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 **?** (**?**), 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 (**?**), online recommendation systems (**?**) and computational advertisement (**?**). In the classical multi-armed bandit problem (**?**), 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 **?** (**?**). Several algorithms have been proposed and proven to achieve $O(\log T)$ regret, such as Thompson Sampling (**?**), ϵ_n-greedy and Upper Confidence Bound (UCB) (**?**). Variants of these bandit policies can be found in **?** (**?**).

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 (**?**) 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* (**?**; **?**; **?**) and *actively adaptive policies* (**?**; **?**; **?**).

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 **?** (**?**), 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 $O(\sqrt{\gamma_T T}\log T)$, where γ_T is the number of change points up to time T (**?**). 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 $O(\sqrt{\gamma_T T \log T})$. Exp3.S (**?**) also achieves the same regret bound, where a uniform exploration is mixed with the standard Exp3 (**?**) algorithm. Similarly, **?** (**?**) 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.[1] The increased regret of Rexp3 comes from the adversarial nature of the algorithm, which assumes that the environment changes every time slot in the worst case.

Second, *actively adaptive policies* adopt a change detection algorithm to monitor the varying environment and restart the bandit algorithms when there is an alarm. Adapt-EvE, proposed by **?** (**?**), employs a Page-Hinkley Test (PHT) (**?**) to detect change points and restart the UCB policy. PHT has also been used to adapt the window length of SW-UCL (**?**), which is an extension of SW-UCB in the multi-armed bandit with Gaussian rewards. However, the regret upper bounds of Adapt-EvE and adaptive SW-UCL are still open problems. These works are closely related to our work, as one can regard them as instances of our change-detection based framework. We highlight that one of our contributions is to provide an analytical result for such a framework. **?** (**?**) took a Bayesian view of the non-stationary bandit problem, where a stochastic model of the dynamic environment is assumed and a Bayesian online change detection algorithm is applied. Similar to the work by **?** (**?**), the theoretical analysis of the Change-point Thompson Sampling (CTS) is still open. Exp3.R (**?**) combines Exp3 and a drift detector, and achieves the regret $O(\gamma_T \sqrt{T \log T})$, which is not efficient when the change rate γ_T is high.

In sum, for various passively adaptive policies theoretical guarantees have been obtained, as they are considered more tractable to analyze. However, it has been demonstrated via extensive numerical studies that actively adaptive policies outperform passively adaptive policies (**?**). The intuition behind this is that actively adaptive policies can utilize the balance between exploration and exploitation by bandit algorithms, once a change point is detected and the environment stays stationary for a while, which is often true in real world applications. This observation motivates us to construct a change-detection based framework, where a class of actively adaptive policies can be developed with both good theoretical bounds and good empirical performance. Our main contributions are as follows.

1. We propose a change-detection based framework for a piecewise-stationary bandit problem, which consists of a change detection algorithm and a bandit algorithm. We develop a class of policies, CD-UCB, that uses UCB as a bandit algorithm. We then design two instances of this class, CUSUM-UCB and PHT-UCB, that exploit CUSUM (cumulative sum) and PHT as their change detection algorithms, respectively.

2. We provide a regret upper bound for the CD-UCB class, for given change detection performance. For CUSUM, we obtain an upper bound on the mean detection delay and a lower bound on the mean time between false alarms, and show that the regret of CUSUM-UCB is at most $O(\sqrt{T\gamma_T \log \frac{T}{\gamma_T}})$. 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 are 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 $\mathcal{K} = \{1, \ldots, K\}$ 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 $I_t \in \mathcal{K}$ 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 $\mu_t(i)$ denote the expectation of reward $X_t(i)$ at time slot t, i.e., $\mu_t(i) = \mathbb{E}[X_t(i)]$. Let i_t^* be the arm with highest expected reward at time slot t, denoted by $\mu_t(*) \triangleq \mu_t(i_t^*) = \max_{i \in \mathcal{K}} \mu_t(i)$. Let $\Delta_{\mu_T(i)} \triangleq \min\{\mu_t(*) - \mu_t(i) : t \leq T, i \neq i_t^*\}$, 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 π 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 π is measured in terms of the *regret*. The regret of π after T plays is defined as the expected total loss of playing suboptimal arms. Let $R_\pi(T)$ denote the regret of policy π 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 π during the first T plays.

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

$$\tilde{N}_T(i) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i,\ \mu_t(i) \neq \mu_t(*)\}}. \tag{2}$$

Note that the regret $R_\pi(T)$ of policy π 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 $\mathbb{E}[\tilde{N}_T(i)]$ to obtain a regret upper bound.

2.2 Piecewise-stationary Environment

We consider the notion of a *piecewise*-stationary environment in **?** (**?**), where the distributions of rewards remain constant for a certain period and abruptly change at some unknown time slots, called *breakpoints*. Let γ_T be the number of breakpoints up to time T, $\gamma_T = \sum_{t=1}^{T-1} \mathbb{1}_{\{\exists i \in \mathcal{K}: \mu_t(i) \neq \mu_{t+1}(i)\}}$. In addition, we make three mild assumptions for tractability.

[1] V_T satisfies $\sum_{t=1}^{T-1} \sup_{i \in \mathcal{K}} |\mu_t(i) - \mu_{t+1}(i)| \leq V_T$ for the expected reward of arm i at time t, $\mu_t(i)$.

Assumption 1. (Piecewise Stationarity) *The shortest interval between two consecutive 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 **?** (**?**) and a *switching* environment in **?** (**?**). 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 **?** (**?**) about the detectability in this paper.

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

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 (**?**; **?**; **?**; **?**). By Assumption 3, the empirical average of M Bernoulli random variables must be one of the grid points $\{0, 1/M, \dots, 1\}$. Let $\lambda_T(i) = \min\{(\mu_t(i) - \epsilon) - \lfloor(\mu_t(i) - \epsilon)M\rfloor/M, \lceil(\mu_t(i) + \epsilon)M\rceil/M - (\mu_t(i) + \epsilon) : t \le T\} \setminus \{0\}$ be the minimal non-trival gap between expectation and closest grid point of arm i.[2] We define the minimal gap of all arms as $\lambda = \min_{i \in \mathcal{K}} \lambda_T(i)$.

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 $I_t \in \mathcal{K}$ 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 (**?**). 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.

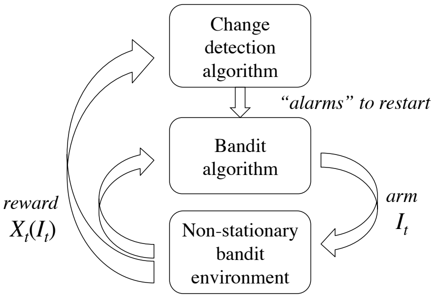


Figure 1: Change-detection based framework for non-stationary bandit problems

(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 tailored CUSUM 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, $\text{CD}(\cdot, \cdot)$, 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 $\text{CD}(i, \cdot)$ 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 $N_t(i)$. Let n_t be the total number of valid observations for the decision maker. For each arm i, let $\bar{X}_t(i)$ be the sample average and $C_t(i)$ be the

[2] Note that $\lfloor\cdot\rfloor$ denotes the floor function and $\lceil\cdot\rceil$ denotes the ceiling function.

Algorithm 1 CD-UCB

```
Require: $T$, $\alpha$ and an algorithm CD$(\cdot,\cdot)$
  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 $X_t(I_t)$.
    if CD$(I_t, X_t(I_t)) == 1$ then
      $\tau_{I_t} = t + 1$; reset 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
```

confidence padding term. In particular,

$$N_t(i) = \sum_{s=\tau_i}^{t} \mathbb{1}_{\{I_s = i\}}, \quad n_t = \sum_{i=1}^{K} N_t(i), \tag{3}$$

$$\bar{X}_t(i) = \sum_{s=\tau_i}^{t} \frac{X_s(i)}{N_t(i)} \mathbb{1}_{\{I_s=i\}}, \quad C_t(i) = \sqrt{\frac{\xi \log n_t}{N_t(i)}}, \tag{4}$$

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

$$I_t = \begin{cases} \arg\max_{i\in\mathcal{K}} \left(\bar{X}_t(i) + C_t(i)\right), & \text{w.p. } 1-\alpha \\ i, & \forall i \in \mathcal{K}, \text{w.p. } \frac{\alpha}{K} \end{cases}. \tag{5}$$

Parameter α controls the fraction of plays we exploit to feed the change detection algorithm. A large α may drive the algorithm to a linear regret performance while a small α can limit the detectability of change detection algorithm. We will discuss the choice of α 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 θ_0 and θ_1 are known for the density function $p(\cdot|\theta)$. In addition, y_k is sampled from distribution under θ_0 (θ_1) 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 (**?**), has been proven to be optimal in detecting abrupt changes in the sense of worst mean detection delay (**?**). 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, $\hat{u}_0 \triangleq (\sum_{k=1}^{M} y_k)/M$. 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

$$(s_k^+, s_k^-) = (y_k - \hat{u}_0 - \epsilon, \hat{u}_0 - y_k - \epsilon)\mathbb{1}_{\{k>M\}}. \tag{6}$$

Let g_k^+ (g_k^-) track the positive drift of upper (lower) random walk. In particular,

$$g_k^+ = \max(0, g_{k-1}^+ + s_k^+), \;\; g_k^- = \max(0, g_{k-1}^- + s_k^-). \tag{7}$$

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$(\cdot,\cdot)$ in CD-UCB. Formal description of CUSUM-UCB can be found in Algorithm **??**, provided in Section **??** in the supplementary material.

4.4 PHT-UCB policy

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

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, $\mathbb{E}[D]$, and the expected number of false alarms up to time T, $\mathbb{E}[F]$. 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 $i \in \{1, \dots, K\}$, the CD-UCB policy achieves,*

$$\mathbb{E}[\tilde{N}_T(i)] \leq \quad (\gamma_T + \mathbb{E}[F]) \cdot \left(\tfrac{4 \log T}{(\Delta_{\mu_T(i)})^2} + \pi^2/3\right) \qquad (8)$$
$$+ \tfrac{\pi^2}{3} + \gamma_T \cdot \mathbb{E}[D] + \tfrac{\alpha T}{K}.$$

Recall that the regret of the CD-UCB policy is upper-bounded by $\sum_{i=1}^{K} \mathbb{E}[\tilde{N}_T(i)]$. Therefore, given the parameter values (e.g., α) and the performance of a change detection algorithm (i.e., $\mathbb{E}[F]$ and $\mathbb{E}[D]$), 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*

$$R_{\pi^{\textit{CD-UCB}}}(T) = O((\gamma_T + \mathbb{E}[F]) \cdot \log T + \gamma_T \cdot \mathbb{E}[D])). \quad (9)$$

Remark 1. *If one can find an oracle algorithm that detects the change point with the properties that $\mathbb{E}[F] \leq O(\gamma_T)$ and $\mathbb{E}[D] \leq O(\log T)$, then one can achieve $O(\gamma_T \log T)$ regret, which recovers the regret result in* **?** *(***?***). We note that the WMD (Windowed Mean-shift Detection) change detection algorithm proposed by* **?** *(***?***) 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 $\mathbb{E}\left[D \middle| |\hat{u}_0 - u_0| < \epsilon\right]$ and the conditional expected number of false alarms $\mathbb{E}\left[F \middle| |\hat{u}_0 - u_0| < \epsilon\right]$ satisfy*

$$\mathbb{E}\left[D||\hat{u}_0 - u_0| < \epsilon\right] \leq \frac{h+1}{|u_1 - \hat{u}_0| - \epsilon}, \qquad (10)$$

$$\mathbb{E}\left[F||\hat{u}_0 - u_0| < \epsilon\right] \leq \frac{2T}{\exp(r(\theta_0)h)}, \qquad (11)$$

where $r(\theta_0) = \min(r^-(\theta_0), r^+(\theta_0))$, $r^-(\theta_0)$ is the non-zero root of $\log \mathbb{E}_{\theta_0}[e^{rs^-_{M+1}}]$ and $r^+(\theta_0)$ is the non-zero root of $\log \mathbb{E}_{\theta_0}[e^{rs^+_{M+1}}]$. 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 $\mathbb{E}[D]$ and $\mathbb{E}[F]$ 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 $\mathbb{E}[D]$ and the expected number of false alarms $\mathbb{E}[F]$ of the Algorithm 2 satisfy*

$$\mathbb{E}[D] \leq C_2(h+1), \qquad (12)$$

$$\mathbb{E}[F] \leq \frac{2T}{(1 - 2\exp(-2\epsilon^2 M))\exp(C_1 h)}, \qquad (13)$$

where $C_2 \triangleq \log(3) + 2\exp(-2\epsilon^2 M)/\lambda$, $C_1^- \triangleq \log\left(\frac{4\epsilon}{(1-\epsilon)^2}\binom{M}{\lfloor 2\epsilon M \rfloor}(2\epsilon)^M + 1\right)$, $C_1^+ \triangleq \log\left(\frac{4\epsilon}{(1+\epsilon)^2}\binom{M}{\lceil 2\epsilon M \rceil}(2\epsilon)^M + 1\right)$ and $C_1 \triangleq \min(C_1^-, C_1^+)$.

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 $i \in \{1, \dots, K\}$, the CUSUM-UCB policy achieves,*

$$\mathbb{E}[\tilde{N}_T(i)] \leq R_1 \cdot R_2 + \frac{\pi^2}{3} + \frac{\alpha T}{K}, \qquad (14)$$

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

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

$$R_{\pi^{\textit{CUSUM-UCB}}}(T) = O\left(\frac{\gamma_T \log T}{(\Delta_{\mu_T(i)})^2} + \sqrt{T \gamma_T \log \frac{T}{\gamma_T}}\right). \qquad (15)$$

Remark 2. *The choices of parameters depend on the knowledge of γ_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 γ_T. The batch size of Rexp3 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 γ_T based on history.*

Remark 3. *As shown in* **?** *(***?***), 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 $R_{\pi^{\textit{SW-UCB}}}(T) = O\left(\frac{\sqrt{T\gamma_T \log T}}{(\Delta_{\mu_T(i)})^2}\right)$ (***?***). If $\Delta_{\mu_T(i)}$ is a constant with respect to T, then $\sqrt{T\gamma_T \log T}$ 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 γ_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

Table 1: Comparison of regret bounds in various algorithms.

	Passively adaptive			*Actively adaptive*		
Policy	D-UCB (**?**)	SW-UCB (**?**)	Rexp3 (**?**)	Adapt-EvE (**?**)	CUSUM-UCB	*lower bound* (**?**)
Regret	$O(\sqrt{T\gamma_T}\log T)$	$O(\sqrt{T\gamma_T}\log T)$	$O(V_T^{1/3}T^{2/3})$	Unknown	$O(\sqrt{T\gamma_T\log\frac{T}{\gamma_T}})$	$\Omega(\sqrt{T})$

switching scenarios) and one real-world dataset from Yahoo! (**?**). Yahoo! dataset collected user click traces for 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 α are tuned around $h = \log(T/\gamma_T)$ and $\alpha = \sqrt{\frac{\gamma_T}{T}\log(T/\gamma_T)}$ based on the flipping environment. We suggest the practitioners to take the same approach because the choices of h and α in Corollary 2 are minimizing the regret upper bound rather than the regret. We use the same parameters h and α for CUSUM-UCB and PHT-UCB to compare the performances of CUSUM and PHT. Parameters are listed in Table **??** in Section **??** of the appendix. Note that ϵ and M are obtained based on the prior knowledge of the datasets. The baseline algorithms are tuned similarly with the knowledge of γ_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 $t \le T$ and

$$\mu_t(2) = \begin{cases} 0.5 - \Delta, & \frac{T}{3} \le t \le \frac{2T}{3} \\ 0.8, & \text{otherwise} \end{cases}. \qquad (16)$$

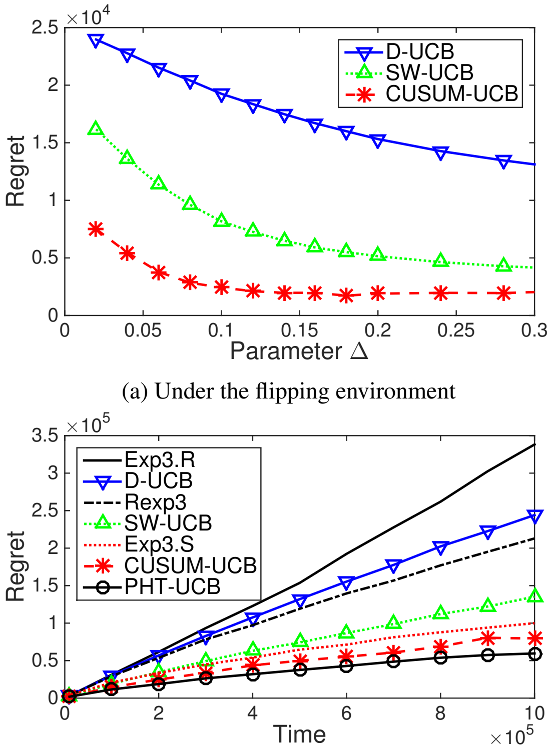
The two change points are at $\frac{T}{3}$ and $\frac{2T}{3}$. Note that Δ is equivalent to $\Delta_{\mu_T(2)}$. We let Δ vary within the interval $[0.02, 0.3]$, and compare the regrets of D-UCB, SW-UCB and CUSUM-UCB to verify Remark 4. For this reason, results of other algorithms are omitted. As shown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

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

$$\mu_t(i) = \begin{cases} \mu_{t-1}(i), & \text{with probability } 1 - \beta(t) \\ \mu \sim U[0,1], & \text{with probability } \beta(t) \end{cases}. \qquad (17)$$

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, Rexp3, SW-UCB, Exp3.S, CUSUM-UCB and PHT-UCB are 0.92, 0.89, 0.85, 0.84, 0.83, 0.72 and 0.69, respectively. 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.



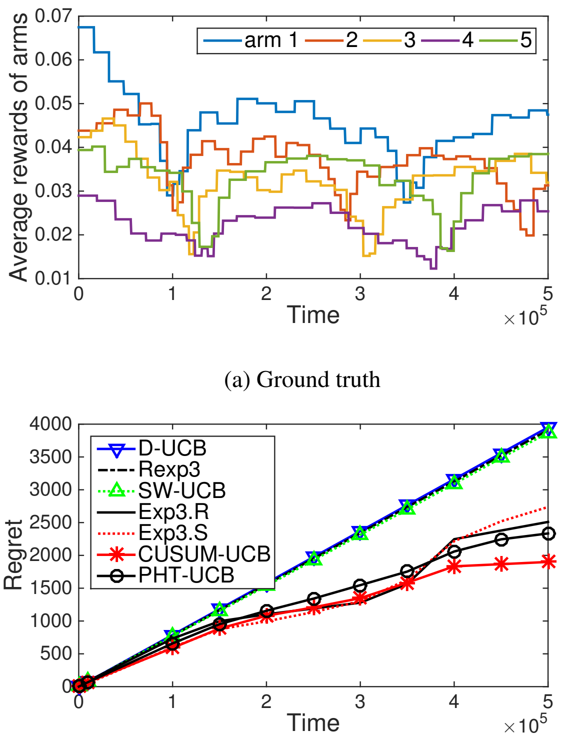
(a) Under the flipping environment

(b) Under the switching environment

Figure 2: Regret over synthetic datasets

6.2 Yahoo! Dataset

Yahoo! Experiment 1 ($K = 5$). Yahoo! has published a benchmark dataset for the evaluation of bandit algorithms (**?**). The dataset is the user click log for news articles



(a) Ground truth

(b) Regret

Figure 3: Rewards and regret over the Yahoo! dataset with $K = 5$

displayed on the Yahoo! Front Page (?). 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 **?** (**?**), 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 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 γ_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.

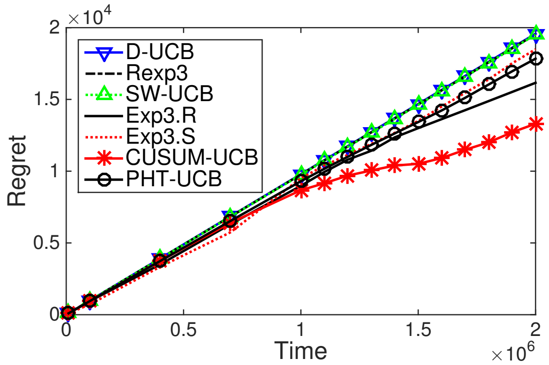


Figure 4: Regret over the Yahoo! dataset with $K = 100$

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 show robust performance in this larger scale experiment.

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.

Ducimus velit quam esse adipisci quas voluptatibus, neque impedit nobis dolor dolorem in eius obcaecati distinctio, pariatur odio aperiam cumque qui asperiores libero nesciunt fugiat vero quam id?Voluptatum laudantium necessitatibus enim quos eaque nemo officiis sit dolorum aliquam, impedit accusantium corrupti id saepe vitae possimus reiciendis.Ducimus ad nostrum esse odio perferendis ratione iure rerum, dolore error quas atque ad neque debitis vitae iusto, at quos inventore.Eius corporis quaerat, obcaecati mollitia deleniti repellendus et odit voluptatibus