \geq \frac{\epsilon}{u_0 - \epsilon} \left( \frac{M}{|2\epsilon M|} \right) (2\epsilon)^M + 1 \quad (90)
\]

\[
\geq \frac{4\epsilon}{(1 - \epsilon)^2} \left( \frac{M}{|2\epsilon M|} \right) (2\epsilon)^M + 1 \quad (91)
\]

Now, we derive a lower bound for (82). Similarly, we have that

\[
\mathbb{E}_{\hat{u}_0} \left[ \exp(\hat{r}^+(\theta_0)) | \hat{u}_0 - u_0 | < \epsilon \right] \geq \frac{4\epsilon}{(1 - \epsilon)^2} \left( \frac{M}{|2\epsilon M|} \right) (2\epsilon)^M + 1 \quad (92)
\]

Note that $\lfloor \cdot \rfloor$ denotes the floor function and $\lceil \cdot \rceil$ denotes the ceiling function.
By the Proposition 1, the average run length is at most
\[
\frac{h + 1}{\hat{u}_0 - u_0} - \epsilon.
\] (110)
Suppose that \(\hat{u}_0 > u_0 + \epsilon\), then
\[
E_{\hat{u}_0} \left[ \frac{h + 1}{\hat{u}_0 - u_0} \left| \hat{u}_0 - u_0 > \epsilon \right. \right] = \mathbb{E}_{\hat{u}_0} \left[ \frac{h + 1}{\hat{u}_0 - u_0} \left| \hat{u}_0 - u_0 > \epsilon \right. \right] \leq \mathbb{E}_{\hat{u}_0} \left[ \frac{h + 1}{(u_0 + \epsilon)M/M - u_0 - \epsilon} \left| \hat{u}_0 - u_0 > \epsilon \right. \right] \leq \frac{h + 1}{\lambda}. \] (114)
The same result holds when \(\hat{u}_0 < u_0 - \epsilon\). Summing the results (78) (108) (109) (114), we derive the upper bound for the mean detection delay and the lower bound for number of false alarms within the horizon \(T\).
\[
E[D] \leq E_{\hat{u}_0} \left[ \frac{h + 1}{u_1 - \hat{u}_0} | \hat{u}_0 - u_0 | < \epsilon \right] \mathbb{P}\{ | \hat{u}_0 - u_0 | < \epsilon \} \leq \frac{h + 1}{\lambda}. \] (117)
Note that the mean time between false alarm is \(E_{\hat{u}_0}[L(\theta_0)]\). By the law of total expectation, we have that
\[
E_{\hat{u}_0}[L(\theta_0)] \geq E_{\hat{u}_0}[L(\theta_0) | | \hat{u}_0 - u_0 | < \epsilon] \mathbb{P}\{| \hat{u}_0 - u_0 | < \epsilon \} \geq \frac{1}{2} \exp(C_1 h) \] (119)
Hence, we have that
\[
E[F] \leq \frac{T}{2} \exp(C_1 h) \mathbb{P}\{| \hat{u}_0 - u_0 | < \epsilon \} \leq \frac{1}{2T} \exp(C_1 h)(1 - 2 \exp(-2\epsilon^2 M)). \] (121)

E Proof of Theorem 3
The result follows the Theorem 1 and the Theorem 2. Since the results of the Theorem 2 do not include the first \(M\) time slots, we additionally count the regret of \(M\) for each detection point (alarm). The mean detection delay is at most scaled by dividing the sampling rate \(\alpha/K\).

F CUSUM-UCB policy
CUSUM-UCB policy is a CD-UCB policy with CUSUM as a change detection algorithm. In particular, it takes \(K\) parallel CUSUM algorithms as CD(\(\cdot\), \(\cdot\)) in CD-UCB. Since our

Algorithm 3 CUSUM-UCB

Require: time horizon \(T\), parameters \(\alpha, \epsilon, M\) and \(h\)
Initialize \(\tau_i = 1\) and \(cnt(i) = M\) for each \(i \in \mathcal{K}\).

for \(t\) from 1 to \(T\) do
    Update according to equations (3).
    if \(cnt(i) > 0\) for some \(i \in \mathcal{K}\) then
        \(I_t = i;\) \(cnt(i) = cnt(i) - 1\).
    else
        Update \(I_t\) according to equation (5).
    end if
    Play arm \(I_t\) and observe \(X_t(I_t)\).
    if CUSUM(I_t, X_t(I_t)) == 1 then
        \(\tau_{I_t} = t + 1;cnt(I_t) = M;\) reset CUSUM(I_t, \(\cdot\)).
    end if
end for

CUSUM algorithm needs to estimate the mean first, we let the policy finish the estimation with \(M\) observations as soon as possible. That is why we introduce the countdown timers \(cnt(\cdot)\) in Algorithm 3. The performance of CUSUM-UCB depends on the parameters \(\alpha\) and \(h\). We discuss the joint choices of \(\alpha\) and \(h\) in Sections 5 and 6.

G Simulation Parameters

Table 2: Parameter setting in the simulation

| Environment | \(K\) | \(T\)     | \(\gamma_T\) | \(\epsilon\) | \(M\) | \(h\) | \(\alpha\) |
|-------------|------|----------|--------------|-------------|------|------|---------|
| Flipping    | 2    | \(10^5\) | 2            | 0.1         | 100  | 50   | 0.001   |
| Switching   | 5    | \(10^5\) | 10           | 0.1         | 100  | 20   | 0.01    |
| Yahoo! 1    | 5    | \(5 \times 10^5\) | 32           | 0.005      | 100  | 200  | 0.024   |
| Yahoo! 2    | 100  | \(2 \times 10^5\) | 216          | 0.005      | 100  | 200  | 0.024   |

\(\dagger\): We count breakpoints when the difference in mean rewards is greater than \(\epsilon = 0.005\).