Finite-time Analysis of Globally Nonstationary Multi-Armed Bandits

Junpei Komiyama
Stern School of Business
New York University
44 West 4th Street, New York, NY, USA

Edouard Fouché
Institute for Program Structures and Data Organization
Karlsruhe Institute of Technology
Am Fasanengarten 5, 76131 Karlsruhe, Germany

Junya Honda
Department of Systems Science, Graduate School of Informatics
Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

Editor: (editor names)

Abstract

We consider nonstationary multi-armed bandit problems where the model parameters of the arms change over time. We introduce the adaptive resetting bandit (ADR-bandit), which is a class of bandit algorithms that leverages adaptive windowing techniques from the data stream community. We first provide new guarantees on the quality of estimators resulting from adaptive windowing techniques, which are of independent interest in the data mining community. Furthermore, we conduct a finite-time analysis of ADR-bandit in two typical environments: an abrupt environment where changes occur instantaneously and a gradual environment where changes occur progressively. We demonstrate that ADR-bandit has nearly optimal performance when the abrupt or global changes occur in a coordinated manner that we call global changes. We demonstrate that forced exploration is unnecessary when we restrict the interest to the global changes. Unlike the existing nonstationary bandit algorithms, ADR-bandit has optimal performance in stationary environments as well as nonstationary environments with global changes. Our experiments show that the proposed algorithms outperform the existing approaches in synthetic and real-world environments.

Keywords: multi-armed bandits, adaptive windows, nonstationary bandits, change-point detection, sequential learning.

1. Introduction

1.1 Motivations

The multi-armed bandit (MAB; Thompson (1933); Robbins (1952)) is a fundamental model capturing the dilemma between exploration and exploitation in sequential decision making. This problem involves $K$ arms (i.e., possible actions). At each time step, the decision-maker selects a set of arms and observes a corresponding reward. The goal of the decision-maker...
is to maximize the cumulative reward over time. The performance of a bandit algorithm is usually measured via the notion of “regret”: the difference between the obtained rewards, and the rewards one would have obtained by choosing the best arms. Minimizing the regret corresponds to maximizing the expected reward.

The MAB has been used to solve numerous problems, such as experimental clinical design (Thompson, 1933), online recommendations (Li et al., 2010), online advertising (Chapelle and Li, 2011; Komiyama et al., 2015), and stream monitoring (Fouché et al., 2019). The most widely studied version of this model is the stochastic MAB, which assumes that the reward for each arm is drawn from an unknown but fixed distribution. In the stochastic MAB, several algorithms, such as the upper confidence bound (UCB; Lai and Robbins (1985); Auer et al. (2002)) and Thompson sampling (TS; Thompson (1933)) are known to have $\Theta(\log T / \Delta_{\text{min}})$ regret, which is optimal (Lai and Robbins, 1985). While it is reasonable to assume in some cases that the reward-generating process does not change, as in these algorithms, the distribution of rewards may change over time in many applications.

To observe this, we consider the following two examples:

**Example 1** (Online advertising) A website has several advertisement slots. Based on each user’s query, the website decides which ads to display from a set of candidates (i.e., “relevant advertisements”). Some advertisements are more appealing to a user than others. Each advertisement is associated with a click-through rate (CTR), the number of clicks per view. Websites receive revenue from clicks on advertisements; thus, maximizing the CTR maximizes the revenue. This problem can be considered an instance of the bandit problem, in which advertisements and clicks correspond to arms and rewards, respectively. However, it is well known that the CTR of some advertisements may change over time for several reasons, such as seasonality or changing user interests. In this case, naively applying a stochastic MAB algorithm leads to suboptimal rewards.

**Example 2** (Predictive maintenance) Correlation often results from physical relationships, for example, between the temperature and pressure of a fluid. Changes in correlation often indicate that the system is transitioning into another state (e.g., the fluid solidifies) or that equipment deteriorates or fails (e.g., a leak). When monitoring large factories, maintaining an overview of correlations can lower the maintenance costs by predicting abnormalities. However, continuously updating the full correlation matrix is computationally impractical because the data are typically high-dimensional and are ever-evolving. A more efficient solution consists of updating only a few elements of the matrix based on a notion of utility (e.g., high correlation values). The system must minimize the cost of monitoring while maximizing the total utility in a possibly nonstationary environment. In other words, correlation monitoring can be considered an instance of the bandit problem (Fouché et al., 2019), in which pairs of sensors and correlation coefficients correspond to arms and rewards.

In such settings, the reward may evolve over time (i.e., it is nonstationary\(^2\)). The nonstationary MAB (NS-MAB) describes a class of MAB algorithms addressing this particular setting. Most of the NS-MAB algorithms rely on passive forgetting methods based on a

---

1. The value $\Delta_{\text{min}}$ is a distribution-dependent constant quantifying the hardness of the problem instance.
2. The use of the term “nonstationary” in the bandit literature is different from the literature on time series analysis. We formally define stationary and nonstationary streams in Definition 1.
sliding window (Garivier and Moulines, 2008) or fixed-time resetting (Gur et al., 2014). Recent work has proposed more sophisticated change detection mechanisms based on adaptive windows (Srivastava et al., 2014) or sequential likelihood ratio tests (Besson et al., 2020). However, the existing methods come with several drawbacks, as we describe later.

1.2 Challenges in nonstationary bandits

Although change detectors can help MAB algorithms to adapt to changing rewards, they often come with costs. Let us consider the case of an abruptly changing environment, where the reward distributions change drastically at some time steps. Previous work by Garivier and Moulines (2008) indicated that the regret of any \( K \)-armed bandit algorithm for such a case is \( \Omega(\sqrt{T}) \). This finding implies that the performance of NS-MAB algorithms in stationary (i.e., non-changing) environments is inferior to the performance of standard stochastic MAB algorithms, such as the UCB and TS, because \( O(\sqrt{T}) \) is much larger than \( O((\log T)/\Delta_{\min}) \) given a moderate value of \( \Delta_{\min} \). Virtually all NS-MAB algorithms conduct \( O(\sqrt{T}) \) forced exploration for all the arms, which is the leading factor of regret—no matter what change-point detection algorithm is used.

Another drawback of the existing methods is that they require several parameters that are highly specific to the problem and require unreasonable assumptions about the environment, such as the number of changes or an estimation of the amount of “nonstationarity” in the stream. If such parameters are not set correctly, then the actual performance may widely deviate from the given theoretical bounds.

In this paper, we solve these issues by introducing adaptive resetting bandit (ADR-bandit) algorithms, which is a new class of bandit algorithms. Several algorithms, such as UCB and TS, have been established in the stationary case; therefore, our idea is to extend them without introducing any forced exploration. For this purpose, we combine them with an adaptive windowing technique. For example, ADR-TS, which is an instance of the ADR-bandit, combines the adaptive windowing with TS. Our method deals with a subclass of changes that we call global changes. Intuitively, if all arms change in a coordinated manner, we can avoid forced exploration to improve performance. This type of change is natural in the predictive maintenance of Example 2, as illustrated in Figure 1, where the nonstationarity often results from changes in the entire system.

For the sake of generality, our analysis considers the more general setting of the multiple-play MAB (MP-MAB), in which the forecaster may play \( L > 1 \) arms per time step (Komiyama et al., 2015). One of the main challenges in MP-MAB is comparing each arm with the top \( L \)-th arm to identify the suboptimality of it. Moreover, in a nonstationary environment, we must maintain the candidate of the best arms for successful change detection, which is reflected in our design of the algorithm in Section 5.3. Naturally, our results also hold in the traditional single-play MAB setting. We demonstrate that the proposed method has optimal performance for stationary streams and comparable bounds with existing NS-MAB algorithms for abruptly changing and gradually changing environments under the global change assumption. Finally, our method is nearly parameter-free and very

---

3. This holds even when the change is sufficiently large, regardless of the value of a distribution-dependent constant \( \Delta_{\min} \). See details in Theorem 13, Corollary 14, and Remark 17 in Garivier and Moulines (2008).
Figure 1: In the Bioliq dataset, each arm corresponds to the correlation coefficient between a pair of sensors. As we can see, the reward between different arms tend to change in a coordinated manner. We describe this dataset in details in Section 6.

Table 1: The table below compares the achievable performance of the existing NS-MAB algorithms, ADR-bandit, and stationary MAB algorithms (e.g., TS and UCB). ADR-bandit has optimal performance in the stationary environments as well as nonstationary environments with global changes. Most existing NS-MAB algorithms only handle abrupt changes and do not provide any performance guarantees under gradual changes. Performance bounds under the gradual changes are provided in Besbes et al. (2014); Allesiardo and Féraud (2015); Wei et al. (2016); Wei and Srivastava (2018); Trovè et al. (2020). Here, $O$ is a Landau notation that ignores a polylogarithmic factor.

|                     | Stationary | Abrupt | Gradual |
|---------------------|------------|--------|---------|
| Existing NS-MAB     | $O(\sqrt{T})$ | $O(\sqrt{T})$ | $O(T^{1-d/3})$ |
| ADR-bandit          | $O(\log T/\Delta_{\min})$ | $O(\sqrt{T})$ (Under GC) | $O(T^{1-d/3})$ (Under GC) |
| Stationary MAB      | $O(\log T/\Delta_{\min})$ | $O(T)$ | $O(T)$ |
1.3 Contributions

We articulate our contributions as follows. First, we provide an extensive analysis of adaptive windowing techniques. We focus in particular on the adaptive windowing (ADWIN) algorithm (Bifet and Gavaldà, 2007), which is still considered state-of-the-art in the data stream mining community. The ADWIN algorithm performs well regarding various types of streams (Gama et al., 2014). Bifet and Gavaldà (2007) provided false positive and false negative rate bounds. However, existing analyses are not sufficient for our aim. For the analysis of bandit algorithms using ADWIN, we need an estimate on the accuracy of the estimator $\hat{\mu}_t$. Thus, we conduct a finite-time analysis on the estimation error $|\mu_t - \hat{\mu}_t|$ (Section 3). As a by-product, this analysis explains why adaptive windowing methodologies perform well in many stream learning problems by bounding the error for abrupt and gradual changes.

Then, after formalizing the MP-MAB problem (Section 4), we introduce the ADR (ADaptive Resetting) bandit algorithm (Section 5). Our study builds on a recent paper by Fouché et al. (2019) that applies the MAB to sensor streams. Although Fouché et al. (2019) proposed a way to combine the stationary bandit algorithms (e.g., UCB and TS) with adaptive windows, they did not provide any theoretical guarantees for the nonstationary case. We slightly modify their framework to enable rigorous analysis. Our paper fills the gap between the practical utility and theoretical understanding of such techniques. To our knowledge, the characterization of global changes and associated analyses are novel in the literature on nonstationary bandit algorithms.

Finally, we demonstrate the performance of our proposed method concerning synthetic and real-world environments (Section 6). The proposed method outperforms the existing nonstationary bandits not only in stationary streams but also in abruptly and gradually changing environments.

2. Related Work

2.1 Change detection

Change detection is an important problem in data mining. The goal is to detect when the statistical properties (e.g., the mean) of a stream of observations change. Such changes are commonly attributed to the phenomenon known as concept drift (Gama et al., 2014; Lu et al., 2019), which characterizes unforeseeable changes in the underlying data distribution. Detecting such changes is crucial to virtually any monitoring tasks in streams, such as controlling the performance of online machine learning algorithms (Bifet and Gavaldà, 2007) or detecting correlation changes (Seliniotaki et al., 2014).

There are numerous methods to detect changes. The fundamental idea is to measure whether the estimated parameters of the current data distribution (e.g., its mean) have changed at any time. In other words, change detection is about separating the signal from the noise (Gama and Castillo, 2006).

Change-point detection approaches can be classified into three major categories (cf. Table II in Gama et al. (2014)). These categories are sequential analysis approaches (e.g., CUSUM (Page, 1954) and its variant Page-Hinkley (PH) testing (Hinkley, 1971)), statisti-

---

4. Although, rigorously speaking, their analysis is not correct, as we will discuss later.
process control (e.g., the Drift Detection Method (DDM) (Gama et al., 2004) and its variants), and monitoring two distributions (e.g., ADWIN (Bifet and Gavaldà, 2007)).

In this work, we primarily consider ADWIN (Bifet and Gavaldà, 2007) because it works with any bounded distribution and has a good affinity with online learning analyses. Moreover, ADWIN monitors the mean from a sequence of observations over a sliding window of adaptive size. When ADWIN detects a change between two subwindows, the oldest observations are discarded. Otherwise, the window grows indefinitely. The success of this approach is due to the quasi-absence of parameters, making it highly adaptive.

Gonçalves et al. (2014) empirically compared ADWIN with other drift detectors. They found that ADWIN is one of the fastest detectors and is one of the only methods to provide false positive and false negative (fp/fn) rate guarantees. Many of the existing methods do not provide any performance guarantees, which is the case for DDM (Gama et al., 2004), EDDM (Baena-García et al. 2006), and ECDD (Ross et al., 2012) (per Blanco et al. (2015)). Although we believe that one can provide fp/fn rate guarantees for the statistical process control and CUSUM (Page, 1954) approaches by choosing the appropriate parameters, limited discussions are available on the theoretical properties of these algorithms. For a thorough history and comparison of change-detection methods, we refer to the surveys by Gama et al. (2014); Krawczyk et al. (2017).

2.2 Nonstationary bandits

In comparison, theoretical performance guarantees (i.e., regret bounds) are much more emphasized in the MAB literature. Traditionally, NS-MAB algorithms are divided into two categories. Active methods actively seek to detect changes, and passive methods do not. Another type of algorithms, known as adversarial bandits (Auer et al., 1995), such as Exp3, can deal with changing environments. However, the guarantee of adversarial bandit algorithms is limited when no arm is consistently good. In the following paragraphs, we discuss the related work on active and passive methods.

Active methods: Hartland et al. (2007) proposed Adapt-EvE, which combines the PH test with UCB, and Mellor and Shapiro (2013) suggested change-point TS. However, these two studies do not provide any regret bounds. Liu et al. (2018) proposed CUSUM-UCB and PH-UCB, which combine the UCB algorithm with a CUSUM-based (or PH-based) resetting and forced exploration. They derived a regret bound of $\tilde{O}(\sqrt{MT})$ in an abruptly changing environment with $M$ known change points. Cao et al. (2019) suggested the M-UCB algorithm, which combines UCB with an adaptive-window-based resetting method with a regret bound of $\tilde{O}(\sqrt{MT})$ for an abruptly changing environment. Allesiardo and Féraud (2015) proposed Exp3.R, which combines Exp3 (Auer et al., 1995) with change-point detection and provides a $\tilde{O}(M\sqrt{T})$ regret bound. Although the assumptions are slightly different among algorithms, many of the regret bounds are of the order of $\tilde{O}(\sqrt{MT})$ concerning the number of change points $M$ and the number of time steps $T$. Moreover, Seznec et al. (2020) applied the adaptive window to the rotting bandit problem where the reward of the arms only decreases. One of the closest work to ours is a manuscript by Mukherjee and Maillard (2019). They consider global changes in abrupt environments and proposed ImpCPD that combines successive rejection and UCB without introducing forced exploration. Our ADR-bandit is more general as we can deal with not only abrupt but
also gradual changes. Besson et al. (2020) proposed GLR-klUCB, a combination of KL-UCB (Garivier and Cappé, 2011) with the Bernoulli Generalized Likelihood Ratio (GLR) test (Maillard, 2017) and forced exploration for an abruptly changing environment. They discussed global and local resets and provided a regret bound of $\tilde{O}(\sqrt{MT})$.\(^5\)

**Passive methods:** Kocsis and Szepesvári (2006) proposed Discounted UCB (D-UCB). Garivier and Moulines (2008) suggested Sliding Window UCB (SW-UCB) and analyzed D-UCB to demonstrate that the two algorithms have a $\tilde{O}(\sqrt{MT})$ regret bound for abruptly changing environments. Besbes et al. (2014) considered the case of limited variation $V$ and proposed the Rexp3 algorithm with the regret bound $\Theta(V^{1/3}T^{2/3})$. Wei et al. (2016) generalized the analysis of Besbes et al. (2014) to the case of intervals where each interval has limited variation. Wei and Srivastava (2018) proposed the LM-DSEE and SW-UCB# algorithms with a $\tilde{O}(\sqrt{MT})$ regret bound for an abruptly changing environment. The latter algorithm adopts an adaptive window with a limited length. They also analyzed the case of a slowly changing environment and derived a regret bound of SW-UCB#. Trovò et al. (2020) proposed the Sliding Window TS (SW-TS) algorithm and provided a $\tilde{O}(\sqrt{MT})$ regret bound for an abruptly changing environment. They also provided a distribution-dependent regret bound for a gradually changing environment.

Our approach is an active one; however, it does strikingly differ from the existing methods. Our algorithm does not sacrifice performance in the stationary case, whereas almost all existing NS-MAB algorithms are exclusively designed for nonstationary environments. In real-world settings, users may not know whether a given stream of data must be considered stationary or not. Similarly, the nature of the stream may also change over time. In such settings, our approach is advantageous, as our experiments show. The existing studies typically only consider the single-play MAB problem, whereas our results also hold in the more general MP-MAB setting, as in Komiyama et al. (2015).

Although the analysis that follows is quite involved, our algorithms are conceptually simple and aimed at practical use. In comparison, our competitors tend to require more computation and rely on parameters that are difficult to set.

### 3. Analysis of ADWIN

This section analyzes ADWIN (Bifet and Gavaldà, 2007). This algorithm monitors at any time $t$ an estimate of the mean $\hat{\mu}_t$ from a single stream of univariate observations.

#### 3.1 Data streams

We assume that each observation $x_t \in [0, 1]$ at any time step $t$ is drawn from some distribution with the mean $\mu_t$. The value $\mu_t$ is not known to the algorithm, and ADWIN estimates it from a sequence of (possibly noisy) observations $S : (x_1, x_2, \ldots, x_T)$. The goal of ADWIN is to maintain an estimator of $\mu_t$ at any $t$ based on past observations. Due to concept drift (Gama et al., 2014), the mean $\mu_t$ may change over time, so the task is not trivial.

The data stream literature (Gama et al., 2014) identifies the four main categories of concept drifts: abrupt, recurring, gradual, and incremental drifts. Figure 2 indicates that the mean $\mu_t$ (the dashed line in the figure) changes abruptly and stays the same for some

---

\(^5\) They also discussed dependence on the distribution-dependent constants (Corollary 6 therein).
time with abrupt and recurring drifts, whereas $\mu_t$ changes gradually over time with gradual and incremental drifts. The actual observations $x_t$ can be arbitrarily noisy. We are only interested in the change in the mean $\mu_t$; thus, we simplify our analysis to the two types of changes. In addition, we also consider stationary streams where the mean $\mu_t$ never changes over time. In summary, we consider three types of streams:

**Definition 1** (Stationary, abrupt, and gradual streams)

1. The stationary stream is defined as $\mu_t = \mu$.
2. The abruptly changing stream is defined as $\mu_t = \mu_{t+1}$ except for changepoints $T_C \subset [T]$.
3. The gradually changing stream is defined as $|\mu_{t+1} - \mu_t| \leq b$ for some (unknown) constant $b \in (0, 1)$.

These definitions refer to the mean $\mu_t$. The observations $\{x_1, x_2, \ldots\}$ can be noisy.

### 3.2 The adaptive window algorithm

**Algorithm 1 ADWIN**

**Require:** A univariate stream of values $S : (x_1, x_2, \ldots) \in [0, 1]$, confidence level $\delta \in (0, 1)$.

1. $W(1) = \{\}$
2. for $t = 1, 2, \ldots$ do
3.     $W(t + 1) = W(t) \cup \{t\}$
4.     while $|\hat{\mu}_{W_1} - \hat{\mu}_{W_2}| \geq e_{\text{cut}}^\delta$ holds for some split $W(t + 1) = W_1 \cup W_2$ do
5.         $W(t + 1) = W_2$.
6.     end while
7. end for

---

6. This categorization matches the MAB literature (Wei and Srivastava, 2018; Trovò et al., 2020).

---

Figure 2: We classify the four types of concepts drifts (abrupt, recurring, gradual, incremental) into two types of changes: abrupt and gradual changes. Here, dashed lines represent $\mu_t$, whereas solid lines represent $x_t$. 
Globally Nonstationary Bandits

Figure 3: Terminology of changepoint and detection time. A changepoint of an abruptly changing stream is a time step $t$ where $\mu_t \neq \mu_{t+1}$. A detection time is the time step where ADWIN shrinks the window.

Algorithm 1 is the pseudo-code for ADWIN (Bifet and Gavaldà, 2007). The law of large numbers implies that using more observations helps estimate the parameter. However, the nature of older observations might be different from more recent observations. To determine a good trade-off between these two effects, ADWIN maintains a window $W(t)$ of past time and discards data points outside the window.

We omit the index $(t)$ of $W(t)$ when the time step of interest is clear. At each time step $t$, ADWIN receives a new observation $x_t$ and extends the window $W$ to include the observation. For the current window $W$, we let $\hat{\mu}_W = \frac{1}{|W|} \sum_{t \in W} x_t$ be the corresponding mean. For each new observation, ADWIN tests whether the mean of the underlying distribution has changed. If we can split $W$ into two consecutive disjoint subwindows $W_1 \cup W_2 = W$ whose means are significantly different (i.e., by some threshold $\epsilon_{\text{cut}}$), then ADWIN discards $W_1$ (i.e., ADWIN “shrinks” the window):

$$
\epsilon_{\text{cut}}^2 = \sqrt{\frac{1}{2|W_1|} \log \left(\frac{1}{\delta}\right)} + \sqrt{\frac{1}{2|W_2|} \log \left(\frac{1}{\delta}\right)},
$$

where $|W|$ is the cardinality of a set $W$. The threshold $\epsilon_{\text{cut}}$ is based on Hoeffding’s inequality\(^{7}\). At each time step, ADWIN check every possible split of $W$.

The following definition formalizes the notion of a window and the shrinking of the current window. See also Figure 3 for illustration.

**Definition 2** (Detection times) The time step $t$ is a detection time of ADWIN if $|W(t + 1)| \leq |W(t)|$. In addition, ADWIN “shrinks” the window at time $t$ if time step $t$ is a detection time. We define the breakpoint as the last round of $W_2$ from the split $W = W_1 \cup W_2$ (Figure 3). We let $T_d$ be the set of all detection times.

**Remark 3** Unlike the set of changepoints $T_c$ (Definition 1) that is only defined for an abrupt stream and is independent of any underlying algorithm, $T_d$ is defined for both abrupt and gradual stream. The detection time is a random variable that is defined via ADWIN.

\(^7\) Note that $\epsilon_{\text{cut}}$ above is slightly different from the original ADWIN where the harmonic mean of $|W_1|$ and $|W_2|$ is used.
3.2.1 Bound for the estimator of the mean $\mu_t$

Bifet and Gavaldà (2007) derived a bound on the false positive and false negative rates by replacing $\delta$ with $\delta' = \delta/|W(t)|$, to account for multiple tests\(^8\). However, the existing analysis is incomplete because it implicitly assumes that the window size $|W(t)|$ is deterministic.

In this paper, we use $\delta$ instead of $\delta'$ because it clarifies the analysis. The value of $\delta$ is in accordance with the possible multiplicity. With this aim, we introduce the notion of the window set and split and then introduce a confidence bound that holds with high probability.

**Definition 4** (Window set) The window set $W$ is the set of all the segments, which are the candidates of the current window $W$.

$$W = \{W' : W' = \{t', t' + 1, \ldots, t''\}, 1 \leq t' < t'' < T\}. \quad (2)$$

**Definition 5** (Split and estimator in a window) Letting $W' = \{t', t' + 1, \ldots, t''\}$, a split $W_1, W_2$ of a window $W' \in W$ is defined as two disjoint subsets of $W'$ such that $W_1 = \{t', t' + 1, \ldots, m\}$ and $W_2 = \{m + 1, m + 2, \ldots, t''\}$ for some $m \in W' \setminus \{t', t''\}$. For a window $W' \in W$,

$$\mu_{W'} = \frac{1}{|W'|} \sum_{s \in W'} \mu_s, \quad (3)$$

and its empirical estimator is

$$\hat{\mu}_{W'} = \frac{1}{|W'|} \sum_{s \in W'} x_s. \quad (4)$$

**Proposition 1** (Uniform Hoeffding bound for the window set) Let $p > 0$ be arbitrary. With probability $1 - 2/T^p$, we have the following:

$$\forall W' \in W, \quad |\mu_{W'} - \hat{\mu}_{W'}| \leq \sqrt{\frac{\log(T^2 + p)}{2|W'|}}. \quad (5)$$

Proposition 1 is derived using the Hoeffding inequality (Lemma 26, in Appendix C) over all possible $|W| \leq T^2$ windows. The bound above holds regardless of the randomness of the current window $W = W(t)$. Moreover, it holds for not only $W$ but also for any split $W_1 \cup W_2 = W$. We typically set $p = 1$ because this insures a uniform bound with probability $1 - O(1/T)$, and an event with probability $O(1/T)$ is usually negligible. Our results are based on this bound.

3.2.2 Bound for the total error of the mean estimator

In the following, we provide a bound on the following total error of the estimator $\hat{\mu}_W$:

**Definition 6** (Total error) The total error of estimator $\hat{\mu}_W$ is defined as follows:

$$\text{Err}(T) = \sum_{t=1}^{T} |\hat{\mu}_{W(t)} - \mu_t|. \quad (6)$$

---

8. There are $|W(t)| - 1$ ways to split $W(t) = W_1 \cup W_2$. 

We first derive the error bound of ADWIN for a stationary stream, which directly follows from the false positive rate (Eq. (5)).

**Theorem 7** (ADWIN in stationary environments) Let the stream be stationary. Then, for ADWIN with $\delta = 1/T^3$, we have the following:

$$E[\text{Err}(T)] \leq \tilde{O}(\sqrt{T}).$$

(7)

**Proof** Equation (5) with $p = 1$, together with Definition 2 implies that no shrink occurs with probability $1 - 2/T$. Therefore,

$$E[\text{Err}(T)] \leq T \times \frac{2}{T} + \sum_{t=1}^{T} \sqrt{\frac{\log(T^3)}{2t}}$$

(8)

$$\leq 1 + \sum_{t=1}^{T} \sqrt{\frac{\log(T^3)}{2t}}$$

(9)

$$\leq \sqrt{6T \log(T)} + 1,$$

(by $\sum_{t=1}^{T} (1/\sqrt{t}) \leq 2\sqrt{T}$)

(10)

which completes the proof.

The $\tilde{O}(\sqrt{T})$ error is the optimal rate because we can only identify the true value of $\mu_W$ up to $O(\sqrt{1/|W|})$. The error per time step $|\mu_t - \mu_W|$ is at least $\Omega(1/\sqrt{|W|}) = \Omega(1/\sqrt{t})$, and the total error is $\Omega(\sum_t 1/\sqrt{t}) = \Omega(\sqrt{T})$.

Theorem 7 states that ADWIN can learn from a stationary environment without any unnecessary shrinking. This statement contrasts with such methods as periodic resetting or fixed-size windowing algorithms which discard their entire memory after a fixed period. Having derived the learnability of ADWIN for a stationary stream, we are now interested in the property of ADWIN in the face of a nonstationary stream.

### 3.3 Analysis of ADWIN for abrupt changes

This section derives an error bound of ADWIN in the face of an abrupt change. As (informally) discussed in Bifet and Gavaldà (2007), ADWIN is able to detect abrupt changes if the changes are infrequent and gaps are detectable. Our results here are even more robust. Somewhat surprisingly, the bound in Theorem 8 does not depend on the detectability of the change. No matter how large or small the changes are and in which interval the changes occur, the algorithm’s performance is bounded in terms of the number of changes.

**Theorem 8** (Error bound of ADWIN under abrupt changes) Let the environment be abrupt with $M$ changepoints. The total error of ADWIN with $\delta = 1/T^3$ is bounded as

$$E[\text{Err}(T)] = \tilde{O}(\sqrt{MT}).$$

(11)

**Proof Sketch** Let $c(t)$ be the number of time steps after the last changepoint. First, we demonstrate that $|\mu_W - \mu_t| = \tilde{O}(1/\sqrt{c(t)})$ because ADWIN otherwise would shrink
the window further (i.e., the event $B$ in the proof). Second, the window size $|W|$ is also bounded below. Given a sufficiently high confidence parameter, $|W| = O(c(t))$ and thus $|\mu_W - \hat{\mu}_W| = \tilde{O}(1/\sqrt{c(t)})$. Combining these two yields the bound of

$$ |\hat{\mu}_W - \mu_t| \leq |\mu_W - \mu_t| + |\mu_W - \hat{\mu}_W| = \tilde{O}(1/\sqrt{c(t)}). \tag{12} $$

Using the Cauchy-Schwarz inequality yields the bound of Eq. (11).

The formal proof is found in Appendix D.1.

**Remark 9** Theorem 8 implies the optimality of ADWIN under abrupt drift. To observe this, assume that a changepoint exists every $T/M$ time steps. As discussed in Theorem 7, the optimal rate of error for each interval between changepoints is $\Theta(\sqrt{T/M})$, and the total error should be $\Theta(M \times \sqrt{T/M}) = \Theta(\sqrt{MT})$, which matches Theorem 8.

### 3.4 Analysis of ADWIN for gradual changes

In this section, we analyze ADWIN for a gradually changing stream, where the mean $\mu_t$ changes slowly with constant $b$ (Definition 1). We consider the error for $b = T^{-d}$ by following the framework of Wei and Srivastava (2018).

**Theorem 10** (Error bound of ADWIN under gradual changes) Let the stream be gradually changing with constant $b$ and $b \leq T^{-d}$ for some $d \in (0, 3/2)$. Then, the performance of ADWIN with $\delta \leq 1/T^3$ is bounded as follow:

$$ \mathbb{E}[\text{Err}(T)] = \tilde{O}(T^{1-d/3}). \tag{13} $$

**Proof Sketch** We establish two lemmas in the appendix. Lemma 27 states that the drift $|\mu_s - \mu_t|$ for any two time steps $s, t$ in the current window $W(t)$ is bounded by $\tilde{O}(N + \sqrt{T/N})$ for any $N \geq |W(t)|$. Moreover, Lemma 28 states that the window is likely to grow until $|W| = O(b^{-2/3})$. Combining these two lemmas yields $|\mu_s - \mu_t| = O(b^{1/3}) = O(T^{-d/3})$. The formal proof is found in Appendix D.4. Theorem 10 states that, if the change is slow compared with the current scale of interest $T$ then the error per time step $\text{Err}(T)/T$ approaches zero.

We provide the bound of total error for abruptly changing streams (Section 3.3) and gradually changing streams (Section 3.4), concluding this section. In subsequent sections, we consider the idea of combining ADWIN with the MAB setting where multiple streams are involved, and only a selected subset of streams are observable.

### 4. Multiple-play Multi-armed Bandits

So far, we considered learning from the a single stream of data, from which every value could be observed. From this section, we study the MP-MAB problem (Anantharam et al., 1987; Komiyama et al., 2015). We first formalize the considered problem setting.

For $K$ arms, at each time step $t = 1, \ldots, T$, the forecaster selects a subset of arms $I(t) \subset [K]$ such that $|I(t)| = L$ and receives a reward $X(t) = (x_i(t))_{i \in I(t)}$ for the selected arms $I(t)$. The well-known $K$-armed bandit problem (Robbins, 1952) is a special case of
Algorithm 2 Multiple-play Thompson sampling (MP-TS)

Require: Set of arms $[K]$.
1: Initialize: $W(t) = \emptyset$
2: for $t = 1, \ldots, T$ do
3:     for $i = 1, \ldots, K$ do
4:         $S_i(t) = \sum_{t \in W(t)} 1[i \in I(t)] x_i(t)$, $N_i(t) = \sum_{t \in W(t)} 1[i \in I(t)]$
5:         $\theta_i(t) \sim \text{Beta}(S_i(t) + 1, N_i(t) - S_i(t) + 1)$
6:     end for
7:     Play arms $I(t) := \arg \max_{I \subseteq [K], |I| = L} \sum_{i \in I} \theta_i(t)$. \hfill ▷ Top-$L$ arms of $\theta_i(t)$
8:     Receive reward $X(t)$.
9:     Update window $W(t+1) = W(t) \cup \{t\}$.
10: end for

Algorithm 3 Multiple-play Kullback–Leibler UCB (MP-KL-UCB).

Require: Set of arms $[K]$.
1: Initialize: $W(t) = \emptyset$
2: for $t = 1, \ldots, T$ do
3:     for $i = 1, \ldots, K$ do
4:         $S_i(t) = \sum_{t \in W(t)} 1[i \in I(t)] x_i(t)$, $N_i(t) = \sum_{t \in W(t)} 1[i \in I(t)]$
5:         $\mu_{i,W}(t) = S_i(t)/N_i(t)$ where $0/0 = +\infty$
6:         $U_i(t) = \max\{q \in [0,1] : N_i(t) d_{KL}(\mu_i, q) \leq \log(t/N_i(t))\}$ \hfill ▷ KL-UCB index
7:     end for
8:     Play arms $I(t) := \arg \max_{I \subseteq [K], |I| = L} \sum_{i \in I} U_i(t)$. \hfill ▷ Top-$L$ arms of $U_i(t)$
9:     Receive reward $X(t)$.
10: Update window $W(t+1) = W(t) \cup \{t\}$.
11: end for
this problem with $L = 1$ (single-play). We also use the term “environment” to describe the $K$ streams that generate rewards $x_i(t) \in [0, 1]$. Let $(\mu_i, t)_{t=[T]}$ be the parameters of the $i$-th stream. The stationary (resp. abruptly changing and gradually changing) environment is defined as a set of $K$ stationary (resp. abruptly changing and gradually changing) streams, as in Definition 1. Under the global change assumption, for the abruptly changing environment, the changepoints $T_c$ are defined as the union of all changepoints of the streams. The gradually changing environment is defined as a set of $K$ gradually changing streams with common change speed $b \in (0, 1)$.

The goal of the forecaster is to maximize the sum of the rewards of the selected arms using a suitable algorithm. The regret is defined as the difference between the expected reward of the top-$L$ arms and the expected reward of the arms chosen by the algorithm:

$$\text{reg}(t) = \max_{I \subseteq [K]: |I|=L} \sum_{i \in I} \mu_i, t - \sum_{i \in I(t)} \mu_i, t, \quad \text{Reg}(T) = \sum_{t=1}^{T} \text{reg}(t). \quad (14)$$

In the $K$-armed bandit problem, TS and UCB are widely known algorithms that balance exploration and exploitation. Among several variants of UCB, the one that uses the Kullback-Leibler (KL) divergence, called KL-UCB (Garivier and Cappé, 2011), exhibits state-of-the-art performance in stationary environments. Although TS and KL-UCB are primarily studied in single-play bandit problems, these two algorithms can be naturally extended to the more general case of $L > 1$. Following Komiyama et al. (2015), we call them MP-TS (Algorithm 2) and MP-KL-UCB (Algorithm 3). Here, $1[A] = 1$ if $A$ or 0 otherwise. These algorithms are designed to deal with stationary environments. The regret of MP-KL-UCB and MP-TS is bounded as follows:

$$E[\text{Reg}(T)] = O\left(\frac{K \log T}{\Delta_{\text{min}}}\right), \quad (15)$$

where $\Delta_{\text{min}} > 0$ is the distribution-dependent gap defined later in Definition 12. The logarithmic bound and its dependence on $\Delta_{\text{min}}$ is optimal (Anantharam et al., 1987; Garivier and Cappé, 2011; Komiyama et al., 2015).

Under a changing environment, the performance of these algorithms is no longer guaranteed. The following section extends the MP-MAB framework for the cases of abruptly changing and gradually changing environments by combining the adaptive window technology with the bandit algorithms.

5. Analysis of MP-MAB for Globally Nonstationary Environments

This section proposes the combination of adaptive windowing and bandit algorithms.

5.1 Adaptive resetting bandit algorithm

The ADaptive Resetting Bandit (ADR-bandit; Algorithm 4) combines adaptive windows with an MP-MAB algorithm. In ADR-bandit, an adaptive window change detector is associated with each arm. If one of the detectors detects a split $W = W_1 \cup W_2$, it resets the entire window, both $W_1$ and $W_2$. This is different from ADWIN (Section 3), which leaves $W_2$. A naive application of ADWIN is challenging to analyze because $W_2$ can include some...
Algorithm 4 ADR-bandit

Require: Set of arms $[K]$, confidence level $\delta$, base bandit algorithm
1: Initialize base bandit algorithm.
2: for $t = 1, 2, \ldots, T$ do
3:  $(I(t), X(t)) = \text{BASE-BANDIT}(W(t)) \triangleright$ Do one step of the base bandit algorithm
4:  if there exists $W(t + 1) = W_1 \cup W_2$ such that $|\hat{\mu}_{i,W_1}(t) - \hat{\mu}_{i,W_2}(t)| \geq \epsilon_{\text{cut}}$ then
5:    Re-initialize the base bandit algorithm.
6:  Update the window $W(t + 1) = \emptyset$ of the base bandit algorithm.
7:  end if
8: end for

amount of observations before the changepoint\(^9\). Although keeping $W_2$ intuitively works better, we empirically confirm that this modification does not degrade the performance much; see our experiments in the appendix (Section B).

We explain the steps of our algorithm with MP-TS as a base bandit algorithm. Before the first time step, the base bandit algorithm (Line 1 of Algorithm 2) is initialized. At each time step $t = 1, 2, \ldots$, Line 3 of Algorithm 4 runs an iteration of the base bandit (Line 2 of Algorithm 2) and obtains the set of arms $I(t)$ and corresponding rewards $X(t)$. Afterward, Line 4 of Algorithm 4 checks whether a change occurred for each arm. If at least one change is detected, it resets window $W(t + 1)$ and initializes the base bandit algorithm.

The following remark characterizes the high-level idea behind the algorithm.

Remark 11 (Nonintervening) Assume that time step $s$ is a detection time and ADR-bandit has not detected any change for the next $S$ time steps $s + 1, s + 2, \ldots, s + S$. Then, the regret during these $S$ time steps is the same as the regret we would obtain if we had run the base bandit algorithm for these $S$ time steps.

Remark 11 states that the introduction of change point detectors does not degrade the performance of the base bandit algorithm while the window is growing. Because $S$ above is a random variable, it is still nontrivial to convert the guarantee on the base bandit into the guarantee of the ADR-bandit.

In Section 5.2, we characterize the properties of the base bandit algorithm required in this analysis.

5.2 Properties for base bandit algorithms

We first characterize properties of the base bandit algorithm under the stationary environment. The MP-KL-UCB and MP-TS have a logarithmic regret bound (Equation (15)). See the following definition:

Definition 12 (Distribution-Dependent (DD) regret) For a stationary environment (i.e., $\mu_{i,t} = \mu_i$), for ease of discussion, we assume $\mu_1 > \mu_2 > \cdots > \mu_K$:

$$\Delta_{\text{min}} = \mu_L - \mu_{L+1}. \quad (16)$$

9. We can cope with this problem in the case of single-stream ADWIN (Section 3.3) since $\mu_W$ eventually converges to $\mu_*$ as $W$ grows. However, under bandit feedback, our algorithms focus on optimal arms and draws suboptimal arms less often. Thus, bounding the error of the estimators for suboptimal arms is hard if the current window $W$ still includes observations before the changepoint.
A base bandit algorithm has a logarithmic DD regret if a universal constant $C^\text{st}$ exists such that
\[
\mathbb{E}[\text{Reg}_{\text{base}}(T)] \leq C^\text{st} \frac{K \log T}{\Delta_{\min}} + o(\log T),
\]
where, $\text{Reg}_{\text{base}}(T)$ denotes the regret when we run the base bandit alone.

The bound in Definition 12 is referred to as the DD bound in the sense that it depends on $\Delta_{\min}$. The inverse of $\Delta_{\min}$ defines the hardness of the instance, and a good algorithm has a provable regret bound in terms of this quantity.

While the DD regret is the most widely studied measure of the performance of stationary bandits, large parts of the literature on the nonstationary bandit are interested in another performance measure called the distribution-independent bound. This bound does not depend on the value of $\mu_{i,t}$ because we cannot define $\Delta_{\min}$ under a nonstationary environment.\(^\text{10}\) In the following definition, we define the performance of the base bandit algorithm under a changing environment.

**Definition 13** (Drift-Tolerant (DT) regret) For a stationary environment that is abruptly or gradually changing, we let
\[
\epsilon(t) = \sum_{s < t} \max_i |\mu_{i,s} - \mu_{i,s+1}|,
\]
which is the maximum drift of the arms by time step $t$. A base bandit algorithm has a sublinear DT regret if a factor $C^\text{dt} = \tilde{O}(1)$ exists such that
\[
\mathbb{E}[\text{Reg}_{\text{base}}(T)] \leq C^\text{dt} \left( \sqrt{LK^2T} + L\epsilon(T) \right).
\]

**Remark 14** Definition 13 is a generalization of the distribution-independent regret in the literature of stationary bandits that allows a drift component $\epsilon(t) > 0$.\(^\text{11}\)

We also introduce a slightly stronger bound.

**Definition 15** (Strong DT regret) Let the environment be stationary, abruptly changing, or gradually changing. A base bandit algorithm has a sublinear strong DT regret if $C^\text{dt} = \tilde{O}(1)$ exists such that, with probability at least $1 - 4K/T^2$,
\[
\text{Reg}_{\text{base}}(S) \leq C^\text{dt} \left( \sqrt{LKS} + L\epsilon(S) \right)
\]
holds for all $S \in [T]$.

Definition 15 is a sufficient condition of Definition 13. We use the DT regret in the analysis of abrupt changes, whereas we use the strong DT regret in the analysis of gradual changes.

---

\(^{10}\) In the gradually changing environment, the gap between two arms can be zero in some time steps.

\(^{11}\) It is well-known that the distribution-independent regret of $\tilde{\Theta}(\sqrt{KT})$ is optimal in the stationary $K$-armed bandit (Audibert and Bubeck 2009). Regarding the algorithms where $C^\text{dt} = O(1)$, see, e.g., Audibert and Bubeck (2009); Degenne and Perchet (2016); Zimmert and Seldin (2019).
Remark 16 We hypothesize that CUCB (Chen et al., 2016), MP-KL-UCB (Garivier and Cappé, 2011), and MP-TS (Komiyama et al., 2015) have a DT regret because (i) these algorithms have a distribution-independent regret of $\tilde{O}(\sqrt{KT})$ in the case of $L = 1$ and (ii) the mean of each arm is allowed to drift at most $\epsilon(t)$. In Section 5.3, we introduce a base bandit algorithm that has a strong DT regret (Elimination-UCB, Algorithm 5).

In addition to the performance of the base bandit algorithm, we require the base algorithm to be “greedy enough” (i.e., it keeps drawing optimal arms).

Definition 17 (Monitoring consistency) A base bandit algorithm has the monitoring consistency with $D$ if

$$
\bigcup_{i \in [K]} \cap_{t \in [T-D+1]} \bigcup_{t' \in \{t, t+1, t+2, \ldots, t+D-1\}} \{i \in I(t')\}
$$

holds when we run it on an MP-MAB environment.

Intuitively, Definition 17 states that an arm exists that is drawn periodically at least once in every $D$ time steps. The MAB algorithm draws the best arms more frequently than the others; therefore, it is reasonable to assume this property. With this property, we expect at least one of the $K$ arms are useful in detecting changes.

Remark 18 (Do state-of-the-art bandit algorithms have monitoring consistency?) In general, a bandit algorithm samples good arms more often than in uniform sampling, and it is natural to assume monitoring consistency. However, it is highly nontrivial to provide a specific value of $D$ for existing algorithms, such as MP-TS and MP-KL-UCB.

5.3 Elimination-UCB algorithm

Although MP-TS or MP-KL-UCB are optimal in a stationary environment (Komiyama et al., 2015), it is highly nontrivial to derive the DT regret and monitoring consistency of these algorithms (Remark 18); thus, it is challenging to derive a regret bound. Instead, we introduce the Elimination-UCB (Algorithm 5), an algorithm that has all the properties required for regret analysis. The Elimination-UCB is a combination of UCB and the elimination algorithm (Even-Dar et al., 2006) that has been primarily considered in the best arm identification. It selects an arm using an elimination subroutine that keeps a candidate pool of the best arms to guarantee monitoring consistency. Afterward, it selects the remaining $(L - 1)$ arms in terms of the UCB bound. Regarding the elimination of the suboptimal arms, we use the samples exclusively chosen by the elimination subroutine. The samples, means, and confidence bounds are given as follows:

$$
N_i(t) = \sum_{s \in W(t)} 1[i \in I(s)], \quad N_i^E(t) = \sum_{s \in W(t)} 1[i \in I(s), i = (t \mod^+ K)],
$$

$$
S_i(t) = \sum_{s \in W(t)} 1[i \in I(s)]x_i(s), \quad S_i^E(t) = \sum_{s \in W(t)} 1[i \in I(s), i = (t \mod^+ K)]x_i(s),
$$

12. This is for the sake of deriving Lemma 34. We limit the samples for the elimination subroutine to fix the ratio of draws so that all arms evolve in parallel.
\[ \hat{\mu}_i(t) = \frac{S_i(t)}{N_i(t)}, \quad \hat{\mu}_E^i(t) = \frac{S_i^E(t)}{N_i^E(t)}, \]
\[ B_i(t) = \sqrt{\frac{\log(T^4)}{2N_i(t)}}, \quad B_E^i(t) = \sqrt{\frac{\log(T^4)}{2N_i^E(t)}}, \]
\[ U_i(t) = \hat{\mu}_i, W(t)) + B_i(t), \quad U_E^i(t) = \hat{\mu}_E^i, W(t)) + B_E^i(t), \]
\[ \text{where } 0/0 = +\infty. \]

**Algorithm 5 Elimination-UCB**

**Require:** Set of arms \([K]\), time horizon \(T\)

1. **Initialize:** \(\hat{I}^*(t) = [K]\). \(\triangleright \hat{I}^*(t)\): Candidate of the best (top-1) arm.
2. **for** \(t = 1, 2, \ldots, T\) **do**
3. \(k = (t \mod + K)\) \(\triangleright t \mod + K := (t \mod K) + 1\) so that \(k \in [K]\)
4. **if** \(k \in \hat{I}^*(t)\) **then**
   \(\triangleright \text{Draw arm } k \text{ and Estimated top-}(L-1)\) arms
   5. \(I(t) = \{k\} + \arg \max_{I':|I'|=L-1,k \notin I'} \sum_{i \in I'} U_i\)
   6. **else**
   \(\triangleright \text{Estimated top-}L\) arms
   7. \(I(t) = \arg \max_{I':|I'|=L} \sum_{i \in I'} U_i\)
8. **end if**
9. **if** \(|\hat{I}^*(t)| \geq 2\) **then** \(\triangleright \text{Elimination until } |\hat{I}^*(t)| = 1\)
   10. **if** There exists \(i\) such that \(\hat{\mu}_E^i(t) - B_E^i(t) > \hat{\mu}_E^k(t) + B_E^k(t)\) **then**
    11. \(\hat{I}^*(t) = \hat{I}^*(t) \setminus \{k\}. \triangleright \text{Eliminate arm } k \text{ if there is clearly better arm } i.\)
12. **end if**
13. **end if**
14. **Receive reward** \(X(t)\) and update window \(W(t+1) = W(t) \cup \{t\}\).
15. **end for**

In the following, we demonstrate Elimination-UCB (Algorithm 5) has (i) a logarithmic distribution-dependent (DD) regret bound, (ii) a sublinear drift-tolerant (DT) regret bound, and that (iii) Elimination-UCB has monitoring consistency.

**Lemma 19** The selection sequence \(\{I(t)\}_{t \in [T]}\) of ADR with Elimination-UCB (Algorithm 5) has (A) monitoring consistency (Definition 17) for any environment. Moreover, (B) Elimination has a logarithmic DD regret (Definition 12) and (C) has sublinear strong DT regret (Definition 15).

The proof of Lemma 19 is in Appendix E.2.

**5.4 Regret bounds of ADR-bandit algorithms**

Having defined the ADR-bandit algorithm and the related properties, we state the main results regarding the performance of these algorithms in stationary abruptly or gradually changing environments.

**Theorem 20** (Regret bound of ADR-bandit, stationary case) *Let the environment be stationary. Let the base bandit algorithm has a logarithmic DD regret. Then, the regret of ADR-bandit with \(\delta = 1/T^3\) is bounded as follows:*
\[ \mathbb{E}[\text{Reg}(T)] \leq C_{\text{st}} \frac{K \log T}{\Delta_{\text{min}}} + o(\log T). \]
Theorem 20 states that change-point detectors do not deteriorate the performance of the base bandit algorithm. The proof directly follows from the following facts. First, adaptive windows do not make a false reset with at least a probability of $1 - O(K/T)$. Second, the nonintervening property (Remark 11) ensures that the regret of the ADR-bandit algorithm is equal to the base bandit algorithm when no reset occurs. For completeness, we provide the proof of Theorem 20 in Appendix E.1 This result has a striking difference from most existing nonstationary bandit algorithms, which have the cost of a forced exploration of $\Omega(\sqrt{T})$. In the following sections, we derive the regret bound in abruptly and gradually changing environments.

We next derive a regret bound for an abruptly changing environment. We first define the global changes that we assume on the nature of the streams (Definition 21) and then state the regret bound.

**Definition 21 (Globally abrupt changes)** Let $M$ be the number of change points and $\{T_1, T_2, \ldots, T_M\} = \{t : \exists i \in [K] \mu_{i,t} \neq \mu_{i,t+1}\}$ be the set of changepoints. Moreover, $(T_0, T_{M+1}) = (0, T)$. For the $m$-th changepoint, let $\epsilon_m = \min_i |\mu_{i,t} - \mu_{i,t+1}|$. We let $U(\epsilon) = (\log(T^3))/(2\epsilon^2)$. The $m$-th changepoint is a global change if $T_m - T_{m-1}, T_{m+1} - T_m \geq 32DU(\epsilon_m)$. We also define the following

$$C^{\text{ab}} = \max_{i,j \in [K], t \in T_c} \frac{|\mu_{j,t} - \mu_{j,t+1}|}{|\mu_{i,t} - \mu_{i,t+1}|}$$

and regard $C^{\text{ab}} > 0$ as a (finite) constant.

Intuitively, $C^{\text{ab}}$ is the maximum ratio of changes among arms. Assuming it is a constant indicates that all other arms change proportionally to the change of an arm. The assumption of $T_{m+1} - T_m \geq 32DU(\epsilon_m)$ corresponds to the detectability of the changes, which is essentially the same as those in the literature of active nonstationary bandit algorithms, such as Cao et al. (2019) and Besson et al. (2020).

**Theorem 22 (Regret bound of ADR-bandit under abrupt changes)** Let the base bandit algorithm have a sublinear DT regret (Definition 13) and have monitoring consistency with parameter $D$ (Definition 17). Let the environment have $M$ globally abrupt changes (Definition 21). Then, the regret of the ADR-bandit with $\delta = 1/T^3$ is bounded as follows:

$$\mathbb{E}[\text{Reg}(T)] = \tilde{O}(\sqrt{MLKT}).$$

**Proof Sketch** The proof sketch is as follows.

- With a high probability, the algorithm resets itself within $16DU(\Delta_m)$ time steps after $m$-th changepoint.

- By the nonintervening property and sublinear DT regret of the base algorithm, the expected regret between $m$-th and $(m+1)$-th changepoint is $\sqrt{2LKM(T_{m+1} - T_m)}$.

- The desired bound is yielded by summing the regret over all change points and using the Cauchy–Schwarz inequality and $\sum_{m}(T_m - T_{m-1}) \leq T$.  

19
The formal proof is found in Appendix E.3.

Theorem 22 implies that the ADR-bandit learns the environment when the changes are infrequent (i.e., $M = o(T)$) and the changes are detectable. In the single-play case (i.e., $L = 1$), this bound matches many existing active algorithms, such as those in Cao et al. (2019); Besson et al. (2020).

We next analyze the performance of ADR-bandit with a gradually changing environment. We employ a technique similar to that in Section 3.4 for single-stream ADWIN. We first describe the global changes in a gradually changing environment and then state a regret bound.

**Assumption 23** (Globally gradual changes) A set of $K$ streams has globally gradual changes if a constant $C_{gr} \in (0, 1]$ exists such that

$$|\mu_{i,t} - \mu_{i,s}| \geq C_{gr}|\mu_{j,t} - \mu_{j,s}|$$

(30)

holds for any two arms $i, j \in [K]$ and any two time steps $t, s \in [T]$.

Intuitively, Assumption 23 states that all arms are drifting in a coordinated manner. The drift on the mean of an arm is proportional to the drift on the means of other arms.

**Theorem 24** (Regret bound of the ADR-bandit under gradual changes) Let the base bandit algorithm have a sublinear strong DT regret (Definition 15) and have monitoring consistency (Definition 17). Let the environment be gradual with change speed $b$ and $d$ be such that $b = T^{-d}$. Let the changes be global (Assumption 23). Then the regret of ADR-bandit algorithm with $\delta = 1/T^3$ is bounded as:

$$\mathbb{E}[\text{Reg}(T)] = \tilde{O}\left(\sqrt{LKT^{1-d/3}}\right).$$

(31)

**Proof Sketch** The proof sketch is as follows.

- With a high probability, $|\mu_t - \mu_W| \leq 2bDN + O\left(\sqrt{\log T/N}\right)$ always holds for any $N \geq |W(t)|$ (Lemma 39).

- Using Lemma 40, the current window grows until $|W| = O(b^{-2/3})$ with a high probability, which implies that the number of resets (= detection times) $M_d$ is at most $T/b^{2/3} = T^{1-(2d)/3}$.

- Using the lemmas above, the nonintervening property and the DT regret, the regret is bounded by

$$\tilde{O}(\sqrt{MT}) + \sum_t |\mu_t - \mu_W(t)| = \tilde{O}(T^{1-d/3}).$$

(32)

The formal proof is in Appendix E.4.

We require a stronger version of DT regret (Definition 15). Under gradually changing streams, an anytime bound (i.e., a bound that holds for any time step) is required because the detection times vary. Moreover, a higher confidence level is required because we instantiate the base bandit algorithm more than once (once per reset) and employ a property
that holds for all instances. The number of the instances is at most $T$, and a confidence level that is $T$ times larger is sufficient. Theorem 24 states that ADR-bandit learns the environment when the change is sufficiently slow (i.e., $b = o(1)$). Unlike most nonstationary bandit algorithms, our algorithm design does not require prior knowledge of $b$.

**Remark 25** (Optimality of the regret bound) The rate matches the $\tilde{\Theta}(V^{1/3}T^{2/3})$ lower bound of Besbes et al. (2014) up to a logarithmic order. Under a gradually changing environment, the total variation is $V := \sum |\mu_{i,t} - \mu_{i,t+1}| = bT = O(T^{1-d})$; thus, $V^{1/3}T^{2/3} = O(T^{1-d/3})$. Although Besbes et al. (2014) studied a larger class of drifts where only the total variation is bounded, we hypothesize that the $O(T^{1-d/3})$ bound with respect to the gradual drift of speed $b = T^{-d}$ is tight.\(^{13}\) Unlike Wei and Srivastava (2018); Besbes et al. (2014), our guarantee in the nonstationary setting is limited to the case of global changes.

In summary, ADR-bandit adapts to both the abruptly and gradually changing environments without knowing how fast the streams change if the base bandit satisfies the three properties (DD regret, (strongly) DT regret, and monitoring consistency). We explicitly constructed such a base bandit algorithm called the Elimination-UCB (Algorithm 5).

6. Experiments

This section reports the empirical performance of our proposed algorithms in simulations. We release the source code to reproduce our experiments on GitHub\(^ {14}\).

6.1 Environments

The rewards from arm $i$ are drawn from a Bernouilli distribution with parameter $\mu_i$ that we determine for each arm. We define the following synthetic environments:

- **Stationary**: We define a stationary environment (i.e., with no change) with 100 arms where the mean $\mu_i = (101 - i)/100$ never changes over time, i.e., the rewards of arms are evenly distributed between 0 and 1.

- **Gradual**: We define a gradually changing environment with 100 arms with $\mu_{i,t} = \frac{(T-t+1)}{T}\mu_{i,1} + \frac{t-1}{T}(1-\mu_{i,1})$, where $\mu_{i,1} = (101 - i)/100$. In this environment, the mean of each arm is the same as the stationary environment at time $t = 1$, but evolves gradually toward $1 - \mu_{i,1}$ as $t$ increases up to $T$.

- **Abrupt**: We define an abruptly changing environment with 100 arms and $\mu_{i,1} = (101 - i)/100$ at $t = 1$. Unlike the stationary environment, $\mu_{i,t} = 1 - \mu_{i,1}$ between round $t' = T/3$ and round $t'' = 2T/3$. In other words, the rewards of every arms change abruptly two times at $t = T/3, 2T/3$.

- **Abrupt local**: We define an environment with 100 arms in which only a portion of the arms change abruptly, i.e., the global change assumption does not hold. As in the

---

\(^{13}\) The result from Trov`o et al. (2020) has a better rate of $\tilde{O}(T^{1-d})$. However, they require an additional assumption (Assumption 2, therein) that essentially states that the two arms are not too close to distinguish for a long time.

\(^{14}\) https://github.com/edouardfouche/G-NS-MAB
Figure 4: Rewards in the Zozo dataset. We chose ten ads (= arms, y-axis) among 80 ads. Dark blue and light blue indicate the reward 1 and 0, respectively. White indicates that the corresponding ad is not shown at that time step (no feedback). The ratio of reward 1 to reward 0 is between 0.5% to 3%. We only show the first two days, but the dataset consists of one week of data.

The stationary, abrupt, and gradual settings are in line with the assumptions from our algorithms (ADR-E-UCB and ADR-TS) as the changes are global. The abrupt local setting violates these assumptions, because only a subset of the arms change.

In addition, we consider the following real-world scenarios:

- **Bioliq**: This dataset was released in (Fouché et al., 2019). The rewards are generated from a stream of measurements between 20 sensors in power plant for a duration of 1 week. The authors considered the task of monitoring high-correlation between those sensors and wanted to use bandit algorithms for efficient monitoring. Each pair of sensors is seen as an arm (i.e., there are $20 \times 19/2 = 190$ arms) and the reward is 1 in case the correlation is greater than some threshold over the last 1000 measurements, otherwise 0. Figure 1 is the reward distribution for this environment, as we can see, arms tend to change periodically together.

- **Zozo** (Saito et al., 2020) is a real-world environment where the rewards are generated from an ad recommender on an e-commerce website. We use the data generated by the uniform recommender from the duration of a week and adopted an unbiased offline
Globally Nonstationary Bandits

evaluator (Li et al., 2011) for our experiment\textsuperscript{15}. There are 80 different ads. Due to the sparseness of the rewards, we picked 10 arms. Since only a few of such ads are presented at each round to each user, we concatenated together all the ads presented within a second and assigned a reward of 1 to ads on which at least one user clicked. We assign a reward of 0 to ads that were presented, but no user clicked on them. We skip the round whenever the selected was not presented at all. Figure 4 shows the reward distribution of the dataset.

We considered the single-play setting (i.e. $L = 1$) and report results in the multiple-play setting ($L > 1$) in the appendix (Section B). All the results are averaged over 100 runs. Note that these datasets match with our motivational examples (Example 1 and 2).

6.2 Bandit algorithms

We compared the following bandit algorithms:

- **ADR-E-UCB** is the ADR-bandit algorithm (Algorithm 4) with Elimination-UCB (Algorithm 5) as base bandit algorithm. As discussed in Section 5.3, this algorithm has regret bounds in stationary, globally abrupt and globally gradual environments.

- **ADR-TS** is the ADR-bandit algorithm (Algorithm 4) with TS (Algorithm 2) as base bandit algorithm.

- Passive algorithms: **RExp3** (Besbes et al., 2014) is an adversarial bandit algorithm. Discounted UCB (D-UCB) (Garivier and Moulines, 2008) and Sliding Window TS (SW-TS) (Trovò et al. 2020), **SW-UCB#** (Wei and Srivastava, 2018) (comes in two variants: for Abrupt (A) and Gradual (G) changes) are bandit algorithms with finite memory.

- Active algorithms: **GLR-klUCB** (Besson et al., 2020) is an likelihood-ratio based change detector. It comes in two variants: for local and global changes, and we show the results of the latter. **M-UCB** (Cao et al., 2019) is a bandit algorithm using another change detector as ours. These algorithms involve forced exploration (uniform exploration of size $O(\sqrt{T})$) unlike our algorithms.

6.3 Experimental results

First, Figure 5 compares ADR-bandit and existing algorithms in the stationary, abrupt, and gradual environments. In the stationary environment, ADR-TS has very low regret, which is consistent with the fact that its regret is logarithmic. In the two nonstationary environments with global changes, ADR-TS outperforms the other algorithms. ADR-E-UCB, which comes with regret bounds, does not perform as well as ADR-TS. This is reasonable given that the confidence bound of Elimination-UCB is larger than KL-UCB, which conducts optimal amount of exploration in stationary setting (Garivier et al., 2018) like TS. Second, Figure 6 compares these algorithms in the abrupt local environment. Although the ADR-TS has linear regret as it sometimes do not detect these two local changes, it still outperforms the

\textsuperscript{15} See also https://github.com/st-tech/zr-obp
Figure 5: Regret of algorithms in stationary, abrupt, and gradual environments. Smaller regret ($R_T$: y-axis) indicates better performance.
Globally Nonstationary Bandits

Figure 6: Regret of algorithms in abrupt local environments. Smaller regret ($R_T$: y-axis) indicates better performance. ADR-TS has linear regret after two local changes. However, it still outperforms algorithms that require uniform exploration.

other algorithms. Finally, Figure 7 compares these algorithms against the Bioliq and Zozo datasets. The proposed algorithm consistently outperforms existing nonstationary bandits algorithms.

We provide in Section B additional simulations on the effect of resetting (compared with the shrinking of the original ADWIN), the sensitivity ADR-TS to the value of parameter $\delta$, and results with $L > 1$. We find that ADR-TS is not very sensitive to the choice of parameter $\delta$, as long as it is moderately small.

7. Conclusions

In this paper, we extended stationary bandit algorithms to the nonstationary setting, in a way that does not require forced exploration.

To this aim, we first analyzed the theoretical property of adaptive windows in a single-stream setting (Section 3). After that we combined bandit algorithms (Section 4) with adaptive windowing by introducing ADR-bandit. Unlike existing algorithms, ADR-bandit does not act on the base bandit algorithm unless change points are detected, and thus does not compromise the performance in a stationary environment. Still, we showed its capability to detect global changes (Section 5). We demonstrated the usefulness of our method in simulated and real-world settings via experiments (Section 6).

Although the theoretical properties we proposed in Section 5.2 clarify the analysis, there is still a gap between the theoretically-proven algorithm (i.e., ADR-E-UCB) and a practical algorithm (i.e., ADR-TS). Considering theoretical guarantees of TS and UCB-based algorithms will be an interesting direction for future work.
Figure 7: Performance of algorithms based on two real-world datasets (Bioliq and Zozo). Unlike synthetic environments, we do not have ground truth of $\mu_{i,t}$ and thus we report cumulative reward. Larger cumulative reward (y-axis) indicates better performance. M-UCB tends to have $\gamma \geq 1$ in our setting, which results in a uniform exploration.
References

Robin Allesiardo and Raphaël Féraud. EXP3 with drift detection for the switching bandit problem. Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, (August), 2015.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards. IEEE Transactions on Automatic Control, 32(11):968–976, 1987.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In COLT, 2009.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In Proceedings of IEEE 36th Annual Foundations of Computer Science, pages 322–331, Oct 1995.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2-3):235–256, 2002.

Manuel Baena-Garcia, José del Campo-Ávila, Raúl Fidalgo, Albert Bifet, R Gavalda, and R Morales-Bueno. Early drift detection method. In Fourth international workshop on knowledge discovery from data streams, volume 6, pages 77–86, 2006.

Omar Besbes, Yonatan Gur, and Assaf J. Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. CoRR, abs/1405.3316, 2014.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. CoRR, abs/1902.01575, 2020.

Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In SDM, pages 443–448. SIAM, 2007.

Isvani Inocencio Frías Blanco, José del Campo-Ávila, Gonzalo Ramos-Jiménez, Rafael Morales Bueno, Agustín Alejandro Ortiz Díaz, and Yailé Caballero Mota. Online and Non-Parametric Drift Detection Methods Based on Hoeffding’s Bounds. IEEE Trans. Knowl. Data Eng., 27(3):810–823, 2015.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In AISTATS, volume 89, pages 418–427, 2019.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In NIPS, pages 2249–2257, 2011.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In NIPS, pages 1651–1659, 2016.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In ICML, volume 48, pages 1587–1595, 2016.
Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, December 2006. ISSN 1532-4435.

Edouard Fouché, Junpei Komiyama, and Klemens Böhm. Scaling multi-armed bandit algorithms. In *KDD*, pages 1449–1459. ACM, 2019.

João Gama and Gladys Castillo. Learning with local drift detection. In *ADMA*, volume 4093 of *Lecture Notes in Computer Science*, pages 42–55. Springer, 2006.

João Gama, Pedro Medas, Gladys Castillo, and Pedro Pereira Rodrigues. Learning with drift detection. In *SBIA*, volume 3171 of *Lecture Notes in Computer Science*, pages 286–295. Springer, 2004.

João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, 2014.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, volume 19 of *JMLR Proceedings*, pages 359–376, 2011.

Aurélien Garivier and Eric Moulines. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *CoRR*, abs/0805.3415, 2008.

Aurélien Garivier, Hédi Hadjii, Pierre Menard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*, 2018.

Paulo Mauricio Gonçalves, Silas Garrido Teixeira de Carvalho Santos, Roberto Souto Maior de Barros, and Davi Carnauba de Lima Vieira. A comparative study on concept drift detectors. *Expert Syst. Appl.*, 41(18):8144–8156, 2014.

Yonatan Gur, Assaf J. Zeevi, and Omar Besbes. Stochastic multi-armed-bandit problem with non-stationary rewards. In *NIPS*, pages 199–207, 2014.

Cédric Hartland, Nicolas Baskiotis, Sylvain Gelly, Michèele Sebag, and Olivier Teytaud. Change Point Detection and Meta-Bandits for Online Learning in Dynamic Environments. In *CAP 2007 : 9è Conférence francophone sur l'apprentissage automatique*, pages 237–250, July 2007.

D. V. Hinkley. Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523, 1971. ISSN 00063444.

Levente Kocsis and Csaba Szepesvári. Discounted ucb. In *2nd PASCAL Challenges Workshop*, pages 784–791, 2006.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. In *ICML*, volume 37, pages 1152–1161, 2015.
Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, and Michal Wozniak. Ensemble learning for data stream analysis: A survey. Inf. Fusion, 37:132–156, 2017.

T.L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1):4 – 22, 1985.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In WWW, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In WSDM, pages 297–306. ACM, 2011.

Fang Liu, Joohyun Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In AAAI, pages 3651–3658, 2018.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. IEEE Trans. Knowl. Data Eng., 31(12):2346–2363, 2019.

Odalric-Ambrym Maillard. Boundary crossing for general exponential families. In ALT, volume 76 of Proceedings of Machine Learning Research, pages 151–184, 2017.

Joseph Charles Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change detection. In AISTATS, volume 31, pages 442–450, 2013.

Subhojyoti Mukherjee and Odalric-Ambrym Maillard. Distribution-dependent and time-uniform bounds for piecewise i.i.d bandits. CoRR, abs/1905.13159, 2019.

E. S. Page. Continuous inspection schemes. Biometrika, 41(1/2):100–115, 1954. ISSN 00063444.

Herbert Robbins. Some aspects of the sequential design of experiments. Bull. Amer. Math. Soc., 58(5):527–535, 09 1952.

Gordon J. Ross, Niall M. Adams, Dimitris K. Tasoulis, and David J. Hand. Exponentially weighted moving average charts for detecting concept drift. Pattern Recognit. Lett., 33 (2):191–198, 2012.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Large-scale open dataset, pipeline, and benchmark for bandit algorithms. CoRR, abs/2008.07146, 2020.

Aleka Seliniotaki, George Tzagkarakis, Vassilis Christophides, and Panagiotis Tsakalides. Stream correlation monitoring for uncertainty-aware data processing systems. In IISA, pages 342–347. IEEE, 2014.

Julien Seznec, Pierre Ménard, Alessandro Lazaric, and Michal Valko. A single algorithm for both restless and rested rotting bandits. In AISTATS, volume 108 of Proceedings of Machine Learning Research, pages 3784–3794, 2020.
Vaibhav Srivastava, Paul Reverdy, and Naomi Ehrich Leonard. Surveillance in an abruptly changing world via multiarmed bandits. In *CDC*, pages 692–697. IEEE, 2014.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Francesco Trovò, Marcello Restelli, and Nicola Gatti. Sliding-Window Thompson Sampling for Non-Stationary Settings. *J. Artif. Intell. Res.*, 68:311–364, 2020.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *NIPS*, pages 3972–3980, 2016.

Lai Wei and Vaibhav Srivastava. On abruptly-changing and slowly-varying multiarmed bandit problems. In *ACC*, pages 6291–6296. IEEE, 2018.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *AISTATS 2019*, volume 89, pages 467–475, 2019.
Table 2: List of tested hyperparameters of the algorithms. Bold letters indicate the ones reported in this paper.

| Algorithm(s)          | Hyperparameters                                      |
|-----------------------|------------------------------------------------------|
| ADR-TS and ADR-E-UCB  | $\delta = 0.1, 0.01, 0.001, 0.0001$                 |
| RExp3                 | $\Delta_T = 100, 500, 1000, 5000$                   |
| SW-TS                 | $W = 100, 500, 1000, 5000$                          |
| D-UCB                 | $\gamma = 0.7, 0.8, 0.9, 0.99$                      |
| SW-UCB#-A            | $\nu = 0.1, 0.2$ and $\lambda = 12.3$              |
| SW-UCB#-G            | $\kappa = 0.1, 0.2$ and $\lambda = 4.3$            |
| GLR-klUCB            | $\alpha = \sqrt{kA \log(T)/T}$ and $\delta = 1/\sqrt{T}$ |
| M-UCB                | $w = 1000, 5000$ and $M = 10, 100$                  |

Appendix A. Hyperparameters in Experiments

The hyperparameters of the algorithms are shown in Table 2.

Appendix B. Additional Experiments

This section introduces the adaptive shrinking bandit (ADS-bandit) and reports the results of additional experiments.

B.1 ADS-bandit

Algorithm 6 ADS-bandit

**Require:** Set of arms $[K]$, confidence level $\delta$, base bandit algorithm

1: Initialize the base bandit algorithm.
2: for $t = 1, 2, \ldots, T$ do
3: \quad $(I(t), X(t)) = \text{BASE-BANDIT}(W(t))$ \Comment{One time step of the base bandit algorithm}
4: \quad if there exists a split $W(t + 1) = W_1 \cup W_2$ such that $|\hat{\mu}_{i,W_1} - \hat{\mu}_{i,W_2}| \geq \epsilon_{\text{cut}}^{i}$ then
5: \quad \quad Update the window $W(t + 1) = W_2$ of the base bandit algorithm.
6: \quad end if
7: end for

Recall that ADR-bandit (Algorithm 4) resets the entire memory when it detects a change. While this is required to derive the regret bound (Section 5.4), it is somewhat conservative. ADS-bandit (adaptive shrinking bandit; Algorithm 6) is another bandit algorithm that keeps the right split $W_2$ when it makes a split $W = W_1 \cup W_2$ like ADWIN (Section 3). Although we consider the analysis of ADS-bandit is out of the scopes of this paper, we demonstrate the performance of ADS-bandit in Figure 8 with TS as a base ban-
Figure 8: Comparison of ADR-TS and ADS-TS in synthetic environments. Smaller regret per $L$ ($R_T/L$: y-axis) indicates better performance. The performance gap between ADR-TS and ADS-TS is very small. The effect of hyperparameter $\delta$ in change point detection is not very large, except for very large value of $\delta = 0.1$.

ddit algorithm. As we can see, the performance of ADS-TS is very similar from the one of ADR-TS.

B.2 Results with multiple-play

We extend each algorithm to the multiple-play setting by selecting the top-$L$ arms instead of the single best arm. We let the number of plays $L$ vary with $L = \{1, 2, 5\}$. Figures 9–11 are the results for each setting. Overall, the advantage of ADR-TS is consistent in the multiple-play and single-play settings.
Figure 9: Regret of algorithms in stationary synthetic environments. Smaller regret per $L$ ($R_T/L$: y-axis) indicates better performance.
Figure 10: Regret of algorithms in abrupt synthetic environments. Smaller regret per $L$ ($R_T/L$: y-axis) indicates better performance.
Figure 11: Regret of algorithms in gradual synthetic environments. Smaller regret per $L$ ($R_T/L$: y-axis) indicates better performance.
Appendix C. Lemmas

This section describes the lemmas that are used in the proofs of this paper.

The Hoeffding inequality, which is one of the most well-known concentration inequality, provides a high-probability bound of the sum of bounded independent random variables.

**Lemma 26 (Azuma-Hoeffding inequality)** Let \( x_1, x_2, \ldots, x_n \) be martingale random variables in \([0, 1]\) with their conditional mean \( \mu_m = \mathbb{E}[x_m | x_1, x_2, \ldots, x_{m-1}] \). Let \( \bar{x} = (1/n) \sum_{t=1}^{n} x_t \) and \( \bar{\mu} = (1/n) \sum_{t=1}^{n} \mu_t \). Then,

\[
\Pr \left[ \bar{x} - \bar{\mu} \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right] \leq \delta, \tag{33}
\]

\[
\Pr \left[ \bar{x} - \bar{\mu} \leq -\sqrt{\frac{\log(1/\delta)}{2n}} \right] \leq \delta. \tag{34}
\]

Moreover, taking a union bound over the two inequalities yields

\[
\Pr \left[ |\bar{x} - \bar{\mu}| \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right] \leq 2\delta. \tag{35}
\]

Appendix D. Proofs of ADWIN

We denote \( A, B = A \cap B \) for two events \( A, B \).

**D.1 Proof of Theorem 8**

The overall idea here is as follows. With sufficiently long time after a changepoint, we can expect that ADWIN shrinks the window. However, there might be some time steps left in the current window \( W(t) \) even if a shrink occurs. Still, we can show that

\[
|\hat{\mu}_W(t) - \mu_t| \leq |\mu_W(t) - \mu_t| + |\hat{\mu}_W(t) - \mu_W(t)| \leq \tilde{O} \left( \sqrt{\frac{1}{c(t)}} \right) + \tilde{O} \left( \sqrt{\frac{1}{c(t)}} \right) \tag{36}
\]

where \( c(t) \) be the number of time steps after the changepoint.\(^{16}\) Events \( C \) and \( D \) in the following corresponds to the bounds of \( |\mu_W(t) - \mu_t| \) and \( |\hat{\mu}_W(t) - \mu_W(t)| \), respectively.

**Proof** [Proof of Theorem 8]

By Eq. (5), event

\[
\mathcal{B} = \bigcap_{W' \in W} \left\{ |\mu_{W'} - \hat{\mu}_{W'}| \leq \sqrt{\frac{\log(T^3)}{2|W'|}} \right\} \tag{38}
\]

holds with probability at least \( 1 - 2/T \).

---

\(^{16}\) A formal definition of \( c(t) \) is given in Eq. (39).
For each time step \( t \in [T] \), let
\[
c(t) = t - \max_{s < t, s \in \mathcal{T}_c} s.
\] (39)
Namely, \( c(t) \) is the number of the time steps after the latest changepoint. Moreover, for \( c(t) > 2 \), let
\[
\Delta(c) = 4\sqrt{\log(T^3)}
\] (40)
and \( \Delta(c) = 1 \) for \( c \leq 2 \). Let
\[
\mathcal{C} = \bigcap_t \{|\mu_{W(t)} - \mu_t| < \Delta(c(t))\}.
\] (41)
In the following we show \( \mathcal{C} \) holds under \( \mathcal{B} \). If \( |W(t)| + 1 \leq c(t) \), then all the time steps in \( W(t) \) is after the last change point, which implies \( \mu_{W(t)} = \mu_t \) and thus \( \mathcal{C} \). Otherwise, let
\[
W_2 = \{t - c(t)/2 + 2, t - c(t)/2 + 3, \ldots, t\} \quad \text{(i.e. the last } c(t)/2 - 1 \text{ time steps)}
\] (42)
\[
W_1 = W(t - 1) \setminus W_2
\] (43)
Clearly \( |W_1|, |W_2| \geq c(t)/2 - 1 \) and \( \mu_{W_2} = \mu_t \). ADWIN at time step \( t - 1 \) shrinks the window until
\[
|\hat{\mu}_{W_1} - |\hat{\mu}_{W_2}| \leq \sqrt{\frac{\log(T^3)}{2|W_1|}} + \sqrt{\frac{\log(T^3)}{2|W_2|}}
\] (44)
holds. Let \( n_1, n_2 = |W_1|, |W_2| \).
\[
\sqrt{\frac{\log(T^3)}{2|W_1|}} + \sqrt{\frac{\log(T^3)}{2|W_2|}} \geq |\hat{\mu}_{W_1} - |\hat{\mu}_{W_2}| \quad \text{(by Eq. (44))}
\] (45)
\[
\geq |\mu_{W_1} - |\mu_{W_2}| - \sqrt{\frac{\log(T^3)}{2|W_1|}} - \sqrt{\frac{\log(T^3)}{2|W_2|}} \quad \text{(by } \mathcal{B} \text{)}
\] (46)
\[
= |\mu_{W_1} - |\mu_t| - \sqrt{\frac{\log(T^3)}{2|W_1|}} - \sqrt{\frac{\log(T^3)}{2|W_2|}} \quad \text{(by } |W_2| + 1 \leq c(t) \text{)}
\] (47)
and thus
\[
|\mu_{W(t)} - \mu_t| \leq |\mu_{W_1} - \mu_t| \quad \text{(by Eq. (47))}
\] (48)
\[
\leq 2\sqrt{\frac{\log(T^3)}{2|W_1|}} + 2\sqrt{\frac{\log(T^3)}{2|W_2|}} \quad \text{(by Eq. (47))}
\] (49)
\[
\leq 4\sqrt{\frac{\log(T^3)}{c - 2}} \quad \text{(by } |W_1|, |W_2| \geq (c - 2)/2 \text{)}
\] (50)
In summary, \( \mathcal{B} \) implies \( \mathcal{C} \).
Note that under event $B$, ADWIN never makes a false shrink (i.e., a shrink when $\mu_{W_1} = \mu_{W_2}$). This implies that between two changepoints a shrink that makes $|W(t)| + 1 < c(t)$ occurs at most once, which leads to the fact that event

$$
\mathcal{D} = \bigcap_{n \in \mathbb{N}} \left\{ \sum_t \mathbf{1}\{|W(t)| = n \land |W(t)| + 1 < c(t)\} \leq 2M \right\}.
$$

(51)

By using this, the total error is bounded as

$$
\mathbb{E}[\text{Err}(T)]
\leq T \Pr[\mathcal{B}^c] + \mathbb{E}[\text{Err}(T)\mathbf{1}[\mathcal{B}, \mathcal{C}, \mathcal{D}]] \quad (\mathcal{B} \text{ implies } \mathcal{C} \text{ and } \mathcal{D})
\leq 2 + \mathbb{E}[\text{Err}(T)\mathbf{1}[\mathcal{B}, \mathcal{C}, \mathcal{D}]]
\leq 2 + \mathbb{E}\left[ \sum_t |\mu_t - \hat{\mu}_W| \mathbf{1}[\mathcal{B}, \mathcal{C}, \mathcal{D}] \right]
\leq 2 + \mathbb{E}\left[ \sum_t (|\mu_t - \mu_W| + |\hat{\mu}_W - \mu_W|) \mathbf{1}[\mathcal{B}, \mathcal{C}, \mathcal{D}] \right]
\leq 2 + \sum_t \Delta(c(t)) + \mathbb{E}\left[ \sum_t |\hat{\mu}_W - \mu_W| \mathbf{1}[\mathcal{B}, \mathcal{C}, \mathcal{D}] \right] \quad (\text{by } \mathcal{C})
\leq 2 + \sum_t \Delta(c(t)) + O(M) + \sum_t \sqrt{\frac{\log(T^3)}{2c(t)}}
\leq 2 + \sum_t \Delta(c(t)) + O(M) + \sum_t \sqrt{\frac{\log(T^3)}{2c(t)}} + \mathbb{E}\left[ \sum_n \sum_t \mathbf{1}[|W(t)| = n, |W(t)| + 1 < c(t)] \sqrt{\frac{\log(T^3)}{2n}} \mathbf{1}[\mathcal{D}] \right] \quad (\text{by } \mathcal{B})
\leq O(M) + \sum_t O\left( \sqrt{\frac{\log(T)}{c(t)}} \right) + \mathbb{E}\left[ \sum_n \sum_t \mathbf{1}[|W(t)| = n, |W(t)| + 1 < c(t)] \sqrt{\frac{\log(T^3)}{2n}} \mathbf{1}[\mathcal{D}] \right].
$$

(60)

Here

$$
\sqrt{\frac{\log(T)}{c(t)}} \leq \sqrt{T} \times \sqrt{\sum_t \left( \frac{\log(T)}{c(t)} \right)^2} \quad (\text{by Cauchy-Schwarz inequality})
$$

(61)

$$
= \tilde{O}\left( \sqrt{MT} \right). \quad (c(t) = n \text{ holds at most } M + 1 \text{ times for each } n)
$$

(62)

Another application of the Cauchy-Schwarz inequality, combined with $\mathcal{D}$, yields

$$
\sum_n \sum_t \mathbf{1}[|W(t)| = n, |W(t)| \leq c(t)] \sqrt{\frac{\log(T^4)}{2n}} \mathbf{1}[\mathcal{D}] = \tilde{O}\left( \sqrt{MT} \right),
$$

(63)

which completes the proof.

The following Lemmas 27 and 28 characterize the accuracy of estimator $\hat{\mu}_W$ under gradual drift. These lemmas are used in the proof of Lemma Theorem 10.
\textbf{D.2 Lemma 27}

\textbf{Lemma 27} Let the stream be gradual with its speed of the change \( b \). Let the parameter of ADWIN be \( \delta = 1/T^3 \). Then, with probability at least \( 1 - 2/T \), for all \( N \leq |W| \),

\[ |\mu_s - \mu_s'| \leq 3bN + \tilde{O} \left( \sqrt{\frac{1}{N}} \right) \]  

(64)

for any two time steps \( s, s' \) in the current window \( W = W(t) \), from which it easily follows that

\[ |\mu_t - \mu_W| \leq 3bN + \tilde{O} \left( \sqrt{\frac{1}{N}} \right). \]  

(65)

Lemma 27 is a strong characterization because \( \mu_t - \mu_s \) does not depend on the window size \( |W| \): No matter how long the current window is, \( \mu_t - \mu_s \) and thus \( \mu_t - \mu_W \) is bounded in terms of change speed \( b \): In other words, if \( \mu_t - \mu_W \) is larger, ADWIN shrinks the window.

\textbf{Proof} [Proof of Lemma 27] We first consider the case \( |W| = CN \) for some integer \( C \in \mathbb{N}^+ \). We decompose \( W \) into \( C \) subwindows of equal size \( N \) and let \( W_c \) be the \( c \)-th subwindow for \( c \in [C] \). For \( c \in [C] \setminus \{1\} \), let \( W_{c'} \) be the joint subwindow of \( W \) before \( W_c \). Namely, \( W_{c'} = W_1 \cup W_2 \cup \cdots \cup W_{c-1} \). The fact that the window grows to size \( W \) without a detection implies that each split \( W_{c'} \cup W_c \) satisfies

\[ |\hat{\mu}_{W_{c'}} - \hat{\mu}_{W_c}| \leq 2\sqrt{\frac{\log(T^3)}{N}}. \]  

(66)

Let \( c \in [C] \) be arbitrary. We have

\[ |\hat{\mu}_W - \hat{\mu}_{W_c}| = \left| \frac{1}{C} \hat{\mu}_{W_c} + \frac{C - 1}{C} \hat{\mu}_{W_c} - \hat{\mu}_{W_c} \right| \]  

(67)

\[ \leq |\hat{\mu}_{W,c} - \hat{\mu}_{W_c}| + \frac{2}{C} \sqrt{\frac{\log(T^3)}{N}} \]  

(by Eq.(66))

(68)

\[ \leq |\hat{\mu}_{W,c-1} - \hat{\mu}_{W_c}| + \left( \frac{2}{C - 1} + \frac{2}{C} \right) \sqrt{\frac{\log(T^3)}{N}} \]  

(69)

\[ \quad \ldots \]  

(70)

\[ \leq |\hat{\mu}_{W,c} - \hat{\mu}_{W_c}| + \sum_{c'=c+1}^{C} \frac{2}{C'} \sqrt{\frac{\log(T^3)}{N}} \]  

(71)

\[ \leq \left( 1 + \sum_{c'=c+1}^{C} \frac{2}{C'} \right) \sqrt{\frac{\log(T^3)}{N}} \]  

(by Eq.(66))

(72)

\[ \leq (\log T) \sqrt{\frac{\log(T^3)}{N}}, \]  

(73)
which implies for any \( c, c' \in [C] \) we have

\[
|\hat{\mu}_{W^c} - \hat{\mu}_{W^{c'}}| \leq 2(\log T)\sqrt{\frac{(T^3)}{N}}. \tag{74}
\]

By Eq. (5) we have

\[
|\mu_{W^c} - \hat{\mu}_{W^c}| \leq \sqrt{\frac{(T^3)}{N}}  \\
|\mu_{W^{c'}} - \hat{\mu}_{W^{c'}}| \leq \sqrt{\frac{(T^3)}{N}}. \tag{75}
\]

By the fact that \( \mu_t \) moves \( bN \) within a subwindow of size \( w \), for any \( s \in W^c, s' \in W^{c'} \), we have

\[
|\mu_{W^c} - \mu_s| \leq bN  \\
|\mu_{W^{c'}} - \mu_{s'}| \leq bN. \tag{76}
\]

By using these, we have

\[
|\mu_s - \mu_{s'}| \leq |\mu_{W^c} - \mu_s| + |\mu_{W^{c'}} - \mu_{s'}| + |\mu_{W^c} - \hat{\mu}_{W^c}| + |\mu_{W^{c'}} - \hat{\mu}_{W^{c'}}| + |\hat{\mu}_{W^c} - \hat{\mu}_{W^{c'}}| \tag{77}
\]

\[
\leq 2bN + 2(1 + \log T)\sqrt{\frac{(T^3)}{N}}. \tag{by Eq. (74),(75),(76) (78)}
\]

The general case of \( |W| = CN + n \) for \( n \in \{0, 1, \ldots, N - 1\} \) is easily proven by replacing \( 2bN \) with \( 3bN \) since \( \mu_t \) drift at most \( bN \) in \( n \) time steps.

\[ \blacksquare \]

### D.3 Lemma 28

Lemma 27 characterizes the accuracy of the estimator. However, Lemma 27 only holds when window size \( |W| \geq N \). When ADWIN shrinks the window very frequently, we cannot guarantee the quality of the estimator \( \hat{\mu}_W \). The following Lemma 28 states that this is not the case: With high probability, the current window grows until \( |W| = O(b^{-2/3}) \).

**Lemma 28** (Bound on erroneous shrinking) Let \( C_1 = \tilde{O}(1) \) be a sufficiently large value that is defined in Eq. (92). Let the parameter of ADWIN be \( \delta = 1/T^3 \). Let the drift speed \( b \) be such that \( C_1 b^{-2/3} \leq T \). Let

\[
\mathcal{P}(t) = \bigcup_{W_1, W_2, W(t)=W_1 \cup W_2} \left\{ \left| W_1 \right| \leq C_1 b^{-2/3}, \left| W_2 \right| \leq C_1 b^{-2/3}, |\hat{\mu}_{W_1} - \hat{\mu}_{W_2}| \geq \epsilon \right\}. \tag{79}
\]

Let

\[
\mathcal{P} = \bigcup_{t \in [T]} \mathcal{P}(t). \tag{80}
\]

Then,

\[
\Pr[\mathcal{P}] \leq 2C_1 T^{-1}. \tag{81}
\]
Globally Nonstationary Bandits

**Proof** [Proof of Lemma 28] Let $C_1 = \tilde{O}(1)$ that we define later in Eq. (92). Let

$$W_{C_1} = \{W' \in \mathcal{W} : |W'| \leq C_1 b^{-2/3}\}$$  \hspace{1cm} (82)

be the subset of the window set $\mathcal{W}$ with their sizes at most $C_1 b^{-2/3}$. It is easy to show that $|W_{C_1}| \leq TC_1 b^{-2/3}$. Let $d$ be such that $T - d = b$. Similarly to Eq. (5), by the union bound of the Hoeffding bound over all windows of $W_{C_1}$, with probability at least

$$1 - \frac{2}{T^{2+d}} \times TC_1 b^{-2/3} = 1 - 2C_1 T^{-1/3} b^{1/3} \geq 1 - 2C_1 T^{-1},$$  \hspace{1cm} (83)

we have

$$|\mu_{W_1} - \tilde{\mu}_{W_1}| \leq \sqrt{\frac{\log(T^{2+d})}{2|W_1|}},$$

$$|\mu_{W_2} - \tilde{\mu}_{W_2}| \leq \sqrt{\frac{\log(T^{2+d})}{2|W_2|}}$$  \hspace{1cm} (84)

holds for all $t$ and any split $W_1 \cup W_2 = W(t) : |W_1|, |W_2| \leq C_1 b^{-2/3}$. Let $N = C_1 b^{-2/3}$. By definition of gradual stream,

$$|\mu_{W_1} - \mu_{W_2}| \leq 2bN.$$  \hspace{1cm} (85)

Here,

$$\sqrt{\frac{\log(T^3)}{2|W_1|}} + \sqrt{\frac{\log(T^3)}{2|W_2|}} \leq |\tilde{\mu}_{W_1} - \tilde{\mu}_{W_2}| \quad \text{(by the fact ADWIN detects a change)}$$  \hspace{1cm} (86)

$$\leq |\mu_{W_1} - \mu_{W_2}| + |\tilde{\mu}_{W_1} - \mu_{W_1}| + |\tilde{\mu}_{W_2} - \mu_{W_2}| \quad \text{(by triangular inequality)}$$  \hspace{1cm} (87)

$$\leq 2bN + |\tilde{\mu}_{W_1} - \mu_{W_1}| + |\tilde{\mu}_{W_2} - \mu_{W_2}| \quad \text{(by (85))}$$  \hspace{1cm} (88)

$$\leq 2bN + \sqrt{\frac{\log(T^{2+d})}{2|W_1|}} + \sqrt{\frac{\log(T^{2+d})}{2|W_2|}} \quad \text{(by (84))}$$  \hspace{1cm} (89)

which implies

$$\left(\sqrt{3} - \sqrt{2} + d\right) \left(\sqrt{\frac{\log T}{|W_1|}} + \sqrt{\frac{\log T}{|W_2|}}\right) \leq 2bN,$$  \hspace{1cm} (91)

which does not hold for

$$N = b^{-2/3} \left(\left(\sqrt{3} - \sqrt{2} + d\right) \sqrt{\frac{\log T}{C_1}}\right)^{2/3}, |W_1|, |W_2| \leq N.$$  \hspace{1cm} (92)

In summary, with probability $1 - 2C_1 T^{-1}$, we have $\mathcal{P}^c$. \hfill \blacksquare
D.4 Proof of Theorem 10

Let $d$ be such that $b = T^{-d}$. Let $C_1$ be the value defined in Eq. (92) and $N_1 = C_1 b^{-2/3}$.

By using Lemma 28, we bound the number of shrinks.

**Lemma 29** (Number of shrinks) Let

$$\mathcal{T}_d^{\text{sml}} := \{ t \in \mathcal{T}_d, |W_2(t)| < N_1 \}. \tag{93}$$

Then, under $\mathcal{P}^c$, $|\mathcal{T}_d^{\text{sml}}| \leq T/N_1$ holds.

**Proof** [Proof of Lemma 29] $\mathcal{P}^c$ implies that for any shrink $|W_1| \geq N_1$ or $|W_2| \geq N_1$ holds. A shrink of the latter case is not included in $\mathcal{T}_d^{\text{sml}}$. A shrink of the former case reduces the size of window at least $N_1$, and cannot occur more than $T/N_1$ times. \hfill \blacksquare

**Proof** [Proof of Theorem 10] Let $c^{br}(t) = t - \max_{s<t,s \in \mathcal{T}_d} s$ be the number of time steps after the last detection time. We have

$$\text{Err}(T) \leq T \mathbb{1}[\mathcal{P}] + \sum_{n=1}^{N_1} \sum_{t=1}^{T} |\mu_t - \hat{\mu}_W| 1[c^{br}(t) = n, \mathcal{P}^c] + \sum_{t=1}^{T} |\mu_t - \hat{\mu}_W| 1[c^{br}(t) \geq N_1, \mathcal{P}^c]. \tag{94}$$

By Lemma 28, the expectation of the first term of Eq. (94) is bounded as

$$\mathbb{P}[T \mathbb{1}[\mathcal{P}]] = T \times 2C_1 T^{-1} = \tilde{O}(1). \tag{95}$$

We next bound the second term of Eq. (94). By Eq. (5),

$$\forall t \in [T] \left| \mu_t - \hat{\mu}_W \right| \leq \sqrt{\frac{\log(T^3)}{2(c^{br}(t) - 1)} + b c^{br}(t)} \tag{96}$$

holds with probability $1 - 2/T$. The expectation of the second term is bounded as:

$$\sum_{n=1}^{N_1} \sum_{t=1}^{T} \mathbb{P}[|\mu_t - \hat{\mu}_W| 1[c(t) = n, \mathcal{X}^c]] \tag{97}$$

$$\leq (|\mathcal{T}_d^{\text{sml}}| + 1) \sum_{n=1}^{N_1} \left( \sqrt{\frac{\log(T^3)}{2n} + bn} \right) + T \times \frac{2}{T} \tag{98}$$

(by Eq. (96) and the fact that $c^{br}(t) = n$ for each $n$ occurs at most once between two detection times)

$$\leq \left( \frac{T}{N_1} + 1 \right) \sum_{n=1}^{N_1} \left( \sqrt{\frac{\log(T^3)}{2n} + bn} \right) + 2 \tag{99}$$

(by Lemma 29)

$$\leq \tilde{O} \left( \frac{T}{N_1} \times (\sqrt{N_1} + b N_1^2) \right) = O(T^{1-d/3}) \tag{100}$$
where we used \( b, N_1 = T^{-d}, \tilde{O}(T^{(2d)/3}) \) in the last transformation. Moreover, by using Lemma 27, the expectation of the third term is bounded as

\[
\sum_{t=1}^{T} E[|\mu_t - \hat{\mu}_W|1[c(t) \geq N_1]] \leq T \times \left( 3bN_1 + \tilde{O} \left( \sqrt{\frac{1}{N_1}} \right) \right) = \tilde{O}(T^{1-d/3}),
\]  

(102)

which completes the proof.

**Appendix E. Proofs of MP-MAB**

We first clarify the notation: In bandit streams, let

\[
W^i(t) = \{t \in W(t) : i \in I(t)\}.
\]  

(103)

Let

\[
\mu_{i,W}(t) = \sum_{t \in W^i(t)} \mu_{i,t}
\]  

(104)

and its empirical estimate be

\[
\hat{\mu}_{i,W}(t) = \sum_{t \in W^i(t)} x_{i,t}.
\]  

(105)

In the bandit setting, only subset of arms \( I(t) \subseteq [K] \) are observable, and thus \( W^i \subseteq W \). Still, the following inequality holds with probability \( 1 - 2/T^p \):

\[
\forall W' \in \mathcal{W} |\mu_{i,W'} - \hat{\mu}_{i,W'}| \leq \sqrt{\frac{\log(T^2 + p)}{2|W^i|}}.
\]  

(106)

Eq. (106), which is the same form as Eq. (5), holds because it corrects \( T^2 \) multiplications. The union bound of Eq. (106) over all arms holds with \( 1 - K/T^p \).

**E.1 Proof of Theorem 20**

**Proof** Under stationary environment, Eq. (106) with \( p = 1 \) implies that ADR-bandit never makes a reset with high probability. The regret is bounded by

\[
\text{Reg}(T) \leq T \times \frac{2K}{T} + C_{\text{st}}^\Delta \frac{K \log T}{\Delta_{\text{min}}} + o(\log T)
\]  

(107)

\[
\leq C_{\text{st}}^\Delta \frac{K \log T}{\Delta_{\text{min}}} + o(\log T).
\]  

(108)
E.2 Proof of Lemma 19

In the following, we prove parts (A)–(C) of Lemma 19. Note that Lemma 19 describes the properties of Algorithm 5 as a base bandit algorithm when no exogenous reset occurred. When no reset occurs, the current window includes all the time steps from 1 to \( t \); \( W(t) = \{1, 2, \ldots, t-1\} \). During the proof, we utilize the following high-probability bounds

\[
\forall W' \in W \left| \mu_{i,W'} - \hat{\mu}_{i,W'} \right| \leq \sqrt{\frac{\log(T^4)}{2|W'|}},
\]

\[
\forall W' \in W \left| \mu_{i,W'}^E - \hat{\mu}_{i,W'}^E \right| \leq \sqrt{\frac{\log(T^4)}{2|W_i,E|}},
\]

where \( W_{i,E} = \{ t \in W : i \in I(t), i = (t \mod K) \} \). Eq. (109) holds with probability \( 1 - 2K/T^2 \) by using Eq. (106) with \( p = 2 \). Similar discussion yields the fact that Eq. (110) holds with probability \( 1 - 2K/T^2 \).

E.2.1 Proof of Lemma 19 (A)

Selection by Algorithm 5 has monitoring consistency since it keeps drawing each arm in \( \hat{I}^*(t) \) at least once in every \( K \) time steps, and at least one arm is consistently in \( \hat{I}^*(t) \) for all the time steps until the reset occurs.

E.2.2 Proof of Lemma 19 (B)

In the following, we show that Elimination-UCB has logarithmic DD regret in a stationary bandit environment. For ease of discussion, we assume all the means are distinct (i.e., \( \mu_i \neq \mu_j \) for \( i \neq j \)). Removing this restriction is not very difficult. Without loss of generality, we assume \( \mu_1 > \mu_2 > \mu_3 > \cdots > \mu_K \) and \( [L] \) be the set of optimal arms.

Let \( \Delta_j^+ = \mu_1 - \mu_j \) and \( \Delta_j^- = \mu_L - \mu_j \) and

\[
T_j^+ = \frac{4K \log(T^4)}{(\Delta_j^+)^2},
\]

\[
T_j^- = \frac{\log(T^4)}{(\Delta_j^-)^2}.
\]

Let the events be

\[ E^{(1)} = \bigcap_{i \in [K], t \in [T]} \{|\hat{\mu}_i(t) - \mu_i| \leq B_i(t)\} \]

\[ E^{(2)} = \bigcap_{i \in [K], t \in [T]} \{|\hat{\mu}_i^E(t) - \mu_i| \leq B_i^E(t)\} \]

\[ E^{(3)} = \bigcap_{t \in [T]} \{1 \in \hat{I}^*(t)\} \]

\[ E^{(4)} = \bigcap_{j>L, t \geq T_j^+ + 1} \{j \notin \hat{I}^*(t)\} \]
\[ \mathcal{E}^{(5)} = \bigcap_{j > L; N_j(t) \geq T_j^+} \{ U_j < \mu_L \}. \]  

(117)

**Lemma 30** Let \( \mathcal{E} = \{ \mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \mathcal{E}^{(3)}, \mathcal{E}^{(4)}, \mathcal{E}^{(5)} \} \). Then,

\[
\Pr[\mathcal{E}] \geq 1 - \frac{4K}{T}. 
\]  

(118)

**Proof** [Proof of Lemma 30] \( \mathcal{E}^{(1)} \) and \( \mathcal{E}^{(2)} \) are derived from\(^{17} \) Eqs. (109) and (110), each of which holds with probability \( 1 - 2K/T^2 \). Events \( \mathcal{E}^{(1)} \) and \( \mathcal{E}^{(2)} \) state that all the estimators are in the confidence region. Namely,

\[
\hat{\mu}_i(t) + B_i(t) \geq \mu_i \geq \hat{\mu}_i(t) - B_i(t)
\]

(119)

\[
\hat{\mu}_i^E(t) + B_i^E(t) \geq \mu_i \geq \hat{\mu}_i^E(t) - B_i^E(t).
\]

(120)

In the following, we show that \( \mathcal{E}^{(2)} \) implies \( \mathcal{E}^{(3)}, \mathcal{E}^{(4)}, \mathcal{E}^{(5)} \).

Under \( \mathcal{E}^{(2)} \), for a suboptimal arm \( j > L \)

\[
\hat{\mu}_1^E(t) + B_1^E(t) \geq \mu_1 > \mu_j \geq \hat{\mu}_j^E(t) - B_j^E(t),
\]

(121)

which implies that arm 1 is never eliminated, and thus \( \mathcal{E}^{(3)} \) holds. Moreover, under \( \mathcal{E}^{(2)} \cap \mathcal{E}^{(3)} \), assume that suboptimal arm \( j > L \) is in \( \hat{\mathcal{I}}^*(t) \) at \( t = T_j^+ + 1 \). Then arm \( j \) has been drawn in at least \( T_j^+ / K \) time steps, and thus

\[
B_j^E(t) = \sqrt{\frac{\log(T_j^+)}{4N_j^E(t)}} \leq \frac{\Delta_j^+}{4},
\]

(122)

and similarly for arm 1

\[
B_1^E(t) = \sqrt{\frac{\log(T_1^+)}{4N_1^E(t)}} \leq \frac{\Delta_1^+}{4}.
\]

(123)

Therefore we have

\[
\hat{\mu}_1^E(t) - B_1^E(t) \geq \mu_1 - 2B_1^E(t) \quad \text{(by } \mathcal{E}^{(2)})
\]

(124)

\[
\geq \mu_1 - \frac{\Delta_1^+}{2} \quad \text{(by Eq.\,(123))}
\]

(125)

\[
\geq \mu_j + \frac{\Delta_j^+}{2} \quad \text{(by definition)}
\]

(126)

\[
\geq \mu_j + 2B_j^E(t) \quad \text{(by Eq.\,(122))}
\]

(127)

\[
\geq \hat{\mu}_j^E(t) + B_j^E(t) \quad \text{(by } \mathcal{E}^{(2)})
\]

(128)

which implies \( \mathcal{E}^{(4)} \). Finally,

\[
\mu_L - U_j^E(t) \leq \Delta_j^- - B_j^E(t)
\]

(129)

\(^{17}\) By definition, \( \hat{\mu}_i(t) := \hat{\mu}_i,_{W(t)} \).
Lemma 31 (Regret per time step) Under $\mathcal{E}$, the following inequality holds:

$$\text{reg}(t) \leq \sum_{j > L} 1[j \in \hat{I}^*(t), (t \mod K) = j] \Delta_j^+ + \sum_{j \in I(t), j > L} B_j(t).$$  \hspace{1cm} (131)$$

Proof [Proof of Lemma 31] For time step $t$, let $k = k(t) = (t \mod K)$. Note that $I(t)$ depends on whether $k \in I^*(t)$ or not. In the former case, Elimination-UCB selects $k$ and top-$(L - 1)$ arms in terms of $U_i$. Otherwise, it selects top-$L$ arms in terms of $U_i$. Moreover, the cases of $k \leq L$ (top-$L$) or $k > L$ (suboptimal) require different analyses. We show Eq. (131) holds for all the cases.

(i) Case $k > L$ and $k \in \hat{I}^*(t)$:

$$\text{reg}(t) = \sum_{i \in [L]} \mu_i - \sum_{j \in I(t)} \mu_j$$  \hspace{1cm} (132)

$$= \sum_{i \in [L]} \mu_i - \mu_k - \sum_{j \in I(t), j \neq k} \mu_j \quad \text{(by $k \in \hat{I}^*(t)$)}$$  \hspace{1cm} (133)

$$= \Delta_k^+ + \sum_{i \in \{2, 3, \ldots, L\}} \mu_i - \sum_{j \in I(t), j \neq k} \mu_j$$  \hspace{1cm} (134)

$$\leq \Delta_k^+ + \sum_{i \in \{2, 3, \ldots, L\}} U_i(t) - \sum_{j \in I(t), j \neq k} (U_j(t) - 2B_j(t)) \quad \text{(by $\mathcal{E}$)}$$  \hspace{1cm} (135)

$$\leq \Delta_k^+ + \sum_{j \in I(t), j \neq k} 2B_j(t)$$  \hspace{1cm} (136)

($I(t)$ chooses top-$(L - 1)$ of $U_i$ indicates $\sum_{j \in I(t), j \neq k} U_j(t) - \sum_{i \in \{2, 3, \ldots, L\}} U_i(t) \geq 0$)  \hspace{1cm} (137)

and thus Eq. (131) holds in this case.

(ii) Case $k \leq L$ and $k \in \hat{I}^*(t)$:

$$\text{reg}(t) = \sum_{i \in [L]} \mu_i - \sum_{j \in I(t)} \mu_j$$  \hspace{1cm} (138)

$$= \sum_{i \in [L] \setminus \{k\}} \mu_i - \sum_{j \in I(t), j \neq k} \mu_j \quad \text{(by $k \in \hat{I}^*(t)$)}$$  \hspace{1cm} (139)

$$\leq \sum_{i \in [L] \setminus \{k\}} U_i(t) - \sum_{j \in I(t), j \neq k} (U_j(t) - 2B_j(t)) \quad \text{(by $\mathcal{E}$)}$$  \hspace{1cm} (140)

$$\leq \sum_{j \in I(t), j \neq k} 2B_j(t)$$  \hspace{1cm} (141)
(by the fact that $I(t)$ chooses top-$(L - 1)$ of $U_i$) \hfill (142)

and thus Eq. (131) holds in this case.

(iii) Case $k \notin I(t)$:

$$\text{reg}(t) = \sum_{i \in [L]} \mu_i - \sum_{j \in I(t)} \mu_j$$ \hfill (143)

$$\leq \sum_{i \in [L]} U_i(t) - \sum_{j \in I(t)} (U_j(t) - 2B_j(t)) \quad \text{(by $\mathcal{E}$)}$$ \hfill (144)

$$\leq \sum_{j \in I(t)} 2B_j(t)$$ \hfill (145)

(by the fact that $I(t)$ chooses top-$L$ of $U_i$) \hfill (146)

and thus Eq. (131) holds in this case. Since the cases (i)–(iii) are exhaustive, we conclude the proof of Lemma 31. \hfill \blacksquare

The following lemma bounds the regret from the arms in $\hat{I}^*(t)$.

**Lemma 32** (Regret due to elimination) Under $\mathcal{E}$, for $j > L$,

$$\sum_{t=1}^{T} \mathbb{1}[j \in \hat{I}^*(t), (t \mod K) = j] \Delta_j^+ \leq (T_j^+ + 1) \Delta_j^+. \hfill (147)$$

**Proof** [Lemma 32] $\mathcal{E}^{(4)}$ implies $\{j \in \hat{I}^*(t), (t \mod K) = j\}$ occurs at most $T_j^+ + 1$ times, which immediately completes the proof of Lemma 32. \hfill \blacksquare

The following theorem bounds the total size of the confidence bounds.

**Lemma 33** (Regret due to confidence bound) Under $\mathcal{E}$,

$$\sum_t \mathbb{1}[j \in I(t)] B_j(t) \leq \frac{2 \log(T^4)}{\Delta_j^-}$$ \hfill (148)

holds for $j > K$.

**Proof** [Proof of Lemma 33] Let

$$B(n) = \sqrt{\frac{\log(T^4)}{2n}}.$$ \hfill (149)

That is, the size of $B_j$ at $N_j = n$.

$$\sum_t \mathbb{1}[j \in I(t)] B_j(t)$$ \hfill (150)

$$= \sum_n \sum_t \mathbb{1}[j \in I(t), N_j(t) = n] B(n)$$ \hfill (151)
\[
\leq \sum_n \mathbf{1} \left( \bigcap_t \{ j \in I(t), N_j(t) = n \} \right) B(n)
\]

(by \( j \in I(t), N_j(t) = n \) occurs at most once)

\[
\leq \sum_{n \leq T^-} B(n)
\]

(by \( \mathcal{E}^{(4)}, \mathcal{E}^{(5)} \), for \( N_j(t) \geq T^-_j \), arm \( U_j \) exceeds none of \( \{U_1, \ldots, U_L\} \))

\[
< \frac{2 \log(T^4)}{\Delta_j^-} \quad (\text{by } \sum_{n=1}^N (1/\sqrt{n}) < 2\sqrt{N}),
\]

which completes the proof of Lemma 33.

In the following, we finally bound the regret.

\[
\text{Reg}(T) = \sum_{t=1}^T \text{reg}(T)
\]

\[
\leq T \mathbf{1}[\mathcal{E}] + \mathbf{1}[\mathcal{E}] \sum_t \left( \sum_{j > L} \mathbf{1}[j \in \hat{I}^*(t), (t \mod K) = j] \Delta_j^+ + \sum_{j \in \hat{I}^*(t), j > L} B_j(t) \right)
\]

(by Lemma 31)

\[
\leq T \mathbf{1}[\mathcal{E}] + \sum_{j > L} O \left( \frac{\log T}{\Delta_j^-} \right) \quad (\text{by Lemmas 32 and 33})
\]

Taking expectation of Eq. (157) we have

\[
\mathbb{E}[\text{Reg}(T)] \leq T \mathbb{P}[\mathcal{E}] + \sum_{j > L} O \left( \frac{\log T}{\Delta_j^-} \right)
\]

\leq T \times \frac{4K}{T^2} + \sum_{j > L} O \left( \frac{\log T}{\Delta_j^-} \right) \quad (\text{by Lemma 30})

\leq \sum_{j > L} O \left( \frac{\log T}{\Delta_j^-} \right),
\]

which states that Elimination-UCB has a logarithmic DD regret.
E.2.3 Proof of Lemma 19 (C)

In the following, we derive a strong DT regret bound of Algorithm 5. The discussion here partly follows similar steps to the one for the stationary environment (i.e., part (B) before). The largest difference is that $\mu_{i,t}$ drifts up to $\epsilon(t)$. Accordingly, we modify several definitions. Without loss of generality, we assume arms are indexed by the $\mu_{i,t}$ at time step $t = 1$. Namely, $\mu_{1,1} \geq \mu_{2,1} \geq \cdots \geq \mu_{K,1}$. By the definition of $\epsilon(t)$,

$$|\mu_{i,W(t)} - \mu_{i,t}| \leq \epsilon(t)$$

holds for any $i \in [K]$ and $W(t)$ with $N_i(t) > 0$. Let

$$\Delta^+_j = \mu_{1,1} - \mu_{j,1},$$

$$\Delta_j(t) = \max_{i \in [K]} \mu_{i,t} - \mu_{j,t}.$$ 

Let

$$T^+_j = \frac{8K \log(T^4)}{(\Delta^+_j)^2}.$$ 

Let

$$B(n) = \sqrt{\frac{\log(T^4)}{2n}}.$$ 

That is, the size of $B_j$ at $N_j = n$. Let the events be

$$G^{(1)} = \bigcap_{i \in [K], t \in [T]} \{|\hat{\mu}_{i,W(t)} - \mu_{i,W(t)}| \leq B_i(t)\}$$

$$G^{(2)} = \bigcap_{i \in [K], t \in [T]} \{|\hat{\mu}_{i,E(t)} - \mu_{i,W(t)}| \leq B_i^E(t)\}$$

$$G^{(3)} = \bigcap_{t > K, j \in \hat{I}^*(t)} \{\Delta_j(t) \leq 4B(t/K) + (C^{\text{ch}} + 4)\epsilon(t)\}$$

where $C^{\text{ch}}$ is a universal constant defined later in Lemma 34.

For the sake of analysis, we define the notion of elimination chain below.

**Definition 2 (Elimination chain)** We build a chain as follows. First, append arm 1 to the chain. If there is a time step $s_1 < t$ such that arm 1 is eliminated from $\hat{I}^*$ by another arm $i_2$, then append $i_2$ to the tail of the chain. If arm $i_2$ is eliminated by another arm $i_3$ before $t$, then append the arm $i_3$ to the tail of the chain. Repeat this procedure until the time step reaches at the end of time step $t$. Let $1, i_2, \ldots, i_l \in [K]$ denotes the elements of the chain. Let $s_1 < s_2 < \cdots < s_l$ be the time steps that arms are eliminated. For convenience, let $s_0 = 0$.

We use the elimination chain to clarify to keep track of the drift of $\mu_{i,t}$ over time.

**Lemma 34 (Quality lower bound of the element of the elimination chain)** Under $G^{(2)}$, the following inequality holds:

$$\mu_{i_l,1} \geq \mu_{1,1} - C^{\text{ch}}\epsilon(s_l),$$

where $C^{\text{ch}} = 66.$
Lemma 34 states that $\mu_{i,1} - \mu_{1,1}$ is proportional to $\epsilon(s_t)$. The largest challenge here is that the length of the elimination chain can be $K - L$. However, as long as the number of draws are controlled (i.e., there exists and absolute constant $C$ such that $C \leq \mathcal{N}_i^E(t)/\mathcal{N}_j^E(t)$ for all $i, j, t$), the amount of total drift does not depend on $L, K$.

**Proof** [Proof of Lemma 34] For $m \in [l]$,

$$w(m) = s_m - s_{m-1}$$

$$w(i, m) = \sum_{s=s_{m-1}+1}^{s_m} 1[i \in I(t), i = (s \mod K)]$$

be the number of time steps (resp. the number of time steps arm $i$ is drawn by the monitoring subroutine) between $(m - 1)$-th and $m$-th elimination. Notice that during these time steps, due to the property of Elimination-UCB, we have

$$\frac{w(m)}{2K} \leq \left\lfloor \frac{w(m)}{K} \right\rfloor \leq w(j, m) \leq \left\lfloor \frac{w(m)}{K} \right\rfloor + 1 \leq \frac{w(m)}{K} + 1$$

for all\(^\text{18}\) $j$ that is in $\hat{I}^*$ at time step $w_{l-1} + 1$. For $m \in [l]$, let

$$w(; m) = \sum_{n=1}^m w(n)$$

$$w(i, ; m) = \sum_{n=1}^m w(i, n).$$

For $i \in [K], m \in [l]$, let

$$u_{i,m} = \frac{\sum_{s=s_{m-1}+1}^{s_m} \mu_{i,s} 1[i \in I(t), i = (s \mod K)]}{w(i, m)}$$

be the number of time steps (resp. the number of time steps arm $i$ is drawn by the monitoring subroutine) between $(m - 1)$-th and $m$-th elimination. Notice that during these time steps, due to the property of Elimination-UCB, we have

$$\frac{w(m)}{2K} \leq \left\lfloor \frac{w(m)}{K} \right\rfloor \leq w(j, m) \leq \left\lfloor \frac{w(m)}{K} \right\rfloor + 1 \leq \frac{w(m)}{K} + 1$$

for all\(^\text{18}\) $j$ that is in $\hat{I}^*$ at time step $w_{l-1} + 1$. For $m \in [l]$, let

$$w(; m) = \sum_{n=1}^m w(n)$$

$$w(i, ; m) = \sum_{n=1}^m w(i, n).$$

For $i \in [K], m \in [l]$, let

$$u_{i,m} = \frac{\sum_{s=s_{m-1}+1}^{s_m} \mu_{i,s} 1[i \in I(t), i = (s \mod K)]}{w(i, m)}$$

be the number of time steps (resp. the number of time steps arm $i$ is drawn by the monitoring subroutine) between $(m - 1)$-th and $m$-th elimination. Notice that during these time steps, due to the property of Elimination-UCB, we have

$$\frac{w(m)}{2K} \leq \left\lfloor \frac{w(m)}{K} \right\rfloor \leq w(j, m) \leq \left\lfloor \frac{w(m)}{K} \right\rfloor + 1 \leq \frac{w(m)}{K} + 1$$

for all\(^\text{18}\) $j$ that is in $\hat{I}^*$ at time step $w_{l-1} + 1$. For $m \in [l]$, let

$$w(; m) = \sum_{n=1}^m w(n)$$

$$w(i, ; m) = \sum_{n=1}^m w(i, n).$$

By definition, $\mu_{i, W(s_m)}^E$ can be represented as a weighted average of $u_{i,m}$ as

$$\mu_{i, W(s_m)}^E = \frac{1}{w(i, ; m)} \sum_{m=1}^l w(i, m)u_{i,m}$$

Let $\mu_{i,0} = \mu_{i,1}, \eta(i, m) = \sum_{s=s_{m-1}+1}^{s_m} |\mu_{i,s-1} - \mu_{i,s}|$ and $\eta(m) = \max_i \eta(i, m)$. We have

$$|u_{i,m} - u_{i,m-1}| \leq |\max\{\mu_{i,s} : s = s_{m-1} + 1, s = s_{m-1} + 2, \ldots, s_m\} - \min\{\mu_{i,s} : s = s_{m-2} + 1, s = s_{m-2} + 2, \ldots, s_{m-1}\}|$$

$$\leq \eta(i, m - 1) + \eta(i, m)$$

$$\leq \eta(m - 1) + \eta(m).$$

\(^\text{18}\) Note that we assume $w(j, m) \geq 1$ to derive Eq. (172). Otherwise $D_{j,m} = 0$, and arm $j$ does not compromise the first transformation of Eq. (185).
Globally Nonstationary Bandits

and thus for $n < m$,

$$|u_{i,n} - u_{i,m}| \leq \sum_{m'=m}^{n-1} |u_{i,m'+1} - u_{i,m'}|$$  (181)

$$\leq \sum_{m'=m}^{n-1} (\eta(m') + \eta(m' + 1))$$  (182)

$$\leq 2 \sum_{m'=m}^{n} \eta(m').$$  (183)

For $m \in \{2, 3, \ldots, l\}$, let

$$D_{i,m} = \mu_{i,W(s_m)}^E - \mu_{i,W(s_{m-1})}^E.$$  (184)

By using Eq. (172), we have

$$D_{i,m} = \frac{1}{w(i, : m)} \sum_{n=1}^{m} w(i, n)u_{i,n} - \frac{1}{w(i, : m-1)} \sum_{n=1}^{m-1} w(i, n)u_{i,n} \quad \text{(by Eq. (176))}$$  (185)

$$= \sum_{n=1}^{m-1} \frac{w(i, m)w(i, n)}{w(i, : m-1)w(i, : m)} (u_{i,m} - u_{i,n})$$  (186)

$$\leq 4 \sum_{n=1}^{m-1} \left( \frac{w(m)}{K} + 1 \right) \left( \frac{w(n)}{K} + 1 \right) (u_{i,m} - u_{i,n}) \quad \text{(by Eq. (172))}$$  (187)

$$\leq 16 \sum_{n=1}^{m-1} \frac{w(m)w(n)}{w(:, m-1)w(:, m)} (u_{i,m} - u_{i,n}).$$  (188)

Here,

$$\sum_{m=2}^{l} \max_{j \in [K]} |D_{j,m}| \leq 16 \sum_{m=2}^{l} \sum_{n=1}^{m-1} \frac{w(m)w(n)}{w(:, m-1)w(:, m)} \max_{j \in [K]} |u_{j,m} - u_{j,n}|$$  (189)

$$\leq 16 \sum_{m=2}^{l} \sum_{n=1}^{m-1} \frac{w(m)w(n)}{w(:, m-1)w(:, m)} \left( 2 \sum_{o=n}^{m} \eta(o) \right) \quad \text{(by Eq. (183))}$$  (190)

$$= 32 \sum_{o=1}^{l} \eta(o) \sum_{m=o}^{l} \sum_{n=1}^{\min(o,m-1)} \frac{w(m)w(n)}{w(:, m-1)w(:, m)}$$  (o runs between $n$ and $m)$  (191)

$$\leq 32 \sum_{o=1}^{l} \eta(o) \left( \sum_{m=0}^{l-1} \sum_{n=1}^{o-1} \frac{w(m)w(n)}{w(:, m-1)w(:, m)} + 1 \right)$$  (192)

$$= 32 \sum_{o=1}^{l} \eta(o) \left( \sum_{m=0}^{l-1} \frac{w(o-1)w(m)}{w(:, m-1)w(:, m)} + 1 \right).$$  (193)
For $m \geq o$, let $c_m = \frac{w(m)}{w(o-1)}$. Then,

\begin{align*}
w(:, m) &= w(:, o - 1) + w(o) + \cdots + w(m) \\
&= w(:, o - 1)(1 + c_o + c_{o+1} + \cdots + c_m),
\end{align*}

and

\begin{align*}
\sum_{m=o}^{l} \frac{w(:, o - 1)w(m)}{w(:, m - 1)w(:, m)} &= \sum_{m=o}^{l} \frac{w(m)}{(1 + c_o + c_{o+1} + \cdots + c_{m-1})w(:, m)} \\
&= \sum_{m=o}^{l} \frac{c_m}{(1 + c_o + c_{o+1} + \cdots + c_{m-1})(1 + c_o + c_{o+1} + \cdots + c_m)}.
\end{align*}

We show Eq. (198) is smaller than 1. Namely, let

\begin{align*}
f_l &= f_l(c_o, c_{o+1}, \ldots, c_l) = \sum_{m=o}^{l} \frac{c_m}{(1 + c_o + c_{o+1} + \cdots + c_{m-1})(1 + c_o + c_{o+1} + \cdots + c_m)}
\end{align*}

and we show

\begin{align*}
F_l &= \max \{f_l(c_o, c_{o+1}, \ldots, c_l) : c_o, c_{o+1}, \ldots, c_l > 0\} < 1.
\end{align*}

Letting $1 + c_o = p_o$,

\begin{align*}
f_l &= \sum_{m=o}^{l} \frac{c_m}{(1 + c_o + c_{o+1} + \cdots + c_{m-1})(1 + c_o + c_{o+1} + \cdots + c_m)} \\
&= \frac{p_o - 1}{p_o} + \sum_{m=o+1}^{l} \frac{c_m}{(p_o + c_{o+1} + \cdots + c_{m-1})(p_o + c_{o+1} + \cdots + c_m)} \\
&= \frac{p_o - 1}{p_o} + \frac{1}{p_o} f_{l-1}(c_o/p_o, \ldots, c_m/p_o)
\end{align*}

and thus

\begin{align*}
F_l &\leq \max_{p_o > 1} \left( \frac{p_o - 1}{p_o} + \frac{1}{p_o} F_{l-1} \right).
\end{align*}

(i) $F_o = \max_{c_o > 0} c_o/(1 + c_o) < 1$. (ii) $F_m < 1$ and Eq. (204) implies $F_{m+1} < 1$. By induction with (i)(ii) we have $F_l < 1$. In summary,

\begin{align*}
\sum_{m=2}^{l} \max_{j \in \mathcal{K}} |D_{j,m}| &\leq 32 \sum_{o=1}^{l} \eta(o) \left( \sum_{m=o}^{l} \frac{w(:, o - 1)w(m)}{w(:, m - 1)w(:, m)} + 1 \right) \\
&\leq 32 \sum_{o=1}^{l} \eta(o) (1 + 1) \quad \text{(by Eq. (201))}
\end{align*}
Globally Nonstationary Bandits

\[
\leq 64 \sum_{o=1}^{l} \eta(o). \tag{207}
\]

Let \( \mu_{1,W(s)^t} = \max_{i \in I^*(t)} \mu_{i,W(s)^t} \) and Let \((1,W(s)) = \arg \max_{i \in I^*(t)} \mu_{i,W(s)^t}\) (ties are broken in an arbitrary way). We finally have,

\[
\mu_{j,W(s)} \leq \mu_{1,W(s)} \tag{208}
\]

(by \( G^{(2)} \))

\[
\leq \mu_{(1,W(s)),W(s)} \tag{209}
\]

(by definition)

\[
\leq \mu_{(1,W(s)),W(s-1)} + \max_{j \in [K]} |D_{j,l}| \tag{210}
\]

\[
\leq \mu_{(1,W(s-1)),W(s-1)} + \max_{j \in [K]} |D_{j,l}| \tag{211}
\]

(by \( G^{(2)} \))

\[
\leq \mu_{(1,W(s-1))} + \max_{j \in [K]} |D_{j,l}| \tag{212}
\]

\[
\leq \ldots \tag{213}
\]

\[
\leq \mu_{(1,W(s-2))} + \max_{j \in [K]} |D_{j,l}| + \max_{j \in [K]} |D_{j,l-1}| \tag{214}
\]

\[
\leq \ldots \tag{215}
\]

\[
\leq \mu_{(1,w(s_1))} + \sum_{m=2}^{l} \max_{j \in [K]} |D_{j,l'}| \tag{216}
\]

\[
\leq \mu_{1,1} + \eta(1) + \sum_{m=2}^{l} \max_{j \in [K]} |D_{j,l'}| \tag{217}
\]

\[
\leq \mu_{1,1} + \eta(1) + 64 \sum_{o=1}^{l} \eta(o) \tag{by Eq. (207)} \tag{218}
\]

\[
\leq \mu_{1,1} + 65 \sum_{o=1}^{l} \eta(o) \tag{219}
\]

\[
\leq \mu_{1,1} + 65 \epsilon(s_l) \tag{220}
\]

which, combined with \(|\mu_{j,W(s)} - \mu_{j,l}^t| \leq \epsilon(s_l)\), yields Eq. (169).

\[\]

Lemma 35 Let \( G = \{G^{(1)}, G^{(2)}, G^{(3)}\} \). Then,

\[
\Pr[G] \geq 1 - \frac{4K}{T^2} \tag{221}
\]

Proof [Proof of Lemma 35] Event \( G^{(1)} \) is equivalent to Eq. (109), which holds with probability \( 1 - 2K/T^2 \). Similar discussion shows that Event \( G^{(2)} \) holds with probability \( 1 - 2K/T^2 \). Taking a union bound, we see \( G^{(1)}, G^{(2)} \) hold with probability \( 1 - 4K/T^2 \). We next show that \( G^{(1)}, G^{(2)} \) implies \( G^{(3)} \).

Let \( i_1, i_2, \ldots, i_l \) be the elements in the elimination chain at round \( t \). We have,

\[
\max_i \mu_i - \mu_{j,l} \leq \max_i \mu_{i,1} - \mu_{j,1} + 2 \epsilon(t) \tag{by definition} \tag{222}
\]
\[
\leq \mu_{ii,1} - \mu_{jj,1} + (C^{ch} + 2)\epsilon(t) \quad \text{(by Lemma 34)} \tag{223}
\]
\[
\leq \mu_{ii,W(t)}^E - \mu_{jj,W(t)}^E + (C^{ch} + 4)\epsilon(t) \tag{224}
\]
\[
\leq 4B(t/K) + (C^{ch} + 4)\epsilon(t) \quad \text{(since arm } j \text{ is not eliminated and } G^{(1)}) \tag{225}
\]

which is \(G^{(3)}\) and completes the proof of Lemma 35.

We next bound regret per time step \(\text{reg}(t)\). The following Lemma 36 is a version ofLemma 31.

**Lemma 36** (Regret per time step) Under \(G\), the following inequality holds:

\[
\text{reg}(t) \leq \sum_{j > L} 1 \left[ j \in \hat{I}^*(t), (t \mod + K) = j \right] (4B(t/K)) + \sum_{j \in I(t)} 2B_j(t) + (2L + C^{ch} + 2)\epsilon(t).
\tag{226}
\]

**Proof** [Proof of Lemma 36] Let \(k = t \mod + K\). In the following, we only consider (1) the case \(k > L\) and \(k \in \hat{I}^*(t)\).

\[
\text{reg}(t) = \max_{I:|I|=L} \sum_{i \in I} \mu_{i,t} - \sum_{i \in I(t)} \mu_{i,t} \tag{227}
\]

\[
\leq (4B(t/K) + (C^{ch} + 4)\epsilon(t)) + \max_{i \in I:|I|=L-1} \sum_{i \in I} \mu_{i,t} - \sum_{i \in I(t),i \neq k} \mu_{i,t} \quad \text{(by } G^{(3)}) \tag{228}
\]

\[
\leq (4B(t/K) + (C^{ch} + 4)\epsilon(t)) + \max_{i \in I:|I|=L-1} \sum_{i \in I} \mu_{i,W(t)} - \sum_{i \in I(t),i \neq k} \mu_{i,W(t)} + 2(L - 1)\epsilon(t) \tag{229}
\]

(by definition of \(\epsilon(t)\))

\[
\leq (4B(t/K) + (C^{ch} + 4)\epsilon(t)) + \max_{i \in I:|I|=L-1} \sum_{i \in I} U_i - \sum_{i \in I(t),i \neq k} (U_i - 2B_i) + 2(L - 1)\epsilon(t) \tag{230}
\]

(by \(G^{(1)}\))

\[
\leq (4B(t/K) + (C^{ch} + 4)\epsilon(t)) + \sum_{i \in I(t),i \neq k} 2B_i + 2(L - 1)\epsilon(t) \tag{231}
\]

(by the fact that \(I(t)\) chooses top-(\(L - 1\)) of \(U_i\))

\[
\leq 4B(t/K) + \sum_{i \in I(t)} 2B_i + (2L + C^{ch} + 2)\epsilon(t) \tag{232}
\]

and Eq. (226) holds. The case (2) \(k \leq L\) and \(k \in \hat{I}^*(t)\) and (3) \(k \notin \hat{I}^*(t)\) can be shown easily, which completes the proof of Lemma 36.
Lemma 37  For any $S \leq T$, $\mathcal{G}$ implies

$$
\sum_{t=1}^{S} \sum_{j>L} 1[j \in \hat{I}^*(t), (t \mod + K) = j](4B(t/K)) \leq 4K \sqrt{2 \log(T^4)(S/K)}.
$$

(233)

Proof [Proof of Lemma 37]

$$
\sum_{t=1}^{S} \sum_{j>L} 1[j \in \hat{I}^*(t), (t \mod + K) = j](4B(t/K)) \leq 4K \sum_{s=1}^{S/K} B(s) \leq 4K \sqrt{2 \log(T^4)(S/K}).
$$

(234)

(by $\sum_{n=1}^{N} (1/\sqrt{n}) < 2\sqrt{N}$)

(235)

(236)

Lemma 38  For any $S \leq T$, $\mathcal{G}$ implies

$$
\sum_{t \leq S} \sum_{j \in I(t)} 2B_j < 2\sqrt{2LK}S \log(T^4).
$$

(237)

Proof [Proof of Lemma 38] For arbitrary $j \in [K]$, we have

$$
\sum_{t \leq S} 1[j \in I(t)]B_j = \sum_{n=1}^{N_i(S)} \sum_{t} 1[j \in I(t), N_j(t) = n]B(n)
$$

(238)

$$
\leq \sum_{n=1}^{N_i(S)} B(n) \quad \text{(by $\{j \in I(t), N_j(t) = n\}$ occurs at most once)}
$$

(239)

$$
= \sum_{n=1}^{N_i(S)} B(n)
$$

(240)

$$
< \sqrt{2 \log(T^4)N_i(S)}
$$

(241)

and thus

$$
\sum_{t \leq S} \sum_{j \in I(t)} 2B_j < 2 \sum_{j>L} \sqrt{2 \log(T^4)N_i(S)} \quad \text{(by Eq. (241))}
$$

(242)

$$
\leq 2\sqrt{2 \log(T^4)} \sqrt{K \sum_{j} N_i(S)} \quad \text{(by Cauchy-Schwarz inequality)}
$$

(243)

$$
\leq 2\sqrt{2 \log(T^4)\sqrt{LKS}} \quad \text{(by MP-MAB draws $L$ arms at each time step)}
$$

(244)
Finally, the regret at any round $S \leq T$ under $G$ is bounded as

$$\text{Reg}(S) = \sum_{t=1}^{S} \text{reg}(t) \leq \sum_{t=1}^{S} \left( \sum_{j>L} 1[j \in \hat{I}^*(t), (t \mod^+ K) = j](4B(t/K)) + \sum_{j \in I(t)} 2B_j + (2L + C^{ch} + 2)\epsilon(t) \right)$$

(by Lemma 36)

$$\leq 4K\sqrt{2 \log(T^4)(S/K)} + 2\sqrt{2LK}S\log(T^4) + \sum_{t \leq S} (2L + C^{ch} + 2)\epsilon(t)$$

(by Lemmas 37 and 38)

$$\leq \tilde{O}(\sqrt{2KL}) + O(L) \sum_{t \leq S} \epsilon(t),$$

which bounds the DT regret of Algorithm 5 with high probability.

E.3 Proof of Theorem 22

Proof [Proof of Theorem 22]

Similar to the case of single stream, we define detection times, which is the subset of rounds where the reset occurs. Let $T_d = \{t \in [T] : |W(t+1)| \leq |W(t)|\}$ be the set of detection times, and $M_d = |T_d|$ be the number of detection times\(^{19}\). Let $T_{d,n}$ be the $n$-th element of $T_d$. For convenience, let $(T_{d,0}, T_{d,M_d+1}) = (0, T)$. We denote $S_n = \{T_{d,n} + 1, T_{d,n} + 2, \ldots, T_{d,n+1}\}$ for $n \in \{0, \ldots, M_d\}$. $S_n$ is the interval between $n$-th and $(n+1)$-th detection times. Let $S_n = |S_n|$. By the nonintervening property (Remark 11), we can decompose the regret into the regret of base bandit algorithm for each interval.

We define an event

$$\mathcal{V} = \{\forall m \in [M] \ \exists s \in T_d \ 0 \leq s - T_m \leq 16DU(\epsilon_m) \} \cap \{M_d = M\}. \quad (249)$$

Event $\mathcal{V}$ states that for each changepoint $T_m$, there exists a detection time $T_{d,m}$ within $16DU(\epsilon_m)$ time steps. We first show that $\mathcal{V}$ holds with high probability by induction. Assume that for every changepoints $m'$ up to $1, 2, \ldots, m-1$, there exists unique detection time $s_{m'}$ such that $0 \leq s_{m'} - T_{m'} \leq 16DU(\epsilon_m)$. By assumption, $T_m - s_{m-1} \geq 16DU(\epsilon_m)$. We show that $0 \leq s_m - T_m \leq 16DU(\epsilon_m)$. During $t \leq T_m$ (these time steps are between $m$-th

19. Remember the difference between changepoints $T$ and detection times $T_d$. The former is defined on an abrupt environment, whereas the latter is defined for the ADR-bandit algorithm (and thus the latter is a random variable).
and \((m - 1)\)-th changepoint), \(\mu_{i,t}\) stays the same. We denote \(\mu_i = \mu_{i,t}\) during these time steps. Eq. (106) implies that, with probability \(1 - 2K/T\), for any split \(W(t) = W_1 \cup W_2\),

\[
\forall i \ |\mu_i - \hat{\mu}_{i,W_1}| \leq \sqrt{\frac{\log(T^3)}{2|W_{i,1}|}}
\]

\[
\forall i \ |\mu_i - \hat{\mu}_{i,W_2}| \leq \sqrt{\frac{\log(T^3)}{2|W_{i,2}|}}
\]

where \(W_{i,1} = \{t \in W_1 : i \in I(t)\}\) and \(W_{i,2} = \{t \in W_2 : i \in I(t)\}\). This implies \(ADR\)-bandit with \(\delta = T^3\) never makes a split before the \(m\)-th changepoint (i.e., \(0 \leq s_m - T_m\)). Let \(s = T_m + 16DU(\epsilon_m)\). Assume that there is no detection between time step \(T_m - 16DU(\epsilon_m)\) and \(T_m + 16DU(\epsilon_m)\). Then for a split \(W(s) = W_1 \cup W_2, W_1 = W \cap [T_m], W_2 = W \setminus W_1\),

\[|W_{i,1}|, |W_{i,2}| \geq 16DU(\epsilon_m).\]

By definition of the monitoring consistency, there exists an arm \(i \in [K]\) that

\[|W_{i,1}|, |W_{i,2}| \geq 16U(\epsilon_m).\]

Again by Eq. (106) we have

\[
\forall i \ |\mu_{i,T_m} - \hat{\mu}_{i,W_1}| \leq \sqrt{\frac{\log(T^3)}{2|W_{i,1}|}}
\]

\[
\forall i \ |\mu_{i,T_{m+1}} - \hat{\mu}_{i,W_2}| \leq \sqrt{\frac{\log(T^3)}{2|W_{i,2}|}}.
\]

By definition of the global changepoint,

\[|\mu_{i,T_m} - \mu_{i,T_{m+1}}| = \epsilon_m = \sqrt{\frac{\log(T^3)}{2U(\epsilon_m)}}\]

Combining these yields

\[|\hat{\mu}_{i,W_1} - \hat{\mu}_{i,W_2}| \geq |\mu_{i,T_m} - \mu_{i,T_{m+1}}| - |\mu_{i,T_m} - \hat{\mu}_{i,W_1}| - |\mu_{i,T_{m+1}} - \hat{\mu}_{i,W_2}|\]

(by triangular inequality)

\[
\geq \sqrt{\frac{\log(T^3)}{2U(\epsilon_m)}} - |\mu_{i,T_m} - \hat{\mu}_{i,W_1}| - |\mu_{i,T_{m+1}} - \hat{\mu}_{i,W_2}|
\]

\[
\geq 4 \max \left( \sqrt{\frac{\log(T^3)}{2|W_{i,1}|}}, \sqrt{\frac{\log(T^3)}{2|W_{i,2}|}} \right) - |\mu_{i,T_m} - \hat{\mu}_{i,W_1}| - |\mu_{i,T_{m+1}} - \hat{\mu}_{i,W_2}|
\]
(by Eq. (253))
\[
\geq \sqrt{\frac{\log(T^3)}{2|W_{i,1}|}} + \sqrt{\frac{\log(T^4)}{2|W_{i,2}|}}.
\]
(by Eq. (254), (255))

which implies that ADR-bandit resets the window at round 1. In summary, under Eq. (106) with \( p = 1 \), \( \mathcal{V} \) holds. Therefore,
\[
\Pr[\mathcal{V}] \geq 1 - 2K/T.
\]

In the following, we bound the regret. Let
\[
\text{Reg}_m = \sum_{t=T_{d,m}}^{T_{d,m+1}} \text{reg}(t).
\]
Namely, \( \text{Reg}_m \) corresponding to the regret between \( m \)-th and \( m+1 \)-th detection times. The regret is decomposed as
\[
\text{Reg}(T) \leq T_1[\mathcal{V}^c] + 1[\mathcal{V}]\text{Reg}(T)
\]
\[
\leq T_1[\mathcal{V}^c] + 1[\mathcal{V}]\sum_{m=0}^{M} \text{Reg}_m. \quad (\mathcal{V} \text{ implies } |T_d| = M)
\]

In what follows, we omit the regret in the case of \( \mathcal{V}^c \) because
\[
T \Pr[\mathcal{V}^c] \leq 2K
\]
is negligible. Under \( \mathcal{V} \),
\[
T_{d,m+1} - T_{d,m} \leq T_{m+1} + T_m + 16DU(\epsilon_m)
\]
By the nonintervening property, the sublinear DT regret of the base bandit algorithm, and Eq. (265), we have
\[
\mathbb{E}[\text{Reg}_m] \leq C_{dt} \left( \sqrt{LK(T_{d,m+1} - T_{d,m} + 16DU(\epsilon_{m+1}))} + LC^{ab}\epsilon_m 16DU(\epsilon_{m+1}) \right)
\]
where we define \( \epsilon_{M+1}, U(\epsilon_{M+1}) = 0 \). We finally obtain
\[
\mathbb{E} \left[ \sum_{m=0}^{M} \text{Reg}_m \right] \leq C_{dt} \left( \sqrt{LK(T_{d,m+1} - T_{d,m} + 16DU(\epsilon_{m+1}))} + LC^{ab}\epsilon_{m+1} 16DU(\epsilon_{m+1}) \right)
\]
(by Eq. (266))
\[
\leq \sum_{m=0}^{M} C^\text{dt} \left( \sqrt{2LK(T_{d,m+1} - T_{d,m}) + LC_{ab}\epsilon_{m+1}16DU(\epsilon_{m+1})} \right)
\]

(by \((T_{m+1} - T_m) \geq 32DU(\epsilon_{m+1})\))
\[
\leq C^\text{dt} \left( \sqrt{2L(M+1)T} + \sum_{m=0}^{M} LC_{ab}\epsilon_{m+1}16DU(\epsilon_{m+1}) \right)
\]

(by the Cauchy-Schwarz inequality and \(\sum_m (T_{m+1} - T_m) = T\))
\[
\leq C^\text{dt} \left( \sqrt{2L(M+1)T} + \sum_{m=0}^{M} LC_{ab}\epsilon_{m+1} \min \left( 16DU(\epsilon_{m+1}), \frac{T_{m+1} - T_m}{2} \right) \right)
\]

(by \((T_{m+1} - T_m) \geq 32DU(\epsilon_{m+1})\))
\[
= C^\text{dt} \left( \sqrt{2L(M+1)T} + \sum_{m=0}^{M} LC_{ab}\epsilon_{m+1} \min \left( 16D\log(T^3), \epsilon_{m+1}, \frac{T_{d,m+1} - T_{d,m}}{2} \right) \right)
\]

(by the Cauchy-Schwarz inequality and \(\sum_m (T_{m+1} - T_m) = T\))
\[
= \tilde{O}(\sqrt{MLKT}),
\]

which concludes the proof.

\[\Box\]

E.4 Proof of Theorem 24

We first state Lemmas 39 and 40, then go to the proof of Theorem 24. A direct consequence of Lemma 40 is that \(M_d := |T_d| = \tilde{O}(T^{1-2d/3})\), which we use in the proof of Theorem 24.

Lemma 39 With probability at least \(1 - 2K/T\), for any \(N < |W(t)|, t \in [T], i \in [K]\) and \(s, s' \in W(t)\)
\[
|\mu_{i,s} - \mu_{i,s'}| \leq C^{\text{gt}} \left( 3bDN + \tilde{O} \left( \sqrt{1/N} \right) \right)
\]
holds.
Lemma 39 is a version of Lemma 27 for bandit setting. Although the proof is very similar to Lemma 27, we demonstrate it below for the sake of completeness.

**Proof** [Proof of Lemma 39] We first consider the case $|W| = CDN$ for some integer $C > 1$. Let $i$ be the arm that is the target of monitoring consistency, which has been drawn at least once in every consecutive $D$ time steps. We decompose $W$ into $C$ subwindows of equal size $DN$ and let $W_c$ be the $c$-th subwindow for $c \in [C]$. For $c \in [C] \setminus \{1\}$, let $W_c$ be the joint subwindow of $W$ before $W_c$. Namely, $W_c = W_1 \cup W_2 \cup \cdots \cup W_{c-1}$. The fact that the window grows up to size $W$ without a breakpoint implies that each split $W_c \cup W_c$ satisfies

$$|\hat{\mu}_{i,W_c} - \hat{\mu}_{i,W_c}| \leq 2 \sqrt{\frac{\log(T^3)}{N}}.$$  \hfill (278)

Let $c \in [C]$ be arbitrary. We have

$$|\hat{\mu}_{i,W} - \hat{\mu}_{i,W_c}| = \left| \frac{1}{C} \hat{\mu}_{i,W} + \frac{C-1}{C} \hat{\mu}_{i,W_c} - \hat{\mu}_{i,W_c} \right|$$  \hfill (279)

$$\leq |\hat{\mu}_{i,W_c} - \hat{\mu}_{i,W_c}| + \frac{2}{C} \sqrt{\frac{\log(T^3)}{N}} \quad \text{(by Eq. (278))}$$  \hfill (280)

$$\leq |\hat{\mu}_{i,W_c} - \hat{\mu}_{i,W_c}| + \left( \frac{2}{C-1} + \frac{2}{C} \right) \sqrt{\frac{\log(T^3)}{N}}$$  \hfill (281)

$$\leq \left( 1 + \sum_{c'=c+1}^{C} \frac{2}{c'} \right) \sqrt{\frac{\log(T^3)}{N}} \quad \text{(by Eq. (278))}$$  \hfill (282)

$$\leq \log(T) \sqrt{\frac{\log(T^3)}{N}},$$  \hfill (283)

which implies for any $c, c' \in [C]$ we have

$$|\hat{\mu}_{i,W_c} - \hat{\mu}_{i,W_{c'}}| \leq 2 \log(T) \sqrt{\frac{\log(T^3)}{N}}.$$  \hfill (284)

By Eq. (106) we have

$$|\mu_{i,W_c} - \hat{\mu}_{i,W_c}| \leq \sqrt{\frac{\log(T^3)}{N}}$$  \hfill (285)

$$|\mu_{i,W_{c'}} - \hat{\mu}_{i,W_{c'}}| \leq \sqrt{\frac{\log(T^3)}{N}}$$  \hfill (286)

with probability at least $1 - 2K/T$. By the fact that $\mu_t$ moves $bDN$ within a subwindow of size $w$, for any $s \in W_c, s' \in W_{c'}$, we have

$$|\mu_{i,W_c} - \mu_{i,s}| \leq bN$$

60
Globally Nonstationary Bandits

\[ |\mu_{i,W_c} - \mu_{i,s'}| \leq bN. \]  

(288)

By using these, we have

\[ |\mu_{i,s} - \mu_{i,s'}| \leq |\mu_{i,W_c} - \mu_{i,s'}| + |\mu_{i,W_c} - \hat{\mu}_{i,W_c}| + |\mu_{i,W_c} - \hat{\mu}_{i,s'}| \]

(289)

\[ \leq 2bDN + 2(1 + \log(T))\sqrt{\frac{\log(T^3)}{N}}. \]  

(290)

(by Eq. (286), (287), (288))

The general case of \(|W| = CN + n\) for \(n \in \{0, 1, \ldots, N-1\}\) is easily proven by replacing \(2bN\) with \(3bN\) since \(\mu_i\) drift at most \(bN\) in \(n\) time steps.

Finally, by using Assumption 23, for any \(j \in [K]\) we have

\[ |\mu_{j,s} - \mu_{j,s'}| \leq C^{gr}\mu_{i,s} - \mu_{i,s'}| \leq C^{gr}\left(3bDN + 2(1 + \log(T))\sqrt{\frac{\log(T^3)}{N}}\right). \]  

(291)

\[ \blacksquare \]

**Lemma 40** Let

\[ \mathcal{X}(t) = \bigcup_{W_1, W_2 : W(t) = W_1 \cup W_2, j \in [K]} \left\{ |W_1| \leq C_1 b^{-2/3}, |W_2| \leq C_1 b^{-2/3}, |\hat{\mu}_{j,W_1} - \hat{\mu}_{j,W_2}| \geq 2\epsilon_{cut} \right\} \]  

(292)

where \(C_1 = \tilde{O}(1)\) is a factor that we define in Eq. (305). Let

\[ \mathcal{X} = \bigcup_{t \in [T]} \mathcal{X}(t). \]  

(293)

Then,

\[ \Pr[\mathcal{X}] \leq 2C_1 KT^{-1}. \]  

(294)

Lemma 40 is a version of Lemma 28 for the bandit setting. In the following, we derive Lemma 40 for completeness. The steps are very similar to the proof of Lemma 28. Under \(\mathcal{X}^c\), we have \(M_d = O(Tb^{-2/3}) = O(T^{1-2d/3})\).

**Proof** [Proof of Lemma 40] Let

\[ \mathcal{W}_{C_1} = \{W_0 \in W : |W_0| \leq C_1 b^{-2/3}\} \]  

(295)

be the set of windows of size at most \(C_1 b^{-2/3}\). It is easy to show that \(|\mathcal{W}_{C_1}| \leq TC_1 b^{-2/3}\).

By the union bound of the Hoeffding bound over all windows of \(\mathcal{W}_{C_1}\) and all arms, with probability at least

\[ 1 - \frac{2K}{T^{2+d}} \times TC_1 b^{-2/3} = 1 - 2C_1 KT^{-1}b^{1/3} \geq 1 - 2C_1 KT^{-1}, \]  

(296)
\[ \forall j \in [K] |\mu_{j,W_1} - \mu_{j,W_2}| \leq \sqrt{\frac{\log(2T^{2+d})}{2|W_{j,1}|}} \]

(297)

holds for all \( t \) and any split \( W_1 \cup W_2 = W(t) : |W_{j,1}|, |W_{j,2}| \leq C_1 b^{-2/3} \). Let \( N = C_1 b^{-2/3} \).

By definition of gradual stream,

\[ |\mu_{j,W_1} - \mu_{j,W_2}| \leq 2bN. \]  

(298)

Here,

\[ \sqrt{\frac{\log(T^3)}{2|W_{j,1}|}} + \sqrt{\frac{\log(T^3)}{2|W_{j,2}|}} \leq |\mu_{j,W_1} - \mu_{j,W_2}| \leq |\mu_{j,W_1} - \mu_{j,W_2}| + |\hat{\mu}_{j,W_1} - \mu_{j,W_1}| + |\hat{\mu}_{j,W_2} - \mu_{j,W_2}| \]  

(by triangular inequality)

(299)

(300)

(301)

\[ \leq 2bN + |\hat{\mu}_{j,W_1} - \mu_{j,W_1}| + |\hat{\mu}_{j,W_2} - \mu_{j,W_2}| \]  

(by (298))

(302)

(303)

which implies

\[ (\sqrt{3} - \sqrt{2 + d}) \left( \frac{\sqrt{\log T}}{|W_{j,1}|} + \frac{\sqrt{\log T}}{|W_{j,2}|} \right) \leq 2bN, \]  

(304)

which does not hold for

\[ N = b^{-2/3} \left( \frac{2 - \sqrt{3 + d}}{\sqrt{\log T}} \right)^{2/3}, |W_{j,1}|, |W_{j,2}| \leq N. \]  

(305)

Since \( |W_{j,1}| \leq |W_1|, |W_{j,2}| \leq |W_2| \), with probability \( 1 - 2C_1 KT^{-1} \), \( \mathcal{X}^c \) holds.

\[ \text{Proof} \quad [\text{Proof of Theorem 24}] \quad \text{Let} \; N \in \mathbb{N}^+ \; \text{be arbitrary. By Lemma 39, the event} \]

\[ \mathcal{Y} = \bigcap_{t \in [T]} \left\{ \epsilon(t) \leq C^\text{gr} \left( 3bDN + \tilde{O} \left( \sqrt{1/N} \right) \right) \right\} \]  

(307)

holds with probability at least \( 1 - 2K/T \). Let

\[ \mathcal{Z}_m = \left\{ \sum_{t=T_{d,m}+1}^{T_{d,m+1}} \text{reg}(t) \leq C^\text{dt} \left( \sqrt{LK(T_{d,m+1} - T_{d,m})} + L \sum_{t=1}^{S} \epsilon(t) \right) \right\}. \]  

(307)
for \( m = \{0, 1, 2, \ldots, M_d\} \), and

\[
Z = \bigcap_{m \in \{0, 1, 2, \ldots, M_d\}} Z_m.
\]  

(308)

By Definition 15, each of \( Z_m \) holds with probability at least \( 1 - \frac{4}{K/T^2} \), and thus \( Z \) holds with probability at least \( 1 - \frac{4}{K/T} \).

The regret is bounded as:

\[
\text{Reg}(T) = \sum_{t=T_{d,m}+1}^{T_{d,m+1}} \text{reg}(t)
\]

(309)

\[
\leq \sum_{m=0}^{M_d} C_{dt} \left( \sqrt{LKT(T_{d,m+1} - T_{d,m})} + \sum_{t} L \max_{i \in [K], s_1, s_2 \in W(t)} |\mu_{i,s_1} - \mu_{i,s_2}| \right) 1[\mathcal{X}, \mathcal{Y}, Z]
\]

+ \( T(1[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c]) \)

(310)

\[
\leq \sum_{m=0}^{M_d} C_{dt} \left( \sqrt{LKT(M_d + 1)} + \sum_{t} L \max_{i \in [K], s_1, s_2 \in W(t)} |\mu_{i,s_1} - \mu_{i,s_2}| \right) 1[\mathcal{X}, \mathcal{Y}, Z]
\]

+ \( T(1[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c]) \)

(311)

\[
\leq C_{dt} \left( O \left( \sqrt{LKT^2 - 2d/3} \right) + \sum_{t} LC_{gr} \left( 3bDN + \mathcal{O} \left( \sqrt{1/N} \right) \right) \right) 1[\mathcal{X}, \mathcal{Y}, Z]
\]

+ \( T(1[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c]) \)

(312)

(by Cauchy–Schwarz inequality)

\[
\leq C_{dt} \left( O \left( \sqrt{LKT^2 - 2d/3} \right) + \sum_{t} L \max_{i \in [K], s_1, s_2 \in W(t)} |\mu_{i,s_1} - \mu_{i,s_2}| \right) 1[\mathcal{X}, \mathcal{Y}, Z]
\]

+ \( T(1[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c]) \)

(313)

(by \( \mathcal{X}^c \))

\[
\leq C_{dt} \left( O \left( \sqrt{LKT^2 - 2d/3} \right) + \sum_{t} LC_{gr} \left( 3bDN + \mathcal{O} \left( \sqrt{1/N} \right) \right) \right) 1[\mathcal{X}, \mathcal{Y}, Z]
\]

+ \( T(1[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c]) \)

(314)

(by \( \mathcal{Y} \))

By taking the expectation, we have

\[
\mathbb{E}[\text{Reg}(T)]
\]

(315)

\[
\leq C_{dt} \left( O \left( \sqrt{LKT^2 - 2d/3} \right) + \sum_{t} LC_{gr} \left( 3bDN + \mathcal{O} \left( \sqrt{1/N} \right) \right) \right) + T(\Pr[\mathcal{X} \cup \mathcal{Y}^c \cup Z^c])
\]

(316)
\[
\leq C^{dt}\left( O\left( \sqrt{LKT^2-2d/3} \right) + \sum_t LC^{gr}\left( 3bDN + \tilde{O}\left( \sqrt{1/N} \right) \right) \right) + T \times O\left( \frac{K}{T} \right)
\] (317)

(by Lemma 40 and the fact that \( \mathcal{Y}, \mathcal{Z} \) hold with probability \( 1 - O(K/T) \))

\[
\leq C^{dt}\left( O\left( \sqrt{LKT^2-2d/3} \right) + \sum_t LC^{gr}\tilde{O}\left( (bD)^{1/3} \right) \right)
\] (318)

(by choosing \( N = O((bD)^{-2/3}) \))

\[
\leq \tilde{O}\left( \left( \sqrt{LK + LD^{1/3}} \right) T^{1-d/3} \right)
\] (319)

where we regard \( C^{dt}, C^{gr} \) as constants in the last transformation.