A Modularized Framework for Piecewise-Stationary Restless Bandits

Kuan-Ta Li
Institute of Communications Engineering
National Yang Ming Chaio Tung University
Hsinchu, Taiwan

Chia-Chun Lin
Institute of Communications Engineering
National Yang Ming Chaio Tung University
Hsinchu, Taiwan

Ping-Chun Hsieh
Department of Computer Science
National Yang Ming Chaio Tung University
Hsinchu, Taiwan

Yu-Chih Huang
Institute of Communications Engineering
National Yang Ming Chaio Tung University
Hsinchu, Taiwan

Abstract

We study the piecewise-stationary restless multi-armed bandit (PS-RMAB) problem, where each arm evolves as a Markov chain but *mean rewards may change across unknown segments*. To address the resulting exploration–detection delay trade-off, we propose a modular framework that integrates arbitrary RMAB base algorithms with change detection and a novel diminishing exploration mechanism. This design enables flexible plug-and-play use of existing solvers and detectors, while efficiently adapting to mean changes without prior knowledge of their number.

To evaluate performance, we introduce a refined regret notion that measures the *excess regret due to exploration and detection*, benchmarked against an oracle that restarts the base algorithm at the true change points. Under this metric, we prove a regret bound of $\tilde{O}(\sqrt{LMKT})$, where L denotes the maximum mixing time of the Markov chains across all arms and segments, M the number of segments, K the number of arms, and T the horizon. Simulations confirm that our framework achieves regret close to that of the segment oracle and consistently outperforms base solvers that do not incorporate any mechanism to handle environmental changes.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Restless multi-armed bandits (RMABs) provide a powerful abstraction for sequential decision-making under resource constraints. In contrast to the classical stochastic multi-armed bandit (MAB), where unselected arms remain frozen, RMABs model each arm as an evolving Markov chain that continues to transition regardless of whether it is activated. This restless property enables RMABs to capture a wide range of dynamic allocation problems where opportunities or risks evolve continuously, even in the absence of intervention.

Since the seminal work of Whittle (1988), RMABs have been studied extensively in both theory and practice. Whittle introduced the Whittle index, derived from a Lagrangian relaxation of the original optimization problem, which provides a computationally tractable priority rule under the indexability condition. Although finding the exact optimal policy is PSPACE-hard (Papadimitriou and Tsitsiklis, 1999), this line of work has inspired a rich literature on both theoretical guarantees and practical algorithms.

Beyond its theoretical significance, the RMAB framework has proven highly impactful in practice, demonstrating utility across multiple domains. In communication systems, RMAB formulations underpin opportunistic scheduling tasks such as dynamic spectrum access in cognitive radio networks and downlink scheduling in wireless networks (Liu and Zhao, 2008). The same framework has been applied in healthcare, where RMAB-based policies help optimize limited intervention resources (e.g., scheduling follow-up calls to patients) in public health outreach programs to maximize engagement and outcomes (Mate et al., 2022). Likewise, modern recommender systems leverage RMAB models to adapt content delivery as user preferences

evolve over time (Meshram et al., 2017). Another important application arises in maintaining information freshness, where scheduling to minimize the *Age of Information* (AoI) can be naturally cast as an RMAB problem (Hsu, 2018).

However, most existing studies assume stationary dynamics, which rarely hold in practice. Real-world environments often undergo gradual or abrupt changes, requiring policies that adapt over time. This motivates the study of RMABs in piecewise-stationary settings, where environments are segmented and the transition dynamics of arms shift at change points. These shifts are typically characterized by changes in the underlying Markov chains. To remain effective, the player's policy must continually adjust to the new segments.

Classical piecewise-stationary bandit problems employ two broad strategies for adaptation: active and passive methods. Active approaches couple standard bandit algorithms with change detection modules, which trigger resets upon detecting environmental shifts, making them well-suited for abrupt changes. In contrast, passive methods exploit mechanisms such as sliding windows or discount factors that gradually down-weight outdated samples, allowing the algorithm to adapt smoothly without explicit resets.

While piecewise-stationary bandits have been extensively studied, extending these ideas to RMABs introduces unique challenges. In RMABs, different states of an arm can yield varying reward means, complicating the distinction between routine state transitions and genuine environmental changes. This variability raises the risk of false alarms, where a detection module may mistakenly interpret state-based fluctuations as true shifts in the environment.

Building on this motivation, our work makes the following contributions:

Modular framework for PS-RMABs. We introduce a modular algorithmic framework that integrates three key components: a stationary base solver, a change detection module, and a diminishing exploration schedule. This design ensures flexibility and plug-and-play adaptability, while isolating the two core sources of cost in piecewise-stationary environments: forced exploration and detection delay.

Refined excess-regret notion. We formalize a new notion of *excess regret*, which explicitly captures the overhead caused by exploration and change detection, distinct from the intrinsic regret of the stationary base solver. Existing results for piecewise-stationary bandits show that the dominant terms in regret already arise from this exploration–delay balance, with the state-of-the-art scaling at $\tilde{O}(\sqrt{MKT})$. Our formulation makes this balance transparent and provides a principled way to analyze its extension to restless settings.

Diminishing exploration with minimal assumptions and low complexity. Our diminishing exploration (DE) requires no prior knowledge of the number of change points M, yet achieves a nearly optimal extra-regret bound of $\tilde{\mathcal{O}}(\sqrt{MKT})$. It adds virtually no computational overhead, relying only on simple scheduling to trigger exploration phases.

Theoretical guarantees and general-case analysis. We establish general excess-regret bounds that hold for arbitrary choices of base solvers and change detectors. These bounds confirm that the exploration–delay overhead is tightly controlled and does not exceed the known minimax rate in the piecewise-stationary setting. Consequently, whenever the stationary base solver is near-optimal, our framework preserves this near-optimality under piecewise-stationary extensions.

Unlike passive adaptation methods such as sliding-window or discounted bandits, which continuously down-weight past observations, our approach actively coordinates exploration and change detection through a diminishing exploration schedule. This mechanism ensures sufficient statistical evidence for reliable detection while avoiding excessive exploration, and crucially eliminates the need for prior knowledge of the number of change points. Moreover, our framework naturally extends these ideas to restless settings, where passive forgetting mechanisms alone may fail due to state-dependent reward fluctuations.

A natural question is whether one can simply combine existing change-detection schemes with learning algorithms designed for stationary restless bandits. However, naive integration may lead to unreliable detection, excessive resets, or degraded regret performance, particularly in restless environments where state evolution can obscure mean shifts. Our framework identifies conditions under which such integration remains stable and preserves regret guarantees.

1.1 Related Work

Restless Bandit Problem. Restless bandits demand a range of algorithmic approaches. In planning settings with known models, index-based policies (Whittle's and its variants) and LP-relaxation methods offer tractable solutions with guarantees like asymptotic optimality or constant-factor approximation (Whittle, 1988; Guha et al., 2010). In online learning settings, even under partial observability, the RMAB community has achieved sublinear regret. A notable early algorithm is Colored UCRL2 (Ortner et al., 2012), which adapts the UCRL2 framework for restless bandits and provides $\tilde{\mathcal{O}}(\sqrt{T})$ regret under mild mixing assumptions, nearly matching

the fundamental lower bound of $\Omega(\sqrt{ST})$ established in (Ortner et al., 2012), where S is the total number of states. More recent methods refine this paradigm using UCB-driven optimism—e.g., UCWhittle (Wang et al., 2023) and Restless-UCB (Wang et al., 2020)—to improve computational efficiency and adaptability. These developments supply both strong theoretical guarantees and broaden the practical applicability of RMAB solutions. These methods assume stationary environments and do not address change detection or adaptation to piecewise-stationary dynamics, which is the focus of our work.

Recent works have begun to address non-stationary RMABs under smoothly varying environments. For instance, Shisher et al. (2025) propose a sliding-window Whittle index policy that adapts to time-varying transition kernels under a bounded variation budget, achieving sublinear dynamic regret. Similarly, Hung et al. (2025) develop an optimistic Whittle-based algorithm with arm-specific sliding windows and establish regret guarantees under a global variation budget constraint. These approaches model non-stationarity through gradual drift in transition dynamics and rely on continuous adaptation mechanisms such as sliding windows.

Another line of work considers more general non-stationarity under adversarial environments. For example, Xiong and Li (2024) study adversarial RMABs with unknown transitions and bandit feedback, where rewards can change arbitrarily across episodes, and develop reinforcement learning algorithms with $\tilde{O}(\sqrt{T})$ regret guarantees. These settings capture worst-case non-stationarity and require fundamentally different techniques based on online learning and occupancy measures.

In contrast, our work focuses on piecewise-stationary environments with abrupt changes, where the main challenge lies in detecting change points and coordinating exploration with detection. Compared to variation-budget settings that assume gradual drift, and adversarial settings that allow arbitrary changes, our formulation imposes a structured non-stationarity model that enables sharper regret guarantees while still capturing realistic abrupt changes.

Piecewise-Stationary Bandits. Existing approaches to PS-MAB can be broadly categorized into *passive* and *active* methods. Passive methods adapt to changes through forgetting mechanisms such as Discounted UCB (Kocsis and Szepesvári, 2006), Sliding-Window UCB (Garivier and Moulines, 2011), and their Thompson Sampling counterparts (Raj and Kalyani, 2017; Qi et al., 2023), all achieving $\mathcal{O}(\sqrt{KMT}\,\mathrm{polylog}(T))$ regret. Active methods, in contrast, integrate change detection with standard bandit algorithms. Examples include CD-UCB (Liu et al., 2018) and M-UCB (Cao et al., 2019), which combine UCB with detectors like CUSUM or simple window-based tests to obtain $\mathcal{O}(\sqrt{KMT\log T})$ regret. More recent efforts, such as AdSwitch (Auer et al., 2019) and GLR-klUCB (Besson et al., 2022), avoid assuming prior knowledge of the number of changes M, advancing practical applicability. Among active methods, complexity is largely dictated by the change detector: M-UCB (Cao et al., 2019) uses a lightweight window-based test, while GLR-klUCB adopts a more statistically refined GLR detector, incurring moderately higher cost but achieving nearly optimal guarantees and strong empirical performance. AdSwitch, in contrast, features a substantially more complex detection scheme with elegant theory but no empirical validation. For example, Manegueu et al. (2021) design detectors based on empirical gaps between arms; Suk and Kpotufe (2022) quantify significant shifts at each step to avoid reliance on prior non-stationarity knowledge; and Abbasi-Yadkori et al. (2023) achieve comparable guarantees under abrupt changes. Complementary to these, multi-scale approaches such as Wei and Luo (2021) maintain multiple competing instances of the base algorithm to strengthen robustness. Although these works do not explicitly target the restless bandit problem, their designs—typically involving change detection or adaptive re-weighting—suggest potential applicability. However, their theoretical analyses usually rely on specific assumptions about the base algorithm (e.g., optimism or structural properties), which narrows the range of admissible base solvers that can be used while still guaranteeing strong regret bounds.

Our framework differs from these works in two key aspects: (i) it introduces a diminishing exploration mechanism to systematically supply evidence for detection without prior knowledge of change frequency, and (ii) it extends these ideas to restless bandits, where state evolution can confound detection.

2 PROBLEM FORMULATION

We study the piecewise-stationary RMAB (PS-RMAB) problem. The system consists of a single learner interacting with K arms. Each arm $k \in \{1, \ldots, K\}$ is modeled as a finite-state Markov chain $\mathcal{M}_k = (\mathcal{S}, P_k, R_k)$, where $\mathcal{S} = \{1, \ldots, S\}$ is the state space, $P_k : \mathcal{S} \times \mathcal{S} \to [0, 1]$ is the transition kernel, and $R_k : \mathcal{S} \to [0, 1]$ is the reward distribution.

At each round $t = 1, 2, \ldots, T$, the learner selects an arm A_t, observes its current state $s_{A_t,t}$, and receives a random reward $R_{A_t}(s_{A_t,t}) \in [0, 1]$. The expectation of this reward is denoted by $\mu_{A_t,s_{A_t,t}} = \mathbb{E}[R_{A_t}(s_{A_t,t})]$, which we call the *per-state mean reward*.

We assume the environment is *piecewise-stationary*.

There exist change points $0 = \nu_0 < \nu_1 < \cdots < \nu_M = T$ such that, within each segment $i \in \{1, \ldots, M\}$, arm k is characterized by a Markov chain $\mathcal{M}_k^{(i)} = (P_k^{(i)}, R_k^{(i)})$ with per-state mean rewards $\mu_{k,s}^{(i)}$.

The object of interest is the *arm-mean* vector $\bar{\boldsymbol{\mu}}^{(i)} = (\bar{\mu}_1^{(i)}, \ldots, \bar{\mu}_K^{(i)})$, which represents the long-run average rewards of all arms in segment i. This vector remains fixed within each segment $(\nu_{i-1}, \nu_i]$ but may change across consecutive segments (i.e., $\bar{\boldsymbol{\mu}}^{(i)} \neq \bar{\boldsymbol{\mu}}^{(i+1)}$ for at least one arm). Any modification that alters an arm mean is regarded as a change point. Figure 1 illustrates the instant reward of one of the arms in a PS-RMAB environment.

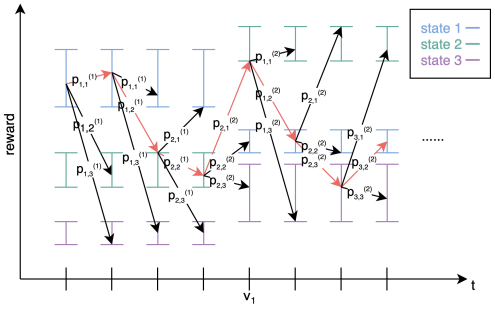


Figure 1: Instant reward of Arm k in PS-RMAB environment

Fix a stationary RMAB base solver $\mathcal{B}$. The segment-wise oracle $\pi^{\text{seg}}(\mathcal{B})$ knows the true change points $\nu_1, \ldots, \nu_{M-1}$ and simply restarts $\mathcal{B}$ at the beginning of each segment (all other details—model class, observability, and action constraints—are identical to those faced by the learner). Let r_t^{π} denote the expected reward at time t under the learner's policy and r_t^{seg} the counterpart under $\pi^{\text{seg}}(\mathcal{B})$

We define the excess regret of a policy π (relative to $\pi^{\text{seg}}(\mathcal{B})$) as

$$\mathcal{R}_{\text{excess}}(T) \;=\; \mathbb{E}\Big[\sum_{t=1}^{T} r_t^{\text{seg}}\Big] \;-\; \mathbb{E}\Big[\sum_{t=1}^{T} r_t^{\pi}\Big]. \qquad (1)$$

This metric depends only on the overhead induced by exploration and change detection (and any reset misalignment), while factoring out the stationary performance of the chosen base solver.

3 PROPOSED FRAMEWORK: DIMINISHING EXPLORATION

Our framework builds upon the active method paradigm (Yu and Mannor, 2009; Liu et al., 2018; Cao et al., 2019), which combines a stationary bandit algorithm with a change-detection module. Whenever a change is detected, the base algorithm is restarted in the new segment. To ensure sufficient environmental sampling for change detection, we further integrate a novel exploration scheme, *diminishing exploration*, which dynamically balances the trade-off between exploration cost and detection delay. This modular design allows arbitrary stationary solvers and detectors to be flexibly combined without altering their internal structures.

3.1 Diminishing Exploration

Most existing works adopt a uniform exploration scheme that allocates a constant fraction of time to exploration (Liu et al., 2018; Cao et al., 2019). While effective for detecting changes, the approach results in regret that scales linearly with the exploration rate and therefore requires prior knowledge of the number of change points M in order to appropriately set this rate. To address this limitation, we introduce *diminishing exploration*, which adaptively reduces the frequency of forced exploration over time, thereby eliminating the need for prior knowledge of M.

Let us define τ_i as the i-th time when the algorithm alarms a change. In the proposed method, a uniform exploration round starts at $u_{i-1}^{(j)}$ for $j \in \{1, 2, \ldots\}$ with $u_{i-1}^{(1)} = \tau_{i-1} + 1$. i.e., the learner chooses to pull the arm $1, 2, \ldots, K$ at time $u_{i-1}^{(j)}, u_{i-1}^{(j)} + 1, \ldots, u_{i-1}^{(j)} + K - 1$, respectively. The process restarts whenever a new change τ_i is detected.

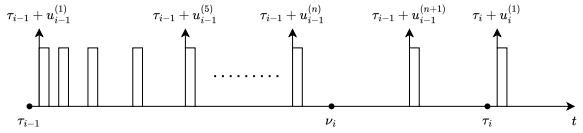


Figure 2: Diminishing exploration.

We aim to balance the regret resulting from exploration and that associated with the performance of change detection by dynamically adjusting the exploration rate *within a segment*. Let $u_{i-1}^{(j)}$ be the start time of the j-th uniform exploration session between two consecutive alarms τ_i and τ_{i-1}. In our approach, these sessions are designed in such a way that $u_{i-1}^{(j+1)} - u_{i-1}^{(j)}$ is greater than $u_{i-1}^{(j)} - u_{i-1}^{(j-1)}$. This means that the inter-session time within the same time segment increases with j, which in turn results in a reduction in the exploration rate. Specifically, for the i-th segment, we choose $u_i^{(1)} = \left\lceil (\alpha - K/4\alpha)^2 \right\rceil$ and $u_i^{(j)} = \left\lceil u_i^{(j-1)} + \frac{K}{\alpha}\sqrt{u_i^{(j-1)}} + \frac{K^2}{4\alpha^2} \right\rceil$, $\forall\ 1 \le i \le M,\ j \ge 2$, without the knowledge of M and the parameter α will

be chosen later. Clearly, we have $u_i^{(j-1)} + K < u_i^{(j)}$ for every $j \geq 2$; thus, these exploration phases will not overlap. Moreover, it is obvious that the duration between two exploration phases $u_i^{(j)} - u_i^{(j-1)} = \mathcal{O}(\sqrt{u_i^{(j-1)}})$ increases with time as Figure 2; hence, the exploration rate decreases. Thus, we term this mechanism *diminishing exploration*.

3.2 Integration of Modular Components

The diminishing exploration scheme is orthogonal to the choice of change-detection algorithm. Any off-the-shelf detector (e.g., CUSUM, GLR, or window-based tests) can be used. Our framework therefore consists of three modular components: (i) A **base algorithm** $\mathcal{B}$ for stationary RMABs (e.g., RestlessUCB, Colored-UCRL2). (ii) A **change detector** $\mathcal{D}$ that raises alarms upon detecting arm-mean changes. (iii) The **diminishing exploration** scheme $\mathcal{E}$ that schedules exploration rounds to feed $\mathcal{D}$.

Whenever $\mathcal{D}$ signals a change, the base algorithm $\mathcal{B}$ is restarted from scratch. The modular design ensures forward compatibility, allowing future advancements in any of the three components (e.g., stronger detectors or more efficient base solvers) to be seamlessly incorporated into the framework.

3.3 Main Algorithm

The proposed framework is described in Algorithm 1, which interleaves diminishing exploration (lines 3–7) with actions taken by a stationary RMAB base solver $\mathcal{B}$ (line 8). $\mathcal{H}_{t-1}$ denotes the available history for arm selection, which can be either base-only ($\mathcal{H}_{t-1}^{\mathcal{B}}$), where the base algorithm relies solely on its own interaction history, or shared ($\mathcal{H}_{t-1}^{\mathcal{B}}, \mathcal{H}_{t-1}^{\mathcal{E}}$), where it additionally incorporates information gathered during the exploration phase (line 9). After each step, observed rewards are passed to the change detector $\mathcal{D}$ (lines 12–14). If $\mathcal{D}$ raises an alarm, the algorithm resets both the exploration schedule and the base solver (lines 15–16).

Although Algorithm 1 is presented in a modular form, our framework is compatible with a wide range of instantiations. For example, $\mathcal{B}$ can be any stationary RMAB solver (e.g., RestlessUCB (Wang et al., 2020), Colored-UCRL2 (Ortner et al., 2012)), while $\mathcal{D}$ can be instantiated by standard change-detection methods such as CUSUM (Liu et al., 2018), GLR (Besson et al., 2022), or window-based tests (Cao et al., 2019). This design highlights the plug-and-play nature of our approach: diminishing exploration can be seamlessly integrated with any $\mathcal{B}$ and $\mathcal{D}$.

Algorithm 1 CD-RMAB with Diminishing Exploration (Modular)

Require: Horizon T, number of arms K, exploration parameter α;
Base solver $\mathcal{B}$ (for stationary RMABs);
Change Detector $\mathcal{D}$ (operates on arm-level reward streams)

1: Initialize segment start $\tau \leftarrow 0$, exploration cursor $u \leftarrow \left\lceil \left(\alpha - \frac{K}{4\alpha}\right)^2 \right\rceil$, and per-arm counters $n_k \leftarrow 0$ for all $k \in \mathcal{K}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: **if** $u \leq t - \tau < u + K$ **then**
4: $A_t \leftarrow (t - \tau) - u + 1$
5: **else**
6: **if** $t - \tau = u + K$ **then**
7: $u \leftarrow \left\lceil u + \frac{K}{\alpha}\sqrt{u} + \frac{K^2}{4\alpha^2} \right\rceil$
8: **end if**
9: $A_t \leftarrow \mathcal{B}.\textsc{Select}(t, \mathcal{H}_{t-1})$
10: **end if**
11: Pull arm A_t; observe current state $s_{A_t,t}$ and reward $r_t \in [0, 1]$
12: $\mathcal{B}.\textsc{Update}(A_t, s_{A_t,t}, r_t)$
13: $n_{A_t} \leftarrow n_{A_t} + 1$; $Z_{A_t,n_{A_t}} \leftarrow r_t$
14: **if** $\mathcal{D}.\textsc{Alarm}(\{Z_{k,\cdot}\}, t) = \textsc{True}$ **then**
15: $\tau \leftarrow t$; $u \leftarrow 1$; $n_k \leftarrow 0$ for all $k \in \mathcal{K}$
16: $\mathcal{B}.\textsc{Reset}()$
17: **end if**
18: **end for**

4 REGRET ANALYSIS

Since the framework allows flexible choices of base solvers $\mathcal{B}$ and change detectors $\mathcal{D}$ at the algorithmic level, our first main result (Theorem 4.1) establishes a general excess regret bound that applies to any such combination. Our analysis treats the base solver as a black box and does not rely on structural properties specific to particular algorithms (e.g., indexability, monotonicity, or other problem-dependent assumptions). To obtain concrete regret guarantees, we then instantiate the framework with specific choices of base solvers and change detectors that satisfy the required conditions. This separation highlights the generality of our framework while ensuring that the final regret bounds remain valid.

To establish an upper bound on the regret, we first introduce the necessary notations and event definitions that characterize the behavior of the change detection module. These events allow us to disentangle the errors introduced by false alarms, detection delay, and the inherent stochasticity of the Markovian reward processes.

Let τ_i denote the time at which the i-th change point is

detected by the change detector, where $0 \le i \le M-1$. We define the set of *false alarm events* as $F_i := \{\tau_i < \nu_i\}$ for $1 \le i \le M-1$, and $F_0 := \{\tau_0 = 0\}$. That is, F_i indicates that the i-th declared change occurs strictly before the true change point ν_i. Similarly, we define the set of *successful detection events* as $D_i := \{\nu_i \le \tau_i \le \nu_i + h_i\}$ for $1 \le i \le M-2$, where h_i is a detection delay parameter determined by the underlying change detector. For completeness, we also set $D_0 := \{\tau_0 = 0\}$ and $D_{M-1} := \{\tau_{M-1} \le T\}$.

In addition, for the regret decomposition we define two gap quantities. For each arm k and segment i, $\Delta_k^{(i)} := \max_{k' \in \mathcal{K}}(\bar{\mu}_{k'}^{(i)} - \bar{\mu}_k^{(i)})$ and $\delta_k^{(i)} := |\bar{\mu}_k^{(i+1)} - \bar{\mu}_k^{(i)}|$. Finally, we define $\delta^{(i)} := \max_{k \in \mathcal{K}} \delta_k^{(i)}$ as the maximum mean shift across arms at change point ν_i.

Recall the excess regret $\mathcal{R}_{\text{excess}}(T) := \mathbb{E}\big[\sum_{t=1}^T r_t^{\text{seg}}\big] - \mathbb{E}\big[\sum_{t=1}^T r_t^{\pi}\big]$, where r_t^{seg} corresponds to the segment–oracle that restarts the same base solver at true change points. We decompose $\mathcal{R}_{\text{excess}}(T)$ into an *exploration cost* and a *change–detection cost*. Throughout the analysis, we focus on the *fully modular* setting, where arm selection relies solely on the base algorithm's history $\mathcal{H}_{t-1}^{\mathcal{B}}$. Incorporating $\mathcal{H}_{t-1}^{\mathcal{E}}$ would make the analysis dependent on the specific base algorithm.

Theorem 4.1 (Extra–regret bound)**.** *For any base solver $\mathcal{B}$ and change detector $\mathcal{D}$, the excess regret of Algorithm 1 satisfies*

$$\mathcal{R}_{\text{excess}}(T) \le \underbrace{2\alpha\sqrt{MT}}_{(a)} + \underbrace{\sum_{i=1}^{M-1} \mathbb{E}\left[\tau_i - \nu_i \middle| D_i \overline{F}_i D_{i-1} \overline{F}_{i-1}\right]}_{(b)} + \underbrace{T\sum_{i=1}^{M} \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) + T\sum_{i=1}^{M-1} \mathbb{P}\left(\overline{D}_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)}_{(c)}, \tag{2}$$

Proof. We rewrite the excess regret as the difference between the total regret of our policy and that of the segment oracle, i.e. $\mathcal{R}_{\text{excess}}(T) = (\mathbb{E}[\sum_{t=1}^T r_t^\star] - \mathbb{E}[\sum_{t=1}^T r_t^\pi]) - (\mathbb{E}[\sum_{t=1}^T r_t^\star] - \mathbb{E}[\sum_{t=1}^T r_t^{\text{seg}}])$. The first term corresponds to the regret of running the base algorithm within each segment, i.e., the sum of the base solver's regrets across all M segments. For the second term, we expand it recursively across change points. In the ideal case of correct detections, this adds the detection delays together with the sum of the base solver's regrets over all M segments. If a false alarm or missed detection occurs, we treat these as worst-case events and charge them up to T regret each. Thus, once a concrete change detector is chosen, the analysis ensures that the probabilities of false alarms and missed detections are sufficiently small, leaving the main contributions from the forced exploration and the detection delay. Finally, since the base solver's regret cancels out between the two terms, we arrive exactly at the decomposition stated in Theorem 4.1. The detailed proof is deferred to Supplementary Materials. □

The bound in Theorem 4.1 is entirely independent of the choice of base solver $\mathcal{B}$. Indeed, the *excess regret* arises solely from two sources: (i) the forced exploration introduced by the diminishing exploration schedule (Algorithm 1), contributing term (a), and (ii) the behavior of the change detection module $\mathcal{D}$, which governs the detection delay and false-alarm probabilities, contributing terms (b)–(c).

4.1 One-State Special Case

In this subsection, we consider the one-state setting, which reduces the PS-RMAB to the classical piecewise-stationary bandit problem. We first present a general upper bound on the excess regret that does not assume any particular base solver or change detector. We then instantiate this bound with specific algorithms to obtain concrete order guarantees. To make the general extra-regret bound in Theorem 4.1 concrete, we now instantiate both the base solver and the change detector.

Base Solver. For the one-state case, the base solver $\mathcal{B}$ can be chosen as the classical UCB algorithm, which is known to achieve near-optimal regret in stationary MABs. Since the one-state PS-RMAB reduces to the piecewise-stationary MAB setting, this choice is natural and ensures that the base solver itself does not inflate the overall regret order.

Change Detector. We instantiate the CD module with the M-UCB Change Detector (Cao et al., 2019), a sliding-window mean-shift test that raises an alarm whenever the empirical means of two consecutive halves of a window differ by more than a threshold. Its parameters (w, b) are chosen according to

$$w = \frac{4}{\delta^2}\left(\sqrt{\log(2KT^2)} + \sqrt{\log(2T)}\right)^2, \tag{3}$$

$$b = \sqrt{w \log(2KT^2/2)}. \tag{4}$$

where δ is a known lower bound on the magnitude of mean shifts. The full pseudocode is in Appendix A.

Remark 4.2. Assumptions such as the existence of a known gap parameter δ (Assumption A.1 in Cao et al. (2019)) and minimum segment lengths (Assumption A.2) are required specifically for the M-UCB detector to guarantee its statistical properties. These are *not* intrinsic to our framework: diminishing exploration itself does not impose additional requirements. This distinction highlights the modularity of our approach,

which allows the plug-and-play use of different CD algorithms under their respective conditions.

Combining the base solver $\mathcal{B}$ with the M-UCB Change Detector, we obtain the following corollary of Theorem 4.1.

Corollary 4.3. *Combining Algorithm 1 (diminishing exploration with resets) with the base solver $\mathcal{B}$ and Algorithm 2 with parameters (w, b) given in Equations 3 and 4, the excess regret is upper-bounded as*

$$\mathcal{R}_{\text{excess}}(T) \;\le\; \mathcal{O}\Big(\sqrt{KMT\log T}\Big). \tag{5}$$

Proof Outline of Corollary 4.3. We decompose the excess regret into three sources: (i) the cost of diminishing exploration within stationary segments, (ii) the overhead caused by false alarms, and (iii) the delay incurred in detecting true changes. The following lemmas bound each component. Finally, combining these bounds with Theorem 4.1 yields the desired order result.

In what follows, the first lemma establishes a bound on the regret incurred during the diminishing exploration phase of the algorithm. We denote by $R_{\text{DE}}(\tau_{i-1}, \nu_i)$ the regret accumulated due to exploration between the previous alarm time τ_{i-1} and the next change point ν_i, and by $N_{\text{DE},k}(\tau_{i-1}, \nu_i)$ the number of times arm k is selected during this exploration phase.

Lemma 4.4 (Diminishing exploration regret)**.** *If the mean values of the arms remain the same during the time interval $[\tau_{i-1}, \nu_i)$, then we have*

$$N_{\text{DE},k}\left(\tau_{i-1}, \nu_i\right) \le \frac{2\alpha\sqrt{\nu_i - \tau_{i-1}}}{K} + \frac{3}{2}, \tag{6}$$

and

$$\mathbb{E}\left[R_{\text{DE}}\left(\tau_{i-1}, \nu_i\right)\right] \le 2\alpha\sqrt{\nu_i - \tau_{i-1}} + \frac{3}{2}K. \tag{7}$$

This lemma shows that, on any stationary interval, the regret incurred purely from diminishing exploration is controlled by a term proportional to the square root of the interval length. Summing over all $M{+}1$ stationary segments and applying Cauchy–Schwarz, we obtain a total contribution of order $\sqrt{MT}$ up to logarithmic factors.

Lemma 4.5 (Probability of false alarm)**.** *Under Algorithm 1 with parameter in Equations 3 and 4, we have*

$$\mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) \le wK\left(1 - \left(1 - \exp\left(-2b^2/w\right)\right)^{\lfloor T/w \rfloor}\right) \le \frac{1}{T}. \tag{8}$$

The above ensures that false alarms occur with vanishing probability. Therefore, the expected regret overhead due to spurious resets is negligible, contributing at most $\mathcal{O}(1)$ in expectation.

Lemma 4.6 (Probability of successful detection)**.** *Consider a piecewise-stationary bandit environment. For any $\boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(i+1)} \in [0,1]^K$ with parameters chosen in equation 3 and equation 4 and*

$$h_i = \left\lceil w\left(\frac{K}{2\alpha} + 1\right)\sqrt{s_i + 1} + \frac{w^2}{4}\left(\frac{K}{2\alpha} + 1\right)^2 \right\rceil, \tag{9}$$

for some $k \in \mathcal{K}, i \ge 1$ and $c > 0$, under the Algorithm 1, we have

$$\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \ge 1 - \frac{1}{T}. \tag{10}$$

This lemma guarantees that each genuine change is detected with high probability, which prevents the algorithm from staying too long in a mismatched segment.

Lemma 4.7 (Expected detection delay)**.** *Consider a piecewise-stationary bandit environment. For any $\boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(i+1)} \in [0,1]^K$ with parameters chosen in equation 3 and equation 4 and*

$$h_i = \left\lceil w\left(\frac{K}{2\alpha} + 1\right)\sqrt{s_i + 1} + \frac{w^2}{4}\left(\frac{K}{2\alpha} + 1\right)^2 \right\rceil, \tag{11}$$

for some $K \in \mathcal{K}, i \ge 1$ and $c > 0$, under the Algorithm 1, we have

$$\mathbb{E}\left[\tau_i - \nu_i \middle| \overline{F}_i D_i \overline{F}_{i-1} D_{i-1}\right] \le h_i. \tag{12}$$

In other words, once a change occurs, the delay before detection is bounded in expectation by h_i, which scales sublinearly with s_i under our parameter choice. This ensures that the regret accumulated during delays is well controlled.

Putting everything together. The regret from diminishing exploration (Lemma 4.4) is $\mathcal{O}(\sqrt{KMT})$; the false alarm contribution (Lemma 4.5) is negligible; and the regret during detection delays (Lemmas 4.6–4.7) is bounded by $\tilde{\mathcal{O}}(\sqrt{KMT})$. Substituting these into the general bound in Theorem 4.1, we obtain

$$\mathcal{R}_{\text{excess}}(T) \le \mathcal{O}\Big(\sqrt{KMT\log T}\Big), \tag{13}$$

which concludes the proof sketch.

In the one-state setting, PS-RMAB reduces to the classical piecewise-stationary MAB. In addition to the extra-regret perspective, our framework naturally extends to a standard regret analysis against a clairvoyant oracle that always pulls the best arm in each segment.

If the base solver $\mathcal{B}$ is chosen as UCB, then under bounded rewards and a change detector with controlled false-alarm probability and expected delay, the total regret of Algorithm 1 remains near-optimal, scaling as $\tilde{\mathcal{O}}(\sqrt{KMT})$. The detailed proof is deferred to Supplementary Materials.

4.2 General Case

Compared to the one-state case, the general PS-RMAB setting introduces additional challenges due to the temporal dependence of rewards induced by the Markovian dynamics. To handle this dependence, our analysis leverages the mixing time of each arm's Markov chain, which allows us to approximate independence and apply concentration inequalities. To make this argument rigorous, we require a technical assumption ensuring that each arm's Markov chain admits well-defined steady-state behavior. In particular, ergodicity guarantees the existence of unique stationary distributions and enables the use of concentration tools in our regret analysis.

Assumption 4.8. All Markov chains $\mathcal{M}_k^{(i)} = (P_k^{(i)}, R_k^{(i)})$ are ergodic. This ensures the existence of a unique steady-state distribution $d_k^{(i)}$ for each arm k in segment i, satisfying $d_k^{(i)} = d_k^{(i)} P_k^{(i)}$ with $\sum_{s\in\mathcal{S}} d_k^{(i)}(s) = 1$. We then define the steady-state arm mean as $\bar{\mu}_k^{(i)} := d_k^{(i)} R_k^{(i)}$.

This assumption is necessary because the reward sequence of each arm arises from a Markov chain, which induces temporal dependence. Ergodicity allows us to approximate independence through the mixing properties of the chains and to apply concentration inequalities (e.g., Hoeffding-type bounds for Markov processes). Without this assumption, the notion of a stationary arm mean would not be well defined, and regret guarantees could not be rigorously established.

To accommodate this dependence within the change detection module, we refine the choice of its parameters. Specifically, the window size w and threshold b are tuned in accordance with the mixing-time bound, ensuring that the detector achieves the required statistical guarantees while remaining compatible with the diminishing exploration schedule. Formally, we select

$$w \;=\; \frac{4}{\delta_{\min}^2}\Big(\sqrt{2\ln(2KT^2)} \;+\; \sqrt{144\,L\,\ln(2KT^2)} \;+\; \sqrt{144\,L\,\ln(2T)}\Big)^2, \quad (14)$$

$$b \;=\; \sqrt{\frac{w}{2}\,\ln(2KT^2)} \;+\; \sqrt{144\,w\,L\,\ln(2KT^2)}. \quad (15)$$

where $\delta_{\min}$ is a known lower bound on the mean shift, and $L = \max_{i\in\{1,\dots,M\}} \max_{k\in\mathcal{K}} L_k^{(i)}$ denotes the maximum mixing time of the Markov chains associated with all arms and segments.

Corollary 4.9. *Combining Algorithm 1 (diminishing exploration with resets) with the base solver $\mathcal{B}$ and Algorithm 2 with parameters (w, b) given in Equations 14 and 15, the excess regret is upper-bounded as*

$$\mathcal{R}_{\text{excess}}(T) \;\le\; \mathcal{O}\Big(\sqrt{MKLT\log T}\Big), \quad (16)$$

where $L = \max_{i\in\{1,\dots,M\}} \max_{k\in\mathcal{K}} L_k^{(i)}$ denotes the maximum mixing time of the Markov chains associated with all arms and segments.

Proof Outline of Corollary 4.9. The overall proof follows the same decomposition as in Corollary 4.3, namely by bounding the regret incurred from (i) diminishing exploration within stationary segments, (ii) regret caused by false alarms, and (iii) regret due to delays in detecting true changes. The bounds from Lemma 4.4 apply directly here as well.

For the false alarm probability, we now invoke Lemma 4.5, which states that under Algorithm 2 with parameters chosen according to Equations 14 and 15, the probability of a false alarm in each stationary period is at most

$$\mathbb{P}\big(F_i \,\big|\, \overline{F}_{i-1} D_{i-1}\big) \;\le\; \frac{\epsilon\|\varphi\|_d + 1}{T}. \quad (17)$$

This again ensures that the contribution of false alarms to the overall regret is negligible in expectation.

For the detection probability and delay, Lemma 4.6 shows that for each change, with the choice of parameters w, b as in Equations 14 and 15 and the delay threshold h_i,

$$\mathbb{P}\big(\overline{D}_i \,\big|\, \overline{F}_{i-1} D_{i-1}\big) \;\le\; \frac{\epsilon\|\varphi\|_d}{T}, \quad (18)$$

so that missed detections occur with vanishing probability. Moreover, the expected detection delay is bounded by h_i, which scales sublinearly with the segment length. These results play the same role as Lemmas 4.6 and 4.7 in the one-state case. Here, $\epsilon > 0$ is an arbitrarily small constant, and φ denotes the initial state distribution of the Markov chains.

Therefore, the regret contributions from diminishing exploration, false alarms, and detection delays remain of the same order as before, and substituting these bounds into Theorem 4.1 yields

$$\mathcal{R}_{\text{excess}}(T) \le \mathcal{O}\Big(\sqrt{KMLT\log T}\Big), \quad (19)$$

which establishes Corollary 4.9.

From Theorem 4.1, the additional cost introduced by diminishing exploration and change detection, captured by $\mathcal{R}_{\text{excess}}$, is strictly controlled and does not alter the order of regret. To argue near-optimality, we turn to the *one-state case*, which reduces to the classical PS-MAB setting. Here, our framework combined with UCB attains the known $\tilde{\mathcal{O}}(\sqrt{KMT})$ bound, confirming that the overhead is indeed compatible with near-optimal performance.

The regret bound depends on problem-specific parameters such as the mixing time of the underlying Markov

chains, reward gaps, and detector parameters. These quantities influence constant factors but do not alter the asymptotic rate. In particular, larger mixing times or smaller mean shifts may increase detection difficulty, leading to longer delays, yet the overall regret order remains unchanged.

Motivated by this evidence, we extend the conclusion to the general PS-RMAB setting:

Principle 4.10 (Transfer Principle)**.** *If the base solver $\mathcal{B}$ achieves near-optimal regret in the stationary RMAB setting, then Algorithm 1 equipped with $\mathcal{B}$ and a suitable change detector $\mathcal{D}$ inherits this near-optimality in the PS-RMAB setting.*

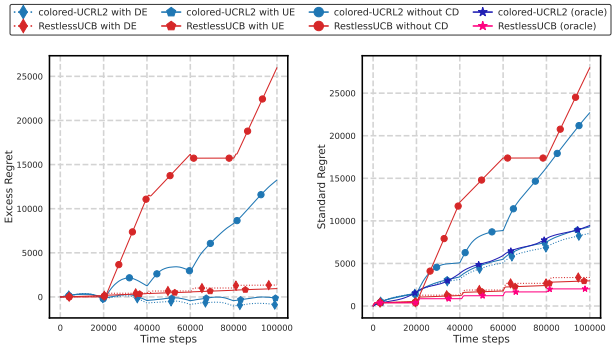
Proof. Since the stationary RMAB already admits a minimax lower bound of $\Omega(\sqrt{ST})$ (Ortner et al., 2012), and our framework in the piecewise-stationary case introduces only an additional $\tilde{\mathcal{O}}(\sqrt{MKT})$ overhead, the overall regret rate remains $\tilde{\mathcal{O}}(\sqrt{T})$. Thus, the near-optimality of stationary solvers is preserved under piecewise-stationary extensions. □

5 SIMULATION RESULTS

We evaluate our framework using `RestlessUCB` and `colorUCRL2` as base algorithms. For each case, we report both the *excess regret* defined in our framework and the standard regret benchmarked against a value-iteration oracle baseline. The simulations adopt the *partial modular* setting, where information is shared between the exploration module and the base algorithm, enabling more accurate and comprehensive estimation for arm selection; this may explain why, as shown in Figures 3a, the observed excess regret can occasionally be smaller than that of the segment oracle. The results are summarized in Figures 3a and 3b.

Environment Configuration. We simulate a PS-RMAB environment with a horizon of $T = 100{,}000$ time steps, partitioned into $M = 5$ stationary segments. The problem consists of $K = 3$ arms, each modeled as a Markov chain with $S = 3$ states. The stationary mean rewards of the arms are configured as $\bar{\mu}_k^{(i)} \in \{0.2, 0.5, 0.8\}$, assigned according to the rule $(i + k) \bmod 3 = 2, 0, 1$, respectively, for segment i and arm k.

Extra vs. Standard Regret. As shown in Figure 3a, the excess regret exhibits only minor variation across different choices of base algorithms. This indicates that the additional overhead induced by our mechanism is essentially independent of the base solver. Comparing with the standard regret in Figure 3b, we observe that the overhead introduced by our framework is negligible relative to the dominant regret incurred by the base algorithms themselves. In particular, the `UE` curve represents the optimal exploration amount achievable when the number of segments is known in advance, while `DE` corresponds to our proposed diminishing exploration scheme without prior knowledge of the number of changes.



(a) Excess Regret. (b) Standard Regret.

Figure 3

Significance of Change Detection. One may wonder: since the regret of the base algorithms already dominates, is it sufficient to simply let them adapt on their own, without any change-detection or exploration mechanism? To examine this, we also simulate the scenario where no change-detection module is used. As shown in Figures 3a and 3b, the regret exhibits a much sharper growth when the environment changes are ignored, indicating that the base algorithms alone struggle to adapt promptly to non-stationarity. This illustrates the potential performance degradation caused by neglecting environmental shifts.

6 CONCLUSION

We addressed the PS-RMAB problem by introducing a modular framework that integrates stationary solvers, change detection, and a novel diminishing exploration schedule. Central to our analysis is a refined notion of excess regret, which isolates the costs of exploration and detection delay. We proved that this overhead matches the minimax order known for piecewise-stationary bandits, and established a transfer principle showing that the near-optimality of stationary solvers carries over to the PS-RMAB setting. This provides both a unifying theoretical perspective and a practical recipe for extending stationary algorithms to dynamic environments.

ACKNOWLEDGMENTS

This research was supported by Wistron Corporation. It was also partially supported by the National Science and Technology Council (NSTC) of Taiwan under

Grant Numbers 113-2223-E-A49-005-MY3, 114-2628-E-A49-002, and 114-2634-F-A49-002-MBK.

References

Peter Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25:287 – 298, 1988.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999. ISSN 0364765X, 15265471. URL `http://www.jstor.org/stable/3690486`.

Keqin Liu and Qing Zhao. A restless bandit formulation of opportunistic access: Indexablity and index policy. In *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pages 1–5, 2008. doi: 10.1109/SAHCNW.2008.12.

Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:12017–12025, 06 2022. doi: 10.1609/aaai.v36i11.21460.

Rahul Meshram, Aditya Gopalan, and D. Manjunath. A hidden markov restless multi-armed bandit model for playout recommendation systems. In Nishanth Sastry and Sandip Chakraborty, editors, *Communication Systems and Networks*, pages 335–362, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67235-9.

Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018.

Sudipto Guha, Kamesh Munagala, and Peng Shi. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):1–50, 2010.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pages 214–228. Springer, 2012.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10131–10139, 2023.

Siwei Wang, Longbo Huang, and John Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33:11878–11889, 2020.

Md Kamran Chowdhury Shisher, Vishrant Tripathi, Mung Chiang, and Christopher G. Brinton. Online learning of whittle indices for restless bandits with non-stationary transition kernels, 2025. URL `https://arxiv.org/abs/2506.18186`.

Yu-Heng Hung, Ping-Chun Hsieh, and Kai Wang. Non-stationary restless multi-armed bandits with provable guarantee, 2025. URL `https://arxiv.org/abs/2508.10804`.

Guojun Xiong and Jian Li. Provably efficient reinforcement learning for adversarial restless multi-armed bandits with unknown transitions and bandit feedback, 2024. URL `https://arxiv.org/abs/2405.00950`.

Levente Kocsis and Csaba Szepesvári. Discounted UCB. In *2nd PASCAL Challenges Workshop*, volume 2, pages 51–134, 2006.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of International Conference on Algorithmic Learning Theory (ALT)*, pages 174–188, 2011.

Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A Bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

Han Qi, Yue Wang, and Li Zhu. Discounted thompson sampling for non-stationary bandit problems. *arXiv preprint arXiv:2305.10718*, 2023.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 418–427, 2019.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of Conference on Learning Theory (COLT)*, pages 138–158, 2019.

Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point

detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(1):3337–3376, 2022.

Anne Gael Manegueu, Alexandra Carpentier, and Yi Yu. Generalized non-stationary bandits. *arXiv preprint arXiv:2102.00725*, 2021.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits. In *Conference on Learning Theory*, pages 2160–2182. PMLR, 2022.

Yasin Abbasi-Yadkori, András György, and Nevena Lazić. A new look at dynamic regret for non-stationary stochastic bandits. *Journal of Machine Learning Research*, 24(288):1–37, 2023.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4300–4354. PMLR, 15–19 Aug 2021. URL `https://proceedings.mlr.press/v134/wei21b.html`.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1177–1184, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. *arXiv preprint arXiv:1201.0559*, 2012.

CHECKLIST

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A CHANGE DETECTION ALGORITHM

Change Detector of M-UCB
The following algorithm is the change detection algorithm for M-UCB (Cao et al., 2019)

Algorithm 2 Change Detection of M-UCB: CD$(w, b, Z_1, \ldots, Z_w)$

Require: An even integer w, w observations $Z_1, \ldots, Z_w$, and a prescribed threshold $b > 0$

1: **if** $\left|\sum_{\ell=w/2+1}^{w} Z_\ell - \sum_{\ell=1}^{w/2} Z_\ell\right| > b$ **then**
2: Return True
3: **else**
4: Return False
5: **end if**

In this algorithm, one requires w observations as input and check whether the difference between the sample average of the first half and that of the second half exceeds a prescribed threshold b (line 1).

Integration with the change detector of M-UCB.

The following two assumptions are required to establish the analytical results when integrating M-UCB into our modularized framework.

Assumption A.1. The algorithm knows a lower bound $\delta > 0$ such that $\delta \leq \min_i \max_{k \in \mathcal{K}} \delta_k^{(i)}$.

Note that Assumption A.1 is Assumption 1(b) of Cao et al. (2019), which is required for the M-UCB detector to determine good w and b in regret analysis. It is worth noting that almost all schemes that actively detect changes share similar assumptions; however, different algorithms may impose distinct sets of assumptions. This assumption is mild since δ may be statistically derived from historical information. Furthermore, even if the lower bound does not hold, and we occasionally encounter changes with expected reward gaps smaller than the assumed δ, such changes may be perceived as too minor to result in significant regret.

Assumption A.2. We assume

$$s_i = \Omega\left((\log KT + \sqrt{K \log KT})\sqrt{s_{i-1}}\right). \tag{20}$$

In particular, if $s_i = \Theta\left(\left(\log KT + \sqrt{K \log KT}\right)^{2(1+\epsilon)}\right)$ for every i, Assumption A.2 holds. This assumption essentially posits that the changes are not overly dense, a condition that holds in many practical scenarios. Simple math would then show that given D_{i-1} is true, with this assumption and the proposed diminishing exploration, each arm will have at least $w/2$ observations before and after a change point. Again, we note that similar assumptions are imposed in other algorithms that actively detect changes, while different algorithms may impose different assumptions. This assumption is necessary with our proof technique, which requires every change to be successfully detected with high probability.

B PROOF DETAIL

In this appendix, we provide detailed proofs of the results presented in Section 4. Each proof is carefully elaborated to ensure clarity and rigor.

B.1 Proof of Section 4

In this subsection, we present the proofs of theorems in Section 4. In what follows, the first lemma bounds the regret accumulated during the diminishing exploration part of the algorithm. We denote by $R_{\mathrm{DE}}(\tau_{i-1}, \nu_i)$ as the regret caused by the exploration part of the algorithm and by $N_{\mathrm{DE},k}(\tau_{i-1}, \nu_i)$ the number of times that the arm k is selected in the exploration phase from the previous alarm time to the next change point.

Lemma B.1 (Diminishing exploration regret)**.** *If the mean values of the arms remain the same during the time interval* $[\tau_{i-1}, \nu_i)$*, then we have*

$$N_{\mathrm{DE},k}\left(\tau_{i-1}, \nu_i\right) \leq \frac{2\alpha\sqrt{\nu_i - \tau_{i-1}}}{K} + \frac{3}{2}, \tag{21}$$

and

$$\mathbb{E}\left[R_{\mathrm{DE}}\left(\tau_{i-1}, \nu_i\right)\right] \leq 2\alpha\sqrt{\nu_i - \tau_{i-1}} + \frac{3}{2}K. \tag{22}$$

Proof. Recall that $u_i^{(j)}$ is the beginning of the j-th uniform exploration session in the i-th segment. In Algorithm 1, the initial time of the first exploration session after each τ_i is given by:

$$u_i^{(1)} = \left\lceil \left(\alpha - \frac{K}{4\alpha}\right)^2 \right\rceil, \tag{23}$$

and subsequent times follow the recursive equation:

$$u_i^{(j)} = \left\lceil u_i^{(j-1)} + \frac{K}{\alpha}\sqrt{u_i^{(j-1)}} + \frac{K^2}{4\alpha^2} \right\rceil \geq u_i^{(j-1)} + \frac{K}{\alpha}\sqrt{u_i^{(j-1)}} + \frac{K^2}{4\alpha^2}. \tag{24}$$

Based on equation 23 and equation 24, one could easily check that the sequence $u_i^{(n)}$ satisfies that for every natural number n,

$$u_i^{(n)} \geq \left(\frac{(2n-3)\,K}{4\alpha} + \alpha\right)^2. \tag{25}$$

Let $u_i^{(m)}$ be the last exploration start time in time interval $[\tau_{i-1}, \nu_i)$. Then, we have

$$\mathbb{E}\left[R_{\mathrm{DE}}\left(\nu_i, \tau_{i-1}\right)\right] \leq mK. \tag{26}$$

Additionally, we have:

$$\nu_i - \tau_{i-1} \geq u_i^{(m)} \geq \left(\frac{(2m-3)\,K}{4\alpha} + \alpha\right)^2 \geq \left(\frac{2\mathbb{E}\left[R_{DE}\left(\nu_i - \tau_{i-1}\right)\right] - 3K}{4\alpha} + \alpha\right)^2. \tag{27}$$

Finally, based on the equation 26 and equation 27, we can conclude that:

$$\mathbb{E}\left[R_{\mathrm{DE}}\left(\nu_i, \tau_{i-1}\right)\right] \leq 2\alpha\sqrt{\nu_i - \tau_{i-1}} - 2\alpha^2 + \frac{3}{2}K \leq 2\alpha\sqrt{\nu_i - \tau_{i-1}} + \frac{3}{2}K, \tag{28}$$

and

$$N_{\mathrm{DE},k}\left(\nu_i, \tau_{i-1}\right) \leq \frac{2\alpha\sqrt{\nu_i - \tau_{i-1}}}{K} + \frac{3}{2}. \tag{29}$$

□

In the following lemma, we aim to explore how many time steps it takes for a given arm to reach a certain number of samples through diminishing exploration.

Lemma B.2 (Samples-time steps transform)**.** *When each arm has accumulated n samples, the required time is as follows: If the counting of the n samples begins immediately after a reset, the required time is given by*

$$T_{reset} \leq \left(\alpha + \frac{(2n-3)\,K}{4\alpha} + n\right)^2 + K. \tag{30}$$

However, if there is a delay of t_d *time steps after the reset before the counting begins, the required time is given by*

$$T_{t_d} \leq 2n\left(\frac{K}{2\alpha} + 1\right)\sqrt{t_d + 1} + n^2\left(\frac{K}{2\alpha} + 1\right)^2. \tag{31}$$

Proof. We can derive the following from Equation 24 in the proof of Lemma B.1:

$$u^{(j)} \leq u^{(j-1)} + \frac{k}{\alpha}\sqrt{u^{(j-1)}} + \frac{k^2}{4\alpha^2} + 1 \leq \left(\sqrt{u^{(j-1)}} + \frac{k}{2\alpha}\right)^2 + 1 \leq \left(\sqrt{u^{(j-1)}} + \frac{k}{2\alpha} + 1\right)^2. \tag{32}$$

First, let us consider the case where the counting of n samples begins immediately after the reset. From Equation 23, we can derive the following:

$$u^{(1)} \leq \left(\alpha - \frac{K}{4\alpha} + 1\right)^2. \tag{33}$$

Using Equation 32, we can further derive:

$$u^{(2)} \leq \left(\alpha - \frac{K}{4\alpha} + 1 + \frac{K}{2\alpha} + 1\right)^2. \tag{34}$$

Finally, we obtain:

$$u^{(n)} \leq \left(\alpha - \frac{K}{4\alpha} + 1 + (n-1)\left(\frac{K}{2\alpha} + 1\right)\right)^2. \tag{35}$$

Here, $u^{(n)}$ represents the starting time of the n-th exploration block. The total time required to guarantee that all K arms have been sampled n times is therefore:

$$T_{\text{reset}} = u^{(n)} + K \leq \left(\alpha + \frac{(2n-3)K}{4\alpha} + n\right)^2 + K. \tag{36}$$

Next, we consider how long it takes for each arm to be sampled n times after t_d time steps following the reset. We first assume that, prior to t_d, each arm has already been sampled x times. For simplicity, we assume an ideal case where the exploration block starts exactly at time $t_d + 1$. Therefore, we have:

$$u^{(x+1)} = t_d + 1. \tag{37}$$

Since, in reality, the exploration block start time may not exactly coincide with $t_d + 1$, we account for the possibility that it could begin at a later time by considering the next exploration block's start time. This allows us to bound the non-ideal case.

Following the same approach as in the first part of the proof, we eventually obtain:

$$u^{(x+n+1)} \leq \left(\sqrt{t_d + 1} + n\left(\frac{K}{2\alpha} + 1\right)\right)^2. \tag{38}$$

Finally, we derive that the total time required for each arm to be sampled n times after t_d time steps is:

$$T_{t_d} = u^{(x+n+1)} - u^{(x+1)} \leq 2n\left(\frac{K}{2\alpha} + 1\right)\sqrt{t_d + 1} + n^2\left(\frac{K}{2\alpha} + 1\right)^2. \tag{39}$$

□

Define $R_{\text{excess}}(r, s) := \mathbb{E}[\sum_{t=r}^{s} r_t^{\text{seg}}] - \sum_{t=r}^{s} r_t^{\pi}$ be the regret accumulated during r and s. In the next lemma, we provide an upper bound on the regret accumulated from the $(i-1)$-th alarm time to the end of $(i-1)$-th segment, given that the previous change was successfully detected.

Lemma B.3 (Excess Regret bound with stationary bandit)**.** *Consider a stationary bandit interval with $\nu_{i-1} < \tau_{i-1} < \nu_i$. Condition on the successful detection events $\overline{F}_{i-1}$ and D_{i-1}, the expected regret accumulated during (τ_{i-1}, ν_i) can be bounded by*

$$\mathbb{E}\left[R_{\text{excess}}\left(\tau_{i-1}, \nu_i\right) \middle| \overline{F}_{i-1} D_{i-1}\right] \leq 2\alpha\sqrt{s_i} + T \cdot \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right), \tag{40}$$

Proof. We begin by rewriting the definition of the excess regret over a stationary segment $(\tau_{i-1}, \nu_i]$ as

$$R_{\text{excess}}(\tau_{i-1}, \nu_i) = \left(\mathbb{E}\Big[\textstyle\sum_{t=\tau_{i-1}}^{\nu_i} r_t^\star\Big] - \sum_{t=\tau_{i-1}}^{\nu_i} r_t^\pi\right) - \left(\mathbb{E}\Big[\textstyle\sum_{t=\tau_{i-1}}^{\nu_i} r_t^\star\Big] - \mathbb{E}\Big[\textstyle\sum_{t=\tau_{i-1}}^{\nu_i} r_t^{\text{seg}}\Big]\right),$$

where the first term corresponds to the standard regret and the second to the standard regret of the segment oracle. We denote the standard regret over this interval by $R(\tau_{i-1}, \nu_i)$ and analyze it first.

For every i, conditioning on the previous successful detection $\overline{F}_{i-1} D_{i-1}$, we decompose the regret depending on whether a false alarm occurs:

$$\begin{aligned}
\mathbb{E}\big[R\left(\tau_{i-1}, \nu_i\right) \big| \overline{F}_{i-1} D_{i-1}\big] &= \mathbb{E}\big[R\left(\tau_{i-1}, \nu_i\right) \big| F_i \overline{F}_{i-1} D_{i-1}\big]\, \mathbb{P}\big(F_i \big| \overline{F}_{i-1} D_{i-1}\big) && \text{(41a)}\\
&\quad + \mathbb{E}\big[R\left(\tau_{i-1}, \nu_i\right) \big| \overline{F}_i \overline{F}_{i-1} D_{i-1}\big]\, \mathbb{P}\big(\overline{F}_i \big| \overline{F}_{i-1} D_{i-1}\big) && \text{(41b)}\\
&\leq T \cdot \mathbb{P}\big(F_i \big| \overline{F}_{i-1} D_{i-1}\big) + \mathbb{E}\big[R\left(\tau_{i-1}, \nu_i\right) \big| \overline{F}_i \overline{F}_{i-1} D_{i-1}\big] && \text{(41c)}\\
&\leq T \cdot \mathbb{P}\big(F_i \big| \overline{F}_{i-1} D_{i-1}\big) + 2\alpha\sqrt{s_i} + \sum_{i=1}^{M} C_i, && \text{(41d)}
\end{aligned}$$

Next, subtracting the segment–oracle standard regret on the same interval cancels out the base regret terms $\sum_{i=1}^{M} C_i$, leaving only the exploration and detection contributions. This yields the desired excess regret bound stated in Lemma B.3.

□

Theorem B.4 (Excess-regret bound)**.** *For any base solver $\mathcal{B}$ and change detector $\mathcal{D}$, the excess regret of Algorithm 1 satisfies*

$$\mathcal{R}_{\text{excess}}(T) \leq 2\alpha\sqrt{MT} + \sum_{i=1}^{M-1} \mathbb{E}\left[\tau_i - \nu_i \middle| D_i \overline{F}_i D_{i-1} \overline{F}_{i-1}\right] + T \sum_{i=1}^{M} \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) + T \sum_{i=1}^{M-1} \mathbb{P}\left(\overline{D}_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right), \tag{42}$$

Proof. Recall that $R_{\text{excess}}(r, s) := \mathbb{E}[\sum_{t=1}^{T} r_t^{\text{seg}}] - \sum_{t=1}^{T} r_t^\pi$, then $\mathcal{R}_{\text{excess}}\left(T\right) = \mathbb{E}\left[R_{\text{excess}}\left(1, T\right)\right]$. We have

$$\begin{aligned}
\mathcal{R}_{\text{excess}}\left(T\right) &= \mathbb{E}\left[R_{\text{excess}}\left(1, T\right)\right] && \text{(43a)}\\
&= \mathbb{E}\left[R_{\text{excess}}\left(1, T\right) \middle| \overline{F}_0 D_0\right] && \text{(43b)}\\
&\leq \mathbb{E}\left[R_{\text{excess}}\left(1, \nu_1\right) \middle| \overline{F}_1 \overline{F}_0 D_0\right] + \mathbb{E}\left[R_{\text{excess}}\left(\nu_1, T\right) \middle| \overline{F}_1 \overline{F}_0 D_0\right] + T \cdot \mathbb{P}\left(F_1 \middle| \overline{F}_0 D_0\right) && \text{(43c)}\\
&\leq 2\alpha\sqrt{(\nu_1 - \nu_0)} + \mathbb{E}\left[R\left(\nu_1, T\right) \middle| \overline{F}_1 \overline{F}_0 D_0\right] + T \cdot \mathbb{P}\left(F_1 \middle| \overline{F}_0 D_0\right), && \text{(43d)}
\end{aligned}$$

where equation 43b holds because $\tau_0 = 0$, equation 43c is due to the law of total expectation and some trivial bounds, and equation 43d follows from Lemmas B.9. The third term in equation 43d is then further bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[R_{\text{excess}}\left(\nu_1, T\right) \middle| \overline{F}_1 \overline{F}_0 D_0\right] &\leq \mathbb{E}\left[R_{\text{excess}}\left(\nu_1, T\right) \middle| D_1 \overline{F}_1 \overline{F}_0 D_0\right] + T \cdot \left(1 - \mathbb{P}\left(D_1 \middle| \overline{F}_1 \overline{F}_0 D_0\right)\right) && \text{(44a)}\\
&\leq \mathbb{E}\left[R_{\text{excess}}\left(\nu_1, T\right) \middle| D_1 \overline{F}_1 \overline{F}_0 D_0\right] + T \cdot \mathbb{P}\left(\overline{D}_1 \middle| \overline{F}_1 \overline{F}_0 D_0\right) && \text{(44b)}\\
&= \mathbb{E}\left[R_{\text{excess}}\left(\tau_1, T\right) \middle| D_1 \overline{F}_1 \overline{F}_0 D_0\right] + \mathbb{E}\left[R\left(\nu_1, \tau_1\right) \middle| D_1 \overline{F}_1 \overline{F}_0 D_0\right] + T \cdot \mathbb{P}\left(\overline{D}_1 \middle| \overline{F}_1 \overline{F}_0 D_0\right) && \text{(44c)}\\
&\leq \mathbb{E}\left[R_{\text{excess}}\left(\tau_1, T\right) \middle| \overline{F}_1 D_1\right] + \mathbb{E}\left[\tau_1 - \nu_1 \middle| \overline{F}_1 D_1 \overline{F}_0 D_0\right] + T \cdot \mathbb{P}\left(\overline{D}_1 \middle| \overline{F}_1 \overline{F}_0 D_0\right) && \text{(44d)}\\
&\leq \mathbb{E}\left[R_{\text{excess}}\left(\tau_1, T\right) \middle| \overline{F}_1 D_1\right] + \mathbb{E}\left[\tau_1 - \nu_1 \middle| \overline{F}_1 D_1\right] + T \cdot \mathbb{P}\left(\overline{D}_1 \middle| \overline{F}_1 \overline{F}_0 D_0\right), && \text{(44e)}
\end{aligned}$$

where equation 44a applies the law of total expectation and some trivial bounds. From here, we can set up the following recursion:

$$\mathbb{E}\left[R_{\text{excess}}\left(1,T\right)\right] = \mathbb{E}\left[R_{\text{excess}}\left(1,T\right)\middle|\overline{F}_0 D_0\right] \tag{45a}$$

$$\leq \mathbb{E}\left[R_{\text{excess}}\left(\tau_1,T\right)\middle|\overline{F}_1 D_1\right] + 2\alpha\sqrt{s_1-1} + \mathbb{E}\left[\tau_1-\nu_1\middle|\overline{F}_1 D_1\right] \tag{45b}$$

$$+ T\cdot\mathbb{P}\left(F_1\middle|\overline{F}_0 D_0\right) + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0 D_0\right) \tag{45c}$$

$$\leq \mathbb{E}\left[R_{\text{excess}}\left(\tau_2,T\right)\middle|\overline{F}_2 D_2\right] + 2\alpha\sum_{i=1}^{2}\sqrt{s_i-1} \tag{45d}$$

$$+\sum_{i=1}^{2}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] + T\sum_{i=1}^{2}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{2}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{45e}$$

$$\vdots$$

$$\leq 2\alpha\sum_{i=1}^{M}\sqrt{s_i} + \sum_{i=1}^{M-1}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] \tag{45f}$$

$$+ T\sum_{i=1}^{M}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{M-1}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{45g}$$

$$\leq 2\alpha\sqrt{MT} + \sum_{i=1}^{M-1}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] \tag{45h}$$

$$+ T\sum_{i=1}^{M}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{M-1}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{45i}$$

where equation 45h follows from the Cauchy–Schwarz inequality

$$\left(\sum_{i=1}^{M}\sqrt{s_i}\right)^2 \leq \left(\sum_{i=1}^{M}s_i\right)\left(\sum_{i=1}^{M}1\right) = M\sum_{i=1}^{M}s_i = MT, \tag{46a}$$

and the fact that $\sum_{i=1}^{M}s_i = T$. This completes the proof. □

B.2 Proof of One-State Special Case

With the general regret bound in Theorem B.4, a regret bound of the change detector with the proposed diminishing exploration can be obtained by bounding $\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right)$, $\mathbb{P}\left(D_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right)$, and $\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_i D_i\overline{F}_{i-1}D_{i-1}\right]$. In what follows, we adopt window-based change detectors (Algorithm A) as the basis for our subsequent analysis.

First, in Lemma B.5, we show that the probability of false alarm is very small; thereby, its contribution to the regret is negligible.

Lemma B.5 (Probability of false alarm)**.** *Under Algorithm 1 with parameter in Equations 3 and 4, we have*

$$\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) \leq wK\left(1-\left(1-\exp\left(-2b^2/w\right)\right)^{\lfloor T/w\rfloor}\right) \leq \frac{1}{T}. \tag{47}$$

Proof. Suppose that at time t, we have gathered w samples of arm $k\in\mathcal{K}$, namely $Y_{k,1}, Y_{k,2}, \ldots, Y_{k,w}$, for change detection in line 17 of Algorithm 1, and we define

$$S_{k,t} = \sum_{\ell=w/2+1}^{w} Y_{k,\ell} - \sum_{\ell=1}^{w/2} Y_{k,\ell}. \tag{48}$$

Note that $S_{k,t}=0$ when there is insufficient (less than w) samples to trigger the change detection algorithm. By definition, we have

$$\tau_{k,i} = \inf\{t\geq\tau_{i-1}+w : |S_{k,t}|>b\}. \tag{49}$$

Given that the events D_{i-1} and $\bar{F}_{i-1}$ hold, we define $\tau_{k,i}$ as the first detection time of the k-th arm after ν_i. Clearly, $\tau_i = \min_{k\in\mathcal{K}}\{\tau_{k,i}\}$ as Algorithm 1 would reset every time a change is detected. Using the union bound, we have

$$\begin{aligned}
\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) =& \mathbb{P}\left(\max_{k\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w, F_i\middle|\overline{F}_{i-1}D_{i-1}\right) && \text{(50a)}\\
&+\mathbb{P}\left(\max_{k\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}< w, F_i\middle|\overline{F}_{i-1}D_{i-1}\right) && \text{(50b)}\\
=&\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}, \max_{k\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\right) && \text{(50c)}\\
&\cdot\mathbb{P}\left(\max_{k\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\middle|\overline{F}_{i-1}D_{i-1}\right) && \text{(50d)}\\
\leq&\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}, \max_{k\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\right) && \text{(50e)}\\
\leq&\sum_{k=1}^{K}\mathbb{P}\left(\tau_{k,i}\leq\nu_i\middle|\overline{F}_{i-1}D_{i-1}, \max_{k'\in\mathcal{K}}\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k'\}}\geq w\right) && \text{(50f)}\\
\leq&\sum_{k=1}^{K}\mathbb{P}\left(\tau_{k,i}\leq\nu_i\middle|\overline{F}_{i-1}D_{i-1}, \sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\right), && \text{(50g)}
\end{aligned}$$

where the term in equation 50b is clearly equal to 0 as there will be no false alarm if we do not even have sufficiently many observations to trigger the alarm as suggested by Algorithm 2. Equation 50c and equation 50d hold by the definition of conditional probability, equation 50e is due to the fact that the term in equation 50d is at most one, and equation 50f follows from the union bound. In equation 50g, if $k\neq k'$, we cannot guarantee that $\sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k'\}}\geq w$. Hence, some k might cause the probability in the equation 50f to be zeros.

For any $0\leq j\leq w-1$, define the stopping time

$$\tau_{k,i}^{(j)} := \inf\{t=\tau_{i-1}+j+nw, n\in\mathbb{Z}^+ : |S_{k,t}|>b\}. \tag{51}$$

Clearly, $\tau_{k,i}=\min\{\tau_{k,i}^{(0)},\ldots,\tau_{k,i}^{(w-1)}\}$. Let us define, for any $0\leq j\leq w-1$,

$$\xi_{k,i}^{(j)} = \frac{\left(\tau_{k,i}^{(j)}-j-\tau_{i-1}\right)}{w}. \tag{52}$$

Note that condition on the events D_{i-1} and $\bar{F}_{i-1}$, $\xi_{k,i}^{(j)}$ is a geometric random variable with parameter $p:=\mathbb{P}(|S_{k,t}|>b)$, because when fixing j, there is no overlap between the samples in the current window and the next.

$$\begin{aligned}
&\mathbb{P}\left(\tau_{k,i}^{(j)}=\tau_{i-1}+nw+j\middle|\overline{F}_{i-1}D_{i-1}, \sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\right)\\
&\qquad=\mathbb{P}\left(\xi_{k,i}=n\middle|\overline{F}_{i-1}D_{i-1}, \sum_{t=\tau_{i-1}+1}^{\nu_i}\mathbf{1}_{\{A_t=k\}}\geq w\right)=p(1-p)^{n-1}. \qquad \text{(53)}
\end{aligned}$$

Here, the inclusion of subsequent events as conditions should not impact the results, as when entering the change

detection algorithm, those events have already occurred. Moreover, by union bound, we have that for any $k \in \mathcal{K}$,

$$\mathbb{P}\left(\tau_{k,i} \leq \nu_i \middle| \overline{F}_{i-1} D_{i-1}, \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\{A_t=k\}} \geq w\right) \leq w\left(1-(1-p)^{\lfloor(\nu_i-\tau_{i-1})/w\rfloor}\right) \tag{54a}$$

$$\leq w\left(1-(1-p)^{\lfloor T/w\rfloor}\right). \tag{54b}$$

We further use the McDiarmid's inequality and the union bound to show that

$$p = \mathbb{P}\left(|S_{k,t}| > b\right) = \mathbb{P}\left(S_{k,t} > b\right) + \mathbb{P}\left(S_{k,t} < -b\right) \tag{55a}$$

$$\leq 2 \cdot \exp\left(-\frac{2b^2}{w}\right). \tag{55b}$$

Using the result in equation 54b and equation 55b into equation 50g,

$$\mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) \leq \sum_{k=1}^{K} w\left(1-\left(1-2\exp\left(-\frac{2b^2}{w}\right)\right)^{\lfloor T/w\rfloor}\right) \tag{56a}$$

$$= wK\left(1-\left(1-2\exp\left(-\frac{2b^2}{w}\right)\right)^{\lfloor T/w\rfloor}\right). \tag{56b}$$

Moreover, applying $(1-x)^a > 1-ax$ for any $a > 1$ and $0 < x < 1$ and plugging the choice of $b = \sqrt{w\log\left(2KT^2\right)/2}$ as in equation 4 shows the second inequality. □

Lemma B.2 ensures that, with high probability, the detection delay is confined within a tolerable interval. That is, each arm is sampled $w/2$ times, and using equation 31 from lemma B.2, we select h_i as

$$h_i = \left\lceil w\left(\frac{K}{2\alpha}+1\right)\sqrt{s_i+1} + \frac{w^2}{4}\left(\frac{K}{2\alpha}+1\right)^2\right\rceil. \tag{57}$$

Lemma B.6 (Probability of successful detection)**.** *Consider a piecewise-stationary bandit environment. For any $\boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(i+1)} \in [0,1]^K$ with parameters chosen in equation 3 and equation 4 and*

$$h_i = \left\lceil w\left(\frac{K}{2\alpha}+1\right)\sqrt{s_i+1} + \frac{w^2}{4}\left(\frac{K}{2\alpha}+1\right)^2\right\rceil, \tag{58}$$

for some $k \in \mathcal{K}, i \geq 1$ and $c > 0$, under the Algorithm 1, we have

$$\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \geq 1 - \frac{1}{T}. \tag{59}$$

Proof.

$$\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) = \mathbb{P}\left(\tau_i \leq \nu_i + h_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{60a}$$

$$\geq \max_{t\in\{\nu_i+1,\ldots,\nu_i+h_i\}} \mathbb{P}\left(S_{\tilde{k},t} > b \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{60b}$$

$$\geq \max_{j\in\{0,\ldots,w/2\}} \left(1-2\exp\left(-\frac{(j\left|\delta_{\tilde{k}}^{(i)}\right|-b)^2}{w}\right)\right) \tag{60c}$$

$$= 1-2\exp\left(-\frac{(w|\delta_{\tilde{k}}^{(i)}|/2-b)^2}{w}\right) \tag{60d}$$

$$\geq 1-2\exp\left(-\frac{wc^2}{4}\right). \tag{60e}$$

where $S_{\tilde{k},t}$ is defined in equation 48, equation 60c follows from the McDiarmid's inequality, and equation 60d is due to the fact that the maximum value is attained when $j = w/2$. Last, equation 60e is true for any choice of w, b and c such that $\delta_{\tilde{k}}^{(i)} \geq 2b/w + c$ holds. We thus set w and b as in equation 3 and equation 4, respectively, and choose $c = 2\sqrt{\log(2T)/w}$, which leads to $\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \geq 1 - 1/T$. □

Lemma B.7 further bounds the expected detection delay in the situation where the change detection algorithm successfully detects the change within the desired interval.

Lemma B.7 (Expected detection delay)**.** *Consider a piecewise-stationary bandit environment. For any $\boldsymbol{\mu}^{(i)}, \boldsymbol{\mu}^{(i+1)} \in [0,1]^K$ with parameters chosen in equation 3 and equation 4 and*

$$h_i = \left\lceil w\left(\frac{K}{2\alpha} + 1\right)\sqrt{s_i + 1} + \frac{w^2}{4}\left(\frac{K}{2\alpha} + 1\right)^2 \right\rceil, \tag{61}$$

for some $k \in \mathcal{K}, i \geq 1$ and $c > 0$, under the Algorithm 1, we have

$$\mathbb{E}\left[\tau_i - \nu_i \middle| \overline{F}_i D_i \overline{F}_{i-1} D_{i-1}\right] \leq h_i. \tag{62}$$

Proof. For any $1 \leq i \leq M$, we have

$$\mathbb{E}\left[\tau_i - \nu_i \middle| \overline{F}_i D_i \overline{F}_{i-1} D_{i-1}\right] = \sum_{j=1}^{h_i} \mathbb{P}\left(\tau_i \geq \nu_i + j \middle| \overline{F}_i D_i \overline{F}_{i-1} D_{i-1}\right) \leq h_i. \tag{63a}$$

□

Plugging the bounds in Lemmas B.5, B.6 and B.7 into Theorem B.4 shows the following regret bound in Corollary B.8.

Corollary B.8. *Combining Algorithm 1 (diminishing exploration with resets) with the base solver $\mathcal{B}$ and Algorithm 2 with parameters (w, b) given in Equations 3 and 4, the excess regret is upper-bounded as*

$$\mathcal{R}_{\text{excess}}(T) \leq 2\alpha\sqrt{MT} + w\left(\frac{K}{2\alpha} + 1\right)\sqrt{M(T+M)} + \frac{w^2 M}{4}\left(\frac{K}{2\alpha} + 1\right)^2 + 2M. \tag{64}$$

By setting $\alpha = c\sqrt{K \log(KT)}$ for some constant c, the expected regret is upper-bounded by $\mathcal{O}(\sqrt{KMT \log T})$.

B.2.1 Proof of standard regret in one-state case

Define $R(r,s) := \sum_{t=r}^{s} \max_{k \in \mathcal{K}} \mathbb{E}[X_{k,t}] - X_{A_t,t}$ be the regret accumulated during r and s. In the next lemma, we provide an upper bound on the regret accumulated from the $(i-1)$-th alarm time to the end of $(i-1)$-th segment, given that the previous change was successfully detected.

Lemma B.9 (Regret bound with stationary bandit)**.** *Consider a stationary bandit interval with $\nu_{i-1} < \tau_{i-1} < \nu_i$. Condition on the successful detection events $\overline{F}_{i-1}$ and D_{i-1}, the expected regret accumulated during (τ_{i-1}, ν_i) can be bounded by*

$$\mathbb{E}\left[R(\tau_{i-1}, \nu_i) \middle| \overline{F}_{i-1} D_{i-1}\right] \leq \tilde{C}_i + 2\alpha\sqrt{s_i} + T \cdot \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right), \tag{65}$$

where $\tilde{C}_i = 8\sum_{\Delta_k^{(i)} > 0} \frac{\log T}{\Delta_k^{(i)}} + \left(\frac{5}{2} + \frac{\pi^2}{3} + K\right)\sum_{k=1}^{K} \Delta_k^{(i)}$.

Proof. For every i, we have

$$\begin{aligned}
\mathbb{E}\left[R(\tau_{i-1}, \nu_i) \middle| \overline{F}_{i-1} D_{i-1}\right] &= \mathbb{E}\left[R(\tau_{i-1}, \nu_i) \middle| F_i \overline{F}_{i-1} D_{i-1}\right] \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) && \text{(66a)} \\
&\quad + \mathbb{E}\left[R(\tau_{i-1}, \nu_i) \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right] \mathbb{P}\left(\overline{F}_i \middle| \overline{F}_{i-1} D_{i-1}\right) && \text{(66b)} \\
&\leq T \cdot \mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) + \mathbb{E}\left[R(\tau_{i-1}, \nu_i) \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right]. && \text{(66c)}
\end{aligned}$$

Now, define $N_k(t_1, t_2) := \sum_{t=t_1}^{t_2} \mathbf{1}_{\{A_t=k\}}$ to be the number of times that arm k is selected by Algorithm 1 from t_1 to t_2. Note that

$$\mathbb{E}\left[R(\tau_{i-1}, \nu_i)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] = \sum_{\Delta_k^{(i)}>0} \Delta_k^{(i)} \cdot \mathbb{E}\left[N_k(\tau_{i-1}, \nu_i)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right]. \tag{67}$$

To bound the second term of equation 66c, we further bound $N_k(\tau_{i-1}, \nu_i)$ as follows,

$$\begin{aligned}
N_k(\tau_{i-1}, \nu_i) =& \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\{A_t=k, \tau_i>\nu_i, N_k(\tau_{i-1},\nu_i)<l\}} + \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\{A_t=k, \tau_i>\nu_i, N_k(\tau_{i-1},\nu_i)\geq l\}} && \text{(68a)}\\
\leq& l + N_{DE,k}(\nu_i - \tau_{i-1}) + \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\left\{k=\arg\max_{k\in\mathcal{K}} \mathrm{UCB}_{k\in\mathcal{K}}, \tau_i>\nu_i, N_k(\tau_{i-1},\nu_i)\geq l\right\}} && \text{(68b)}\\
\leq& l + \frac{2\alpha\sqrt{\nu_i - \tau_{i-1}}}{K} + \frac{3}{2} + \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\left\{k=\arg\max_{k\in\mathcal{K}} \mathrm{UCB}_{k\in\mathcal{K}}, \tau_i>\nu_i, N_k(\tau_{i-1},\nu_i)\geq l\right\}} && \text{(68c)}\\
\leq& l + \frac{2\alpha\sqrt{s_i}}{K} + \frac{3}{2} + \sum_{t=\tau_{i-1}+1}^{\nu_i} \mathbf{1}_{\left\{k=\arg\max_{k\in\mathcal{K}} \mathrm{UCB}_{k\in\mathcal{K}}, \tau_i>\nu_i, N_k(\tau_{i-1},\nu_i)\geq l\right\}}, && \text{(68d)}
\end{aligned}$$

Equation 68c follows from Lemma B.1. Setting $l = \left\lceil 8\log T / \left(\Delta_k^{(i)}\right)^2 \right\rceil$, and following the same steps as in the proof of Theorem 1 of Auer et al. (2002), we arrive at

$$\mathbb{E}\left[N_k(\tau_{i-1}, \nu_i)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] \leq \frac{2\alpha\sqrt{s_i}}{K} + \frac{8\log T}{\left(\Delta_k^{(i)}\right)^2} + \frac{5}{2} + \frac{\pi^2}{3} + K. \tag{69}$$

Putting everything together completes the proof. □

The term $\tilde{C}_i$ in equation 65 involves $\Delta_k^{(i)}$. In what follows, we establish an upper bound of $\tilde{C}_i$ that avoids this dependence.

Lemma B.10 (Regret bound without involving $\Delta_k^{(i)}$)**.** *Consider a stationary bandit interval with $\nu_{i-1} < \tau_{i-1} < \nu_i$. Condition on the successful detection events $\overline{F}_{i-1}$ and D_{i-1}, and let $\Delta > 0$, the expected regret accumulated during (τ_{i-1}, ν_i) can be bounded by*

$$\mathbb{E}\left[R(\tau_{i-1}, \nu_i)\middle|\overline{F}_{i-1}D_{i-1}\right] \leq 2\alpha\sqrt{s_i} + \frac{8K\log T}{\Delta} + \Delta\cdot s_i + \left(\frac{5}{2} + \frac{\pi^2}{3} + K\right)K + T\cdot\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right). \tag{70}$$

Moreover, by selecting $\Delta = \sqrt{MK\log T/T}$, we have

$$\mathbb{E}\left[R(\tau_{i-1}, \nu_i)\middle|\overline{F}_{i-1}D_{i-1}\right] \leq 2\alpha\sqrt{s_i} + 8\sqrt{\frac{KT\log T}{M}} + \sqrt{\frac{MK\log T}{T}}s_i + \left(\frac{5}{2} + \frac{\pi^2}{3} + K\right)K + T\cdot\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right). \tag{71}$$

Proof. We rewite the equation 67 as follows,

$$\mathbb{E}\left[R\left(\tau_{i-1},\nu_i\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] = \sum_{\Delta_k^{(i)}>0}\Delta_k^{(i)}\cdot\mathbb{E}\left[N_k\left(\tau_{i-1},\nu_i\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] \tag{72a}$$

$$= \sum_{\Delta_k^{(i)}\geq\Delta}\Delta_k^{(i)}\cdot\mathbb{E}\left[N_k\left(\tau_{i-1},\nu_i\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] \tag{72b}$$

$$+ \sum_{\Delta_k^{(i)}<\Delta}\Delta_k^{(i)}\cdot\mathbb{E}\left[N_k\left(\tau_{i-1},\nu_i\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right] \tag{72c}$$

$$\leq \sum_{\Delta_k^{(i)}\geq\Delta}\left[\frac{2\alpha\sqrt{s_i}}{K}\cdot\Delta_k^{(i)}+\frac{8\log T}{\Delta_k^{(i)}}+\left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)\cdot\Delta_k^{(i)}\right]+\Delta\cdot s_i \tag{72d}$$

$$\leq 2\alpha\sqrt{s_i}+\sum_{\Delta_k^{(i)}\geq\Delta}\frac{8\log T}{\Delta_k^{(i)}}+\left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)\sum_{\Delta_k^{(i)}\geq\Delta}\Delta_k^{(i)}+\Delta\cdot s_i \tag{72e}$$

$$\leq 2\alpha\sqrt{s_i}+\frac{8K\log T}{\Delta}+\left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)K+\Delta\cdot s_i. \tag{72f}$$

Equations 72b to 72d are derived using equation 69, while the transition from 72c to 72d leverages the inequalities $\Delta > \Delta_k^{(i)}$ and $s_i > \mathbb{E}\left[N_k\left(\tau_{i-1},\nu_i\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right]$.

The second part of the lemma follows straightforwardly by plugging $\Delta = \sqrt{MK\log T/T}$ into then the equation 72f. □

Theorem B.11 can then be proved by recursively applying Lemma B.9.

Theorem B.11 (Standard Regret bound of M-UCB)**.** *Combining Algorithms 1 and 2 with the parameters in Equation 3, and Equation 4 achieves the expected regret upper bound as follows:*

$$\mathbb{E}\left[R\left(1,T\right)\right] \leq \underbrace{9\sqrt{MKT\log T}+\left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)MK}_{(a)}+\underbrace{2\alpha\sqrt{MT}}_{(b)}+\underbrace{w\left(\frac{K}{2\alpha}+1\right)\sqrt{M\left(T+M\right)}}_{(c)}$$
$$+\underbrace{\frac{w^2M}{4}\left(\frac{K}{2\alpha}+1\right)^2}_{(c)}+\underbrace{2M}_{(d)}. \tag{73}$$

By setting $\alpha = c\sqrt{K\log\left(KT\right)}$ for some constant c, the expected regret is upper-bounded by $\mathcal{O}(\sqrt{KMT\log T})$.

Proof. Recall that $R\left(r,s\right) = \sum_{t=r}^{s}\max_{k\in\mathcal{K}}\mathbb{E}\left[X_{k,t}\right] - X_{A_t,t}$, then $\mathcal{R}\left(T\right) = \mathbb{E}\left[R\left(1,T\right)\right]$. We have

$$\mathcal{R}\left(T\right) = \mathbb{E}\left[R\left(1,T\right)\right] \tag{74a}$$

$$= \mathbb{E}\left[R\left(1,T\right)\middle|\overline{F}_0D_0\right] \tag{74b}$$

$$\leq \mathbb{E}\left[R\left(1,\nu_1\right)\middle|\overline{F}_1\overline{F}_0D_0\right]+\mathbb{E}\left[R\left(\nu_1,T\right)\middle|\overline{F}_1\overline{F}_0D_0\right]+T\cdot\mathbb{P}\left(F_1\middle|\overline{F}_0D_0\right) \tag{74c}$$

$$\leq \tilde{C}_1+2\alpha\sqrt{\left(\nu_1-\nu_0\right)}+\mathbb{E}\left[R\left(\nu_1,T\right)\middle|\overline{F}_1\overline{F}_0D_0\right]+T\cdot\mathbb{P}\left(F_1\middle|\overline{F}_0D_0\right), \tag{74d}$$

where equation 74b holds because $\tau_0 = 0$, equation 74c is due to the law of total expectation and some trivial bounds, and equation 74d follows from Lemmas B.9. The third term in equation 74d is then further bounded as follows:

$$\mathbb{E}\left[R\left(\nu_1,T\right)\middle|\overline{F}_1\overline{F}_0D_0\right] \leq \mathbb{E}\left[R\left(\nu_1,T\right)\middle|D_1\overline{F}_1\overline{F}_0D_0\right] + T\cdot\left(1-\mathbb{P}\left(D_1\middle|\overline{F}_1\overline{F}_0D_0\right)\right) \tag{75a}$$
$$\leq \mathbb{E}\left[R\left(\nu_1,T\right)\middle|D_1\overline{F}_1\overline{F}_0D_0\right] + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0D_0\right) \tag{75b}$$
$$= \mathbb{E}\left[R\left(\tau_1,T\right)\middle|D_1\overline{F}_1\overline{F}_0D_0\right] + \mathbb{E}\left[R\left(\nu_1,\tau_1\right)\middle|D_1\overline{F}_1\overline{F}_0D_0\right] + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0D_0\right) \tag{75c}$$
$$\leq \mathbb{E}\left[R\left(\tau_1,T\right)\middle|\overline{F}_1D_1\right] + \mathbb{E}\left[\tau_1-\nu_1\middle|\overline{F}_1D_1\overline{F}_0D_0\right] + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0D_0\right) \tag{75d}$$
$$\leq \mathbb{E}\left[R\left(\tau_1,T\right)\middle|\overline{F}_1D_1\right] + \mathbb{E}\left[\tau_1-\nu_1\middle|\overline{F}_1D_1\right] + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0D_0\right), \tag{75e}$$

where equation 75a applies the law of total expectation and some trivial bounds.From here, we can set up the following recursion:

$$\mathbb{E}\left[R\left(1,T\right)\right] = \mathbb{E}\left[R\left(1,T\right)\middle|\overline{F}_0D_0\right] \tag{76a}$$
$$\leq \mathbb{E}\left[R\left(\tau_1,T\right)\middle|\overline{F}_1D_1\right] + \tilde{C}_1 + 2\alpha\sqrt{s_1-1} + \mathbb{E}\left[\tau_1-\nu_1\middle|\overline{F}_1D_1\right] \tag{76b}$$
$$+ T\cdot\mathbb{P}\left(F_1\middle|\overline{F}_0D_0\right) + T\cdot\mathbb{P}\left(\overline{D}_1\middle|\overline{F}_1\overline{F}_0D_0\right) \tag{76c}$$
$$\leq \mathbb{E}\left[R\left(\tau_2,T\right)\middle|\overline{F}_2D_2\right] + \sum_{i=1}^{2}\tilde{C}_i + 2\alpha\sum_{i=1}^{2}\sqrt{s_i-1} \tag{76d}$$
$$+ \sum_{i=1}^{2}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] + T\sum_{i=1}^{2}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{2}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{76e}$$
$$\vdots$$
$$\leq \sum_{i=1}^{M}\tilde{C}_i + 2\alpha\sum_{i=1}^{M}\sqrt{s_i} + \sum_{i=1}^{M-1}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] \tag{76f}$$
$$+ T\sum_{i=1}^{M}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{M-1}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{76g}$$
$$\leq \sum_{i=1}^{M}\tilde{C}_i + 2\alpha\sqrt{MT} + \sum_{i=1}^{M-1}\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_{i-1}D_{i-1}\right] \tag{76h}$$
$$+ T\sum_{i=1}^{M}\mathbb{P}\left(F_i\middle|\overline{F}_{i-1}D_{i-1}\right) + T\sum_{i=1}^{M-1}\mathbb{P}\left(\overline{D}_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{76i}$$

where equation 76h follows from the Cauchy–Schwarz inequality

$$\left(\sum_{i=1}^{M}\sqrt{s_i}\right)^2 \leq \left(\sum_{i=1}^{M}s_i\right)\left(\sum_{i=1}^{M}1\right) = M\sum_{i=1}^{M}s_i = MT, \tag{77a}$$

Finally, by lemma B.10, we know that

$$\sum_{i}^{M}\tilde{C}_i \leq \sum_{i=1}^{M}\left[8\sqrt{\frac{KT\log T}{M}} + \left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)K + \sqrt{\frac{MK\log T}{T}}\cdot s_i\right] \tag{78a}$$
$$= 8\sqrt{MKT\log T} + \left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)MK + \sqrt{\frac{MK\log T}{T}}\cdot\sum_{i=1}^{M}s_i. \tag{78b}$$
$$= 8\sqrt{MKT\log T} + \left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)MK + \sqrt{\frac{MK\log T}{T}}\cdot T. \tag{78c}$$
$$= 8\sqrt{MKT\log T} + \left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)MK + \sqrt{MKT\log T} \tag{78d}$$
$$= 9\sqrt{MKT\log T} + \left(\frac{5}{2}+\frac{\pi^2}{3}+K\right)MK. \tag{78e}$$

Then, plugging equation 78e into equation 76h completes the proof of Theorem B.11. □

B.3 Proof of General Case

To handle the temporal dependence induced by the Markovian dynamics, we leverage a concentration inequality for ergodic Markov chains. Assumption B.12 ensures the existence of unique stationary distributions for each arm in each segment, allowing us to apply the following Hoeffding-type bound for Markovian rewards Chung et al. (2012), which plays a key role in bounding the probabilities of false alarms $\mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right)$ and successful detections $\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)$.

Assumption B.12. All Markov chains $\mathcal{M}_k^{(i)} = (P_k^{(i)}, R_k^{(i)})$ are ergodic. This ensures the existence of a unique steady-state distribution $d_k^{(i)}$ for each arm k in segment i, satisfying $d_k^{(i)} = d_k^{(i)} P_k^{(i)}$ with $\sum_{s \in \mathcal{S}} d_k^{(i)}(s) = 1$. We then define the steady-state arm mean as $\bar{\mu}_k^{(i)} := d_k^{(i)} R_k^{(i)}$.

Lemma B.13 (Hoeffding Bound for Markovian Reward Chung et al. (2012))**.** *Let $\mathcal{M}$ be an ergodic Markov chain with state space $\mathcal{S}$ and stationary distribution d. Let $L = T(\epsilon)$ be its ε-mixing time for $\varepsilon \leq 1/8$. Let $(V_1, \cdots, V_t)$ denote a t-step random walk on $\mathcal{M}$ starting from an initial distribution ϕ. Let $f_\ell : \mathcal{S} \mapsto [0,1]$ be a weight function at ℓ time step such that expected weighted $\mathbb{E}_{v \leftarrow d} = \mu$ for all ℓ. Define total weight walk $X \triangleq \sum_{\ell=1}^{t} f_\ell(V_\ell)$. There exists some constant c (which is independent to d, δ, ϵ) and let $0 \leq \delta \leq 1$ such that*

$$\begin{aligned} \mathbb{P}[X \geq (1+\delta)\,\mu t] &\leq c\|\phi\|_d \exp\left(-\delta^2 \mu t / (72L)\right) \\ \mathbb{P}[X \leq (1-\delta)\,\mu t] &\leq c\|\phi\|_d \exp\left(-\delta^2 \mu t / (72L)\right). \end{aligned} \tag{79}$$

Using Lemma B.13, we now control the probability of false alarms within each stationary segment. The following lemma shows that, with appropriate choices of the detection window and threshold parameters, the probability of triggering a false alarm is negligible.

Lemma B.14 (Probability of false alarm)**.** *Under Algorithm 2 with the parameter choices specified in Equations 14 and 15, the probability of a false alarm in each stationary segment satisfies*

$$\mathbb{P}\left(F_i \,\middle|\, \overline{F}_{i-1} D_{i-1}\right) \;\leq\; \frac{\epsilon\|\varphi\|_d + 1}{T}, \tag{80}$$

where $\epsilon > 0$ is an arbitrarily small constant, φ is the initial state distribution, and $\|\varphi\|_d$ is the corresponding d-norm.

Proof. We formalize this analysis by defining an equivalent function $S_{k,t}$ and establishing the stopping conditions necessary for detecting such events. The function $S_{k,t}$ serves as a proxy for the behavior of Algorithm 2 and is defined as follows:

$$S_{k,t} = \sum_{\ell=w/2+1}^{w} Z_{k,\ell} - \sum_{\ell=1}^{w/2} Z_{k,\ell}, \quad \text{for any} \quad k \in K, t \geq w, \tag{81}$$

where w is window size.

To formalize the conditions for triggering a change detection, we define the stopping time: $|S_{k,t}| \geq b$, which is equalivent to the change alerts in Algorithm 2. To facilitate analysis, we divide the stopping time events into w subsequences indexed by j with each subsequence corresponding to a distinct offset. The stopping time for the j-th subsequence is given by:

$$\tau_{k,i}^{(j)} := \inf\left\{t = \tau_{i-1} + j + nw, n \in \mathbb{Z}^+ : |S_{k,t}| > b\right\}, \tag{82}$$

where b is the detection threshold. By dividing the stopping time events into these w subsequences, we ensure that the samples within each subsequences are independent over w-time steps. This independence is achieved because the samples are aggregated over disjoint time intervals defined by the periodic offset j. Such independence is crucial for applying concentration inequalities later in the analysis. Clearly, $\tau_{k,i} = \min\left\{\tau_{k,i}^{(0)}, \ldots, \tau_{k,i}^{(w-1)}\right\}$. Let us define $\xi_{k,i}^{(j)}$

$$\xi_{k,i}^{(j)} = \frac{\left(\tau_{k,i}^{(j)} - j - \tau_{k,i-1}\right)}{w}, \quad \text{for any} \quad 0 \leq j \leq w-1. \tag{83}$$

Note that condition on the events D_{i-1} and $F_{i-1}, \xi_{k,i}^{(j)}$ is a geometric random variable with parameter $p := \mathbb{P}(|S_{k,t}| > b)$, because when fixing j, there is no overlap between the samples in the current window and the next.

$$\mathbb{P}\left(\tau_{k,i}^{(j)} = \tau_{i-1} + nw + j \middle| \bar{F}_{i-1} D_{i-1}, \sum_{t=\tau_i+1}^{\nu_i} \mathbf{1}_{A_t=k} \geq w\right) = \mathbb{P}\left(\xi_{k,i} = n \middle| \bar{F}_{i-1} D_{i-1}, \sum_{t=\tau_i+1}^{\nu_i} \mathbf{1}_{A_t=k} \geq w\right) = p(1-p)^{n-1}.$$

Here, the inclusion of subsequent events as conditions should not impact the results, as when entering the change detection algorithm, those events have already occurred. Moreover, by union bound, we have that for any $i \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{P}\left(\tau_{k,i} \leq \nu_i \mid \bar{F}_{i-1} D_{i-1}\right) &\leq w\left(1 - (1-p)^{\lfloor (\nu_i - \tau_i - 1)/w \rfloor}\right) && \text{(85a)} \\
&\leq w\left(1 - (1-p)^{\lfloor T/w \rfloor}\right). && \text{(85b)}
\end{aligned}$$

We further upper bound the probability p_k for each arm k by decomposing the reward into two parts: (i) a *state-dependent mean component*, determined by the Markovian state of the arm; and (ii) an *independent noise term* $n_t \in [-1, 1]$ with zero mean.

The key idea is to separately bound the effects of the *distributional fluctuations* induced by the Markov chain and the *reward noise*. Accordingly, we split the detection threshold into two components, $b = b_d + b_n$, where b_d handles deviations due to distributional shifts and b_n accounts for noise fluctuations.

Let $V_{k,t}$ denote the state of arm k at time t, and let $f_{k,t} : \mathcal{S} \mapsto [0, 1]$ be the corresponding reward mean function. Then the probability of the detection statistic exceeding the threshold can be bounded as follows:

$$\begin{aligned}
p &= \mathbb{P}\left[|S_{k,\ell}| \geq b\right] && \text{(86a)} \\
&= \mathbb{P}\left[\left|\sum_{\ell=w/2+1}^{w} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right) - \sum_{\ell=1}^{w/2} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right)\right| \geq b_d + b_n\right] && \text{(86b)} \\
&\leq 1 - \mathbb{P}\left[\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}(V_{k,\ell}) - \sum_{\ell=1}^{w/2} f_{k,\ell}(V_{k,\ell})\right| \leq b_d\right] \cdot \mathbb{P}\left[\left|\sum_{\ell=1}^{w} n_\ell\right| \leq b_n\right]. && \text{(86c)}
\end{aligned}$$

To apply the concentration inequality, we focus on the case that induces the largest deviation from the steady-state distribution. This occurs when the rewards within the detection window are collected consecutively in time, which leads to the strongest temporal dependence among samples. We then derive the concentration probability of this distributional deviation. To fit the requirement of Lemma B.13, we shift the original reward function f_ℓ to a new function $f'_{k,\ell}$ such that $f'_{k,\ell} : \mathcal{S} \mapsto [0, 1]$:

$$f'_{k,\ell} = \begin{cases} \frac{1 - f_{k,\ell}}{2} & 1 \leq \ell \leq \frac{w}{2}, \\ \frac{1 + f_{k,\ell}}{2} & \frac{w}{2} + 1 \leq \ell \leq w. \end{cases} \tag{87}$$

We first focus on the first probability term in Equation 86c. By substituting Equation 87 into it, we obtain

$$\mathbb{P}\left[\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}(V_{k,\ell}) - \sum_{\ell=1}^{w/2} f_{k,\ell}(V_{k,\ell})\right| \le b_d\right] \tag{88a}$$

$$= \mathbb{P}\left[\left|2\sum_{\ell=w/2+1}^{w} f'_{k,\ell}(V_{k,\ell}) + 2\sum_{\ell=1}^{w/2} f'_{k,\ell}(V_{k,\ell}) - w\right| \le b_d\right] \tag{88b}$$

$$= \mathbb{P}\left[\left|\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) - \frac{w}{2}\right| \le \frac{b_d}{2}\right] \tag{88c}$$

$$= 1 - \mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) - \frac{w}{2} > \frac{b_d}{2}\right] - \mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) - \frac{w}{2} < -\frac{b_d}{2}\right]. \tag{88d}$$

$$= 1 - \mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) > \frac{w}{2} + \frac{b_d}{2}\right] - \mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) < \frac{w}{2} - \frac{b_d}{2}\right]. \tag{88e}$$

Specifically, Equation 88b follows from the definition of $f'_{k,\ell}$ in Equation 87. Next, by simple algebraic rearrangement, we obtain Equation 88c. Equation 88d is derived by decomposing the absolute value event into two separate one-sided probability terms, and Equation 88e then follows from further rearranging these terms.

By applying Lemma B.13 to the two probability terms above, we obtain

$$\mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) > \frac{w}{2} + \frac{b_d}{2}\right] \tag{89a}$$

$$\le \epsilon\|\varphi\|_d \exp\left(-\frac{b_d^2}{144L_k}\right), \tag{89b}$$

$$\mathbb{P}\left[\sum_{\ell=1}^{w} f'_{k,\ell}(V_{k,\ell}) < \frac{w}{2} - \frac{b_d}{2}\right] \tag{89c}$$

$$\le \epsilon\|\varphi\|_d \exp\left(-\frac{b_d^2}{144L_k}\right). \tag{89d}$$

Combining the above two inequalities, we have

$$\mathbb{P}\left[\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}(V_{k,\ell}) - \sum_{\ell=1}^{w/2} f_{k,\ell}(V_{k,\ell})\right| \le b_d\right] \ge 1 - 2\epsilon\|\varphi\|_d \exp\left(-\frac{b_d^2}{144L_k}\right), \tag{90}$$

where L_k is the mixing time of arm k.

If we set $b_d = \sqrt{144L\log\left(2KT^2\|\varphi\|_d^{-1}\right)}$, where $L = \max_k L_k$, then we have

$$\mathbb{P}\left[\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}(V_{k,\ell}) - \sum_{\ell=1}^{w/2} f_{k,\ell}(V_{k,\ell})\right| \le b_d\right] \ge 1 - \frac{\epsilon\|\varphi\|_d}{KT^2}. \tag{91}$$

Next, we bound the second probability term in Equation 86c. Since the noise terms n_ℓ are independent and bounded within $[-1, 1]$, we can apply McDiarmid's inequality to obtain:

$$\mathbb{P}\left[\left|\sum_{\ell=1}^{w} n_\ell\right| \le b_n\right] \ge 1 - \mathbb{P}\left[\left|\sum_{\ell=1}^{w} n_\ell\right| > b_n\right] \ge 1 - \mathbb{P}\left[\sum_{\ell=1}^{w} n_\ell > b_n\right] - \mathbb{P}\left[\sum_{\ell=1}^{w} n_\ell < -b_n\right] \ge 1 - 2\exp\left(-\frac{2b_n^2}{w}\right). \tag{92}$$

By setting $b_n = \sqrt{w\log\left(2KT^2\right)/2}$, we have

$$\mathbb{P}\left[\left|\sum_{\ell=1}^{w} n_\ell\right| \le b_n\right] \ge 1 - \frac{1}{KT^2}. \tag{93}$$

Combining the bounds for both probability terms, we obtain

$$p \leq 1 - \left(1 - \frac{\epsilon\|\varphi\|_d}{KT^2}\right)\left(1 - \frac{1}{KT^2}\right) \leq \frac{\epsilon\|\varphi\|_d + 1}{KT^2}. \tag{94}$$

Finally, we can bound the probability of false alarm as follows:

$$\begin{aligned}
\mathbb{P}\left(F_i \middle| \overline{F}_{i-1} D_{i-1}\right) &\leq \sum_{k=1}^{K} \mathbb{P}\left(\tau_{k,i} \leq \nu_i \mid \overline{F}_{i-1} D_{i-1}\right) && \text{(95a)}\\
&\leq \sum_{k=1}^{K} w\left(1 - \left(1 - \frac{\epsilon\|\varphi\|_d + 1}{KT^2}\right)^{\lfloor T/w \rfloor}\right) && \text{(95b)}\\
&\leq \sum_{k=1}^{K} w \cdot \frac{\epsilon\|\varphi\|_d + 1}{KT^2} \cdot \lfloor T/w \rfloor && \text{(95c)}\\
&\leq \frac{\epsilon\|\varphi\|_d + 1}{T}. && \text{(95d)}
\end{aligned}$$

Here, equation 95a follows from the union bound; equation 95b follows from the earlier derived bound on p; equation 95c uses the inequality $(1-x)^n \geq 1 - nx$ for $x \in [0,1]$ and $n \geq 1$; and equation 95d simplifies the expression to yield the final result. □

We next use the same concentration tool to establish that, with high probability, each true change is successfully detected within a sublinear delay.

Lemma B.15 (Probability of successful detection)**.** *Under Algorithm 2 with the parameter choices specified in equation 14 and equation 15, and with the delay threshold*

$$h_i = \left\lceil w\left(\frac{K}{2\alpha} + 1\right)\sqrt{s_i + 1} + \frac{w^2}{4}\left(\frac{K}{2\alpha} + 1\right)^2 \right\rceil, \tag{96}$$

for any $i \geq 1$, the probability of successful detection satisfies

$$\mathbb{P}\big(D_i \,\big|\, \overline{F}_i \overline{F}_{i-1} D_{i-1}\big) \;\geq\; 1 - \frac{\epsilon\|\varphi\|_d}{T}, \tag{97}$$

where $\epsilon > 0$ is an arbitrarily small constant, φ is the initial state distribution, and d is the stationary distribution.

Proof.

$$\begin{aligned}
\mathbb{P}\left(D_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) &= 1 - \mathbb{P}\left(\tau_i > \nu_i + h_i \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) && \text{(98a)}\\
&\geq \max_{t \in \{\nu_i+1, \ldots, \nu_i+h_i\}} \mathbb{P}\left(\exists k \in \mathcal{K}, |S_{k,t}| > b \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) && \text{(98b)}\\
&\geq \max_{j \in \{0,\ldots,w/2\}} \mathbb{P}\left(\left|\sum_{\ell=j+1}^{w} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right) - \sum_{\ell=1}^{j} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right)\right| > b \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) && \text{(98c)}\\
&\geq \mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right) - \sum_{\ell=1}^{w/2} \left(f_{k,\ell}\left(V_{k,\ell}\right) + n_\ell\right)\right| > b \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) && \text{(98d)}
\end{aligned}$$

Here, equation 98a follows from the definition of successful detection; equation 98b follows from the definition of the stopping time; equation 98c mirrors the derivation from equation 86a to equation 86b; and equation 98d follows from noting that the maximum is attained at $j = w/2$ and decomposing the absolute value event into two separate probability terms, one capturing the distributional deviation and the other the noise fluctuation.

First, we bound the first probability term in equation 98d.

$$\mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right)-\sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)\right|>b+b_d\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99a}$$

$$=\mathbb{P}\left(\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right)-\sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)>b+b_d\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99b}$$

$$+\mathbb{P}\left(\sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)-\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right)>b+b_d\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99c}$$

$$\geq\mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)-\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)>b+b_d-\frac{w}{2}\left(\bar{\mu}_k^{(i)}-\bar{\mu}_k^{(i-1)}\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99d}$$

$$+\mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)-\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)>b+b_d-\frac{w}{2}\left(\bar{\mu}_k^{(i-1)}-\bar{\mu}_k^{(i)}\right)\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99e}$$

$$\geq\mathbb{P}\left(\bar{\mu}_k^{(i)}>\bar{\mu}_k^{(i-1)}\right)\mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)<\frac{\left(\bar{\mu}_k^{(i)}-\bar{\mu}_k^{(i-1)}\right)w}{4}-\frac{(b+b_n)}{2}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99f}$$

$$\cdot\,\mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)>\frac{(b+b_n)}{2}-\frac{\left(\bar{\mu}_k^{(i)}-\bar{\mu}_k^{(i-1)}\right)w}{4}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99g}$$

$$+\mathbb{P}\left(\bar{\mu}_k^{(i)}<\bar{\mu}_k^{(i-1)}\right)\mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)<-\frac{\left(\bar{\mu}_k^{(i-1)}-\bar{\mu}_k^{(i)}\right)w}{4}-\frac{(b+b_n)}{2}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99h}$$

$$\cdot\,\mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)>\frac{(b+b_n)}{2}+\frac{\left(\bar{\mu}_k^{(i-1)}-\bar{\mu}_k^{(i)}\right)w}{4}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99i}$$

$$\geq\mathbb{P}\left(\bar{\mu}_k^{(i)}>\bar{\mu}_k^{(i-1)}\right)\mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)<\frac{\delta w}{4}-\frac{(b+b_n)}{2}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99j}$$

$$\cdot\,\mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)>\frac{(b+b_n)}{2}-\frac{\delta w}{4}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99k}$$

$$+\mathbb{P}\left(\bar{\mu}_k^{(i)}<\bar{\mu}_k^{(i-1)}\right)\mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i)}\right)<-\frac{\delta w}{4}-\frac{(b+b_n)}{2}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99l}$$

$$\cdot\,\mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right)-\bar{\mu}_k^{(i-1)}\right)>\frac{(b+b_n)}{2}+\frac{\delta w}{4}\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right) \tag{99m}$$

Here, equation 99b and equation 99c follow from decomposing the absolute value event into two separate one-sided probability terms; equation 99d and equation 99e follow from rearranging the terms and adding and subtracting the respective means; equation 99f and equation 99h follow from conditioning on whether the mean reward has increased or decreased; equation 99g and equation 99i follow from rearranging the terms; and equation 99j and equation 99l follow from the definition of the minimum change size δ.

Let $\zeta_k^{(i)} = \frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i-1)}}$ and $\zeta_k^{(i-1)} = \frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i-1)}}$. Then, equation 99j–equation 99m become

$$\mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right) - \sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)\right| > b + b_d \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100a}$$

$$\geq \mathbb{P}\left(\bar{\mu}_k^{(i)} > \bar{\mu}_k^{(i-1)}\right) \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) < \frac{1}{2}\left(\frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i-1)}}\right)\bar{\mu}_k^{(i-1)} w \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100b}$$

$$\cdot \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) > -\frac{1}{2}\left(\frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i)}}\right)\bar{\mu}_k^{(i)} w \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100c}$$

$$+ \mathbb{P}\left(\bar{\mu}_k^{(i)} < \bar{\mu}_k^{(i-1)}\right) \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) < \frac{1}{2}\left(\frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i)}}\right)\bar{\mu}_k^{(i)} w \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100d}$$

$$\cdot \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) > -\frac{1}{2}\left(\frac{\delta/2-(b+b_n)/w}{\bar{\mu}_k^{(i-1)}}\right)\bar{\mu}_k^{(i-1)} w \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100e}$$

$$\geq \mathbb{P}\left(\bar{\mu}_k^{(i)} > \bar{\mu}_k^{(i-1)}\right) \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) < \frac{\zeta_k^{(i-1)}\bar{\mu}_k^{(i-1)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100f}$$

$$\cdot \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) > -\frac{\zeta_k^{(i)}\bar{\mu}_k^{(i)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100g}$$

$$+ \mathbb{P}\left(\bar{\mu}_k^{(i)} < \bar{\mu}_k^{(i-1)}\right) \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) < \frac{\zeta_k^{(i)}\bar{\mu}_k^{(i)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100h}$$

$$\cdot \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) > -\frac{\zeta_k^{(i-1)}\bar{\mu}_k^{(i-1)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{100i}$$

Here, equation 100a–equation 100g follow from substituting the definitions of $\zeta_k^{(i)}$ and $\zeta_k^{(i-1)}$; and equation 100h and equation 100i follow from rearranging the terms. Before applying Lemma B.13, we rewrite the above as

$$\mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right) - \sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)\right| > b + b_d \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right) \tag{101a}$$

$$\geq \mathbb{P}\left(\bar{\mu}_k^{(i)} > \bar{\mu}_k^{(i-1)}\right)\left(1 - \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) \geq \frac{\left(1+\zeta_k^{(i-1)}\right)\bar{\mu}_k^{(i-1)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)\right) \tag{101b}$$

$$\cdot \left(1 - \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) \leq \frac{\left(1-\zeta_k^{(i)}\right)\bar{\mu}_k^{(i)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)\right) \tag{101c}$$

$$+ \mathbb{P}\left(\bar{\mu}_k^{(i)} < \bar{\mu}_k^{(i-1)}\right)\left(1 - \mathbb{P}\left(\sum_{\ell=w/2+1}^{w}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i)}\right) \geq \frac{\left(1+\zeta_k^{(i)}\right)\bar{\mu}_k^{(i)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)\right) \tag{101d}$$

$$\cdot \left(1 - \mathbb{P}\left(\sum_{\ell=1}^{w/2}\left(f_{k,\ell}\left(V_{k,\ell}\right) - \bar{\mu}_k^{(i-1)}\right) \leq \frac{\left(1-\zeta_k^{(i-1)}\right)\bar{\mu}_k^{(i-1)} w}{2} \middle| \overline{F}_i \overline{F}_{i-1} D_{i-1}\right)\right) \tag{101e}$$

Here, equation 101a–equation 101d follow from rearranging the terms. Now, we can apply Lemma B.13 to bound

the probability terms in equation 101b–equation 101e. For example, for the term in equation 101b, we have

$$\mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right)-\sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)\right|>b+b_d\left|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right.\right) \tag{102a}$$

$$\geq \mathbb{P}\left(\bar{\mu}_k^{(i)}>\bar{\mu}_k^{(i-1)}\right)\left(1-\epsilon\|\varphi\|_d\exp\left(-\frac{(c/2-b_n/w)^2}{144L}\right)\right)^2 \tag{102b}$$

$$+\mathbb{P}\left(\bar{\mu}_k^{(i)}<\bar{\mu}_k^{(i-1)}\right)\left(1-\epsilon\|\varphi\|_d\exp\left(-\frac{(c/2-b_n/w)^2}{144L}\right)\right)^2 \tag{102c}$$

$$=\left(1-\epsilon\|\varphi\|_d\exp\left(-\frac{(c/2-b_n/w)^2}{144L}\right)\right)^2 \tag{102d}$$

$$\tag{102e}$$

Finally, to satisfy the inequality $\delta_k^{(i)}\geq 2b/w+c$, we set w and b as in equation 14 and equation 15, and choose $c=2\left(\sqrt{144L\ln T/w}+\sqrt{\ln(2KT^2)/2w}\right)$, which leads to

$$\mathbb{P}\left(\left|\sum_{\ell=w/2+1}^{w} f_{k,\ell}\left(V_{k,\ell}\right)-\sum_{\ell=1}^{w/2} f_{k,\ell}\left(V_{k,\ell}\right)\right|>b+b_d\left|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right.\right)\geq 1-\frac{\epsilon\|\varphi\|_d}{T}. \tag{103}$$

The second probability term in equation 98d can be bounded similarly as in equation 93, the successful detection probability is then obtained by substituting equation 103 and equation 93 back into equation 98d.

$$\mathbb{P}\left(D_i\middle|\overline{F}_i\overline{F}_{i-1}D_{i-1}\right)\geq 1-\frac{\epsilon\|\varphi\|_d}{T}. \tag{104}$$

□

Lemma B.16 further bounds the expected detection delay in the situation where the change detection algorithm successfully detects the change within the desired interval.

Lemma B.16 (Expected detection delay)**.** *Consider a piecewise-stationary restless bandit environment. For any $\bar{\boldsymbol{\mu}}^{(i)},\bar{\boldsymbol{\mu}}^{(i+1)}\in[0,1]^K$ with parameters chosen in equation 3 and equation 4 and*

$$h_i=\left\lceil w\left(\frac{K}{2\alpha}+1\right)\sqrt{s_i+1}+\frac{w^2}{4}\left(\frac{K}{2\alpha}+1\right)^2\right\rceil, \tag{105}$$

for some $k\in\mathcal{K}, i\geq 1$ and $c>0$, under the Algorithm 1, we have

$$\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_iD_i\overline{F}_{i-1}D_{i-1}\right]\leq h_i. \tag{106}$$

Proof. For any $1\leq i\leq M$, we have

$$\mathbb{E}\left[\tau_i-\nu_i\middle|\overline{F}_iD_i\overline{F}_{i-1}D_{i-1}\right]=\sum_{j=1}^{h_i}\mathbb{P}\left(\tau_i\geq\nu_i+j\middle|\overline{F}_iD_i\overline{F}_{i-1}D_{i-1}\right)\leq h_i. \tag{107a}$$

□

Plugging the bounds in Lemmas B.14, B.15 and B.7 into Theorem B.4 shows the following regret bound in Corollary B.17.

Corollary B.17. *Combining Algorithm 1 (diminishing exploration with resets) with the base solver $\mathcal{B}$ and Algorithm 2 with parameters (w,b) given in Equations 3 and 4, the excess regret is upper-bounded as*

$$\mathcal{R}_{\text{excess}}(T)\leq 2\alpha\sqrt{MT}+w\left(\frac{K}{2\alpha}+1\right)\sqrt{M\left(T+M\right)}+\frac{w^2M}{4}\left(\frac{K}{2\alpha}+1\right)^2+2\epsilon\|\varphi\|_d+1. \tag{108}$$

By setting $\alpha=c\sqrt{KL\log\left(KT\right)}$ for some constant c, the expected regret is upper-bounded by $\mathcal{O}(\sqrt{KLMT\log T})$.

C ALGORITHMS AND PARAMETERS TUNING

In this appendix, we provide an explanation of our parameter selection.

Algorithm Setting. For change detection mechanism, the window size w is set to $\left\lfloor 100\sqrt{144L} \right\rfloor$, where L is the mixing time for $\epsilon = 1/8$, the threshold $b = \sqrt{w/2 \ln(2KT^2)} + \sqrt{144wL \ln(2KT^2)}$. For uniform sampling scheme, the sample rate $\alpha = \sqrt{M/T \log(T/M)}$ is following the setting in Liu et al. (2018). For diminishing exploration scheme, the sampling parameter $\alpha = 1$. In RestlessUCB algorithm, the minimum sampling amount is set as $m(T) = 100$, and the confidence radius $rad(T) = \sqrt{T/(2Mm(T))}$. For the Colored-UCRL2 algorithm, $\delta = 1/\min(10000, T)$ and $B = S$, following the work in Ortner et al. (2012). To improve computational efficiency, we restrict the value of δ so that the value iteration can run faster without significant loss of accuracy.

Environment Setting. We consider an experiment consisting of three Markov chains (arms), indexed by $k \in \{1, 2, 3\}$, each evolving over five segments, indexed by $i \in \{1, 2, 3, 4, 5\}$. The transitions between states in each Markov chain are governed by transition probability matrices, and each state is associated with a reward mean function.

Markov Chain Transition Kernels. The transition kernel for each segment i is given by a set of transition matrices $P^{(i,k)}$, where each row represents the probability distribution over the next states, given the current state. For each segment i, we define:

$$P^{(i)} = \begin{bmatrix} P_1^{(i)} & P_2^{(i)} & P_3^{(i)} \end{bmatrix} \tag{109}$$

where $P_k^{(i)}$ represents the transition matrix for arm k.

The transition matrices for each segment are:

$$P^{(1)} = \left[\begin{bmatrix} 0.50 & 0.50 & 0.00 \\ 0.17 & 0.66 & 0.17 \\ 0.00 & 0.50 & 0.50 \end{bmatrix}, \begin{bmatrix} 0.36 & 0.64 & 0.00 \\ 0.39 & 0.22 & 0.39 \\ 0.00 & 0.50 & 0.50 \end{bmatrix}, \begin{bmatrix} 0.55 & 0.45 & 0.00 \\ 0.43 & 0.36 & 0.21 \\ 0.00 & 0.50 & 0.50 \end{bmatrix} \right] \tag{110}$$

$$P^{(2)} = \left[\begin{bmatrix} 0.43 & 0.57 & 0.00 \\ 0.45 & 0.30 & 0.25 \\ 0.00 & 0.64 & 0.36 \end{bmatrix}, \begin{bmatrix} 0.33 & 0.67 & 0.00 \\ 0.32 & 0.42 & 0.26 \\ 0.00 & 0.50 & 0.50 \end{bmatrix}, \begin{bmatrix} 0.50 & 0.50 & 0.00 \\ 0.11 & 0.47 & 0.42 \\ 0.00 & 0.67 & 0.33 \end{bmatrix} \right] \tag{111}$$

$$P^{(3)} = \left[\begin{bmatrix} 0.62 & 0.38 & 0.00 \\ 0.50 & 0.44 & 0.06 \\ 0.00 & 0.62 & 0.38 \end{bmatrix}, \begin{bmatrix} 0.50 & 0.50 & 0.00 \\ 0.43 & 0.21 & 0.36 \\ 0.00 & 0.38 & 0.62 \end{bmatrix}, \begin{bmatrix} 0.17 & 0.83 & 0.00 \\ 0.50 & 0.28 & 0.22 \\ 0.00 & 0.44 & 0.56 \end{bmatrix} \right] \tag{112}$$

$$P^{(4)} = \left[\begin{bmatrix} 0.57 & 0.43 & 0.00 \\ 0.20 & 0.47 & 0.33 \\ 0.00 & 0.57 & 0.43 \end{bmatrix}, \begin{bmatrix} 0.60 & 0.40 & 0.00 \\ 0.18 & 0.47 & 0.35 \\ 0.00 & 0.50 & 0.50 \end{bmatrix}, \begin{bmatrix} 0.53 & 0.47 & 0.00 \\ 0.20 & 0.20 & 0.60 \\ 0.00 & 0.73 & 0.27 \end{bmatrix} \right] \tag{113}$$

$$P^{(5)} = \left[\begin{bmatrix} 0.38 & 0.62 & 0.00 \\ 0.31 & 0.13 & 0.56 \\ 0.00 & 0.56 & 0.44 \end{bmatrix}, \begin{bmatrix} 0.55 & 0.45 & 0.00 \\ 0.45 & 0.46 & 0.09 \\ 0.00 & 0.64 & 0.36 \end{bmatrix}, \begin{bmatrix} 0.33 & 0.67 & 0.00 \\ 0.56 & 0.25 & 0.19 \\ 0.00 & 0.40 & 0.60 \end{bmatrix} \right] \tag{114}$$

Reward Mean Functions. The rewards are randomly generated based on the current state in each segment, denoted as $R_k^{(i)}(s)$, where $k = 1, 2, 3$ represents the arm, $i = 1, 2, 3, 4, 5$ represents the segment, and $s = 1, 2, 3$ represents the state.

$$R^{(i)} = \begin{bmatrix} R_1^{(i)}(1) & R_1^{(i)}(2) & R_1^{(i)}(3) \\ R_2^{(i)}(1) & R_2^{(i)}(2) & R_2^{(i)}(3) \\ R_3^{(i)}(1) & R_3^{(i)}(2) & R_3^{(i)}(3) \end{bmatrix} \tag{115}$$

The values for each segment are:

$$R^{(1)} = \begin{bmatrix} 0.24090 & 0.21567 & 0.11303 \\ 0.68377 & 0.55163 & 0.29024 \\ 0.88837 & 0.87993 & 0.40863 \end{bmatrix} \quad R^{(2)} = \begin{bmatrix} 0.76023 & 0.42959 & 0.15431 \\ 0.95580 & 0.81347 & 0.63100 \\ 0.37325 & 0.24685 & 0.06446 \end{bmatrix} \tag{116}$$

$$R^{(3)} = \begin{bmatrix} 0.83859 & 0.80774 & 0.19539 \\ 0.29037 & 0.23361 & 0.08248 \\ 0.95163 & 0.45959 & 0.03668 \end{bmatrix} \quad R^{(4)} = \begin{bmatrix} 0.52273 & 0.14461 & 0.03640 \\ 0.67324 & 0.54894 & 0.31872 \\ 0.93209 & 0.88988 & 0.62226 \end{bmatrix} \tag{117}$$

$$R^{(5)} = \begin{bmatrix} 0.64763 & 0.56596 & 0.36023 \\ 0.87041 & 0.82968 & 0.08825 \\ 0.22632 & 0.21085 & 0.13084 \end{bmatrix} \tag{118}$$

Mixing Time Calculation. The mixing time L is calculated based on the transition matrices of the Markov chains. For each arm k in segment i, we compute the second largest eigenvalue modulus (SLEM) of the transition matrix $P^{(i,k)}$. The mixing time is then determined using the formula:

$$L^{(i,k)} = \frac{\ln(1/\epsilon)}{1 - \lambda_2^{(i,k)}} \tag{119}$$

where $\lambda_2^{(i,k)}$ is the SLEM of $P^{(i,k)}$ and ϵ is the desired accuracy level (set to 1/8 in our experiments). The overall mixing time L used in the algorithm is the maximum mixing time across all arms and segments:

$$L = \max_{i,k} L^{(i,k)} \tag{120}$$

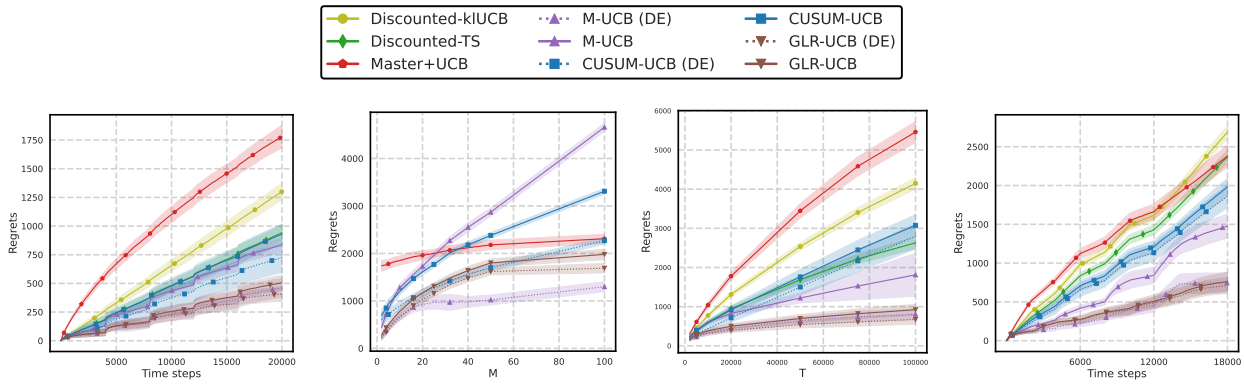
D ONE-STATE SPECIAL CASE SIMULATION RESULTS

In this appendix, we assess the effectiveness of the proposed diminishing exploration scheme across various dimensions, encompassing regret scaling in M, K, and T, regrets in synthetic environments, and regrets in a real-world scenario. In addition to evaluating M-UCB (Cao et al., 2019) with our diminishing exploration, we also examine a variant of CUSUM-UCB (Liu et al., 2018) that incorporates diminishing exploration with CUSUM-UCB, further highlighting the efficacy of the proposed exploration method. We will compare our approach with M-UCB, CUSUM-UCB, GLR-UCB, Discounted-UCB, Discounted Thompson Sampling, and MASTER. Unless stated otherwise, we report the average regrets over 100 simulation trials. Detailed configuration is provided in Appendix E.

Regret in Each Time Step. In this simulation, we consider a multi-armed bandit problem with $T = 20000$ time steps and $M = 5$. Recall that $\mu_k^{(i)}$ represents the expected value for arm k in the i-th segment. Here, we set $\mu_k^{(i)} = 0.2, 0.5, 0.8$ for i with $(i + k) \bmod 3 = 2, 0, 1$, respectively. Figure 4a shows that for both CUSUM-UCB and M-UCB, employing diminishing exploration can effectively reduce the additional regret caused by constant exploration. Moreover, M-UCB with the proposed diminishing exploration achieves the lowest regret. In the figure, the change points are clearly evident by observing the breakpoints in each line. The reason for the overall steeper slope of CUSUM-UCB (both with and without diminishing exploration) is due to the heightened sensitivity of the CUSUM detector itself, resulting in more frequent false alarms.

Regret Scaling in M. Based on the settings outlined above, with the only variation being in the parameter M, Figure 4b illustrates the dynamic regrets across various values of M. In this experiment, adjustments to the exploration parameter settings are required based on the size of M when using a constant exploration rate. However, this is not the case for the proposed diminishing exploration. The result confirms the earlier-discussed rationale that the proposed diminishing exploration can automatically adapt to the environment, resulting in the best regret performance among the algorithms.

Regret Scaling in T. In line with the setting described above, with the only variation being the parameter T, we present the dynamic regrets across different values of T. Observations similar to those made above can again be found in Figure 4c, where the proposed diminishing exploration can effectively reduce the regret.



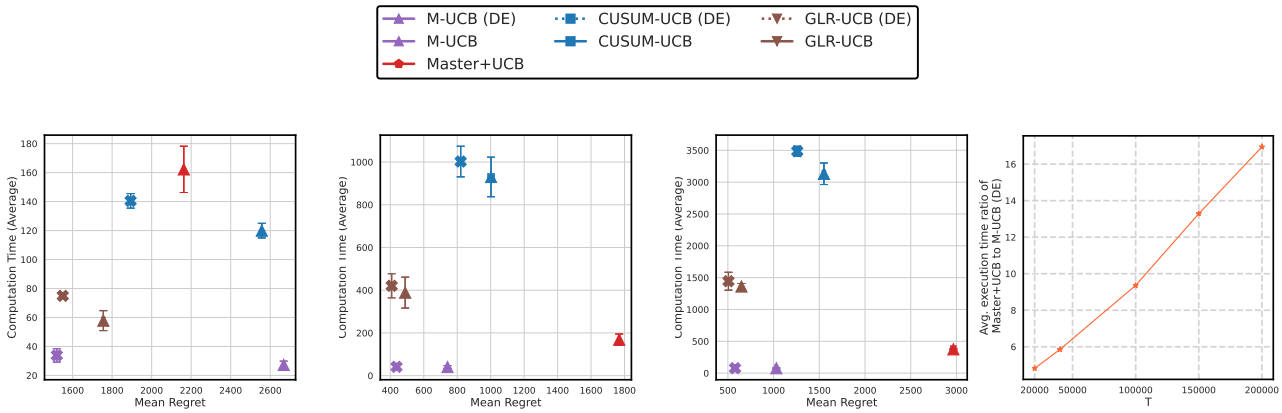
(a) Scaling in t. (b) Scaling in M. (c) Scaling in T. (d) Yahoo Data Set.

Figure 4: Regret in the synthetic environments and under the Yahoo data set.

Regret and Execution Time. Here, we compare the computation time and regret across different algorithms for various choices of M and T with other parameters same as above. Specifically, we conduct experiments under three scenarioso: one where the environment changes rapidly ($M = 50$ and $T = 20000$), one where the environment changes slowly ($M = 5$ and $T = 20000$), and one where the considered time horizon is double ($M = 5$ and $T = 40000$). As shown in Figure 5a to 5c, despite oblivious of M, our algorithm almost always achieves the lowest computation time and regret in all the scenarios. Moreover, comparing to Master+UCB, another algorithm not requiring the knowledge of M, our algorithm is always significantly better as shown in these figures. Figure 5d plots the ratio of average execution time of Master+UCB to that of M-UCB with our DE for various T. It is shown that the growth rate is faster than $0.5 \log T$, and it appears to become even linear in T as T increases.

Regret in an Environment Built from a Real-World Dataset. We further utilize the benchmark dataset publicly published by Yahoo! for evaluation. To enhance arm distinguishability in our simulation, we scale up the data by a factor of 10. The number of segments is set to $M = 9$ and the number of arms is set to $K = 6$. Figure 4d shows the evolution of dynamic regret. Again, we see that the diminishing exploration scheme could help M-UCB, GLR-UCB and CUSUM-UCB achieve similar or better regret even without knowing M.

Scenarios when Assumptions are Violated. In Assumption A.1, we assumes the knowledge of δ to select an appropriate w. We emphasize that our settings in many of the above simulations actually violate this assumption. Take Figure 4a for example, $w = 200$ is chosen, which corresponds to $\delta \approx 0.6$ when back calculating, which is much larger than the actual $\delta = 0.3$ of this scenario. Assumptions A.2 provide guarantees that the segment length is sufficiently long. However, in the last data point of our Figure 4b ($M = 100$), these assumptions are clearly violated. In this case, it becomes challenging for the active methods to promptly detect every change point. For algorithms like M-UCB and CUSUM-UCB, which require knowledge of M, their awareness of quick changes causes them to invest more effort into change detection, leading to a very high exploration rate. However, this does not always guarantee successful detection, resulting in very high regret. Diminishing exploration, on the other hand, continues to decrease the exploration rate regardless of M when no changes are detected. This allows more resources to be invested for UCB, which might gradually adapt to the environment's changes, leading to a regret that is lower than that of uniform exploration.



(a) $M = 50$, $T = 20000$. (b) $M = 5$, $T = 20000$. (c) $M = 5$, $T = 40000$. (d) Ratio of avg. execution time of Master+UCB to that of M-UCB (DE).

Figure 5: Regret and computation times.

E ONE-STATE SPECIAL CASE ALGORITHMS AND PARAMETERS TUNING

In this appendix, we provide an explanation of our parameter selection. For M-UCB, the window size w is set to 200 unless otherwise specified; however, for the last data point ($M = 100$) in Figure 4b, we chose $w = 50$ due to the limitations inherent to change detection. We compute the change detection threshold $b_{\text{M-UCB}} = \sqrt{w/2 \log(2KT^2)}$ following the original formulation in Cao et al. (2019). Additionally, the uniform exploration rate $\gamma_{\text{M-UCB}} = \sqrt{MK \log T/T}$ is determined as initially stated in Besson et al. (2022).Concerning CUSUM-UCB, we adhere to Liu et al. (2018) by fixing $\epsilon = 0.1$, setting the change detection threshold $b_{\text{CUSUM-UCB}} = \log(T/M - 1)$, and establishing the uniform exploration rate $\gamma_{\text{CUSUM-UCB}} = \sqrt{MK \log T/T}$ as initially stated in Besson et al. (2022). Additionally, in CUSUM-UCB, the change point detection involves averaging the first H samples, where H is set to 100. For GLR-UCB (Besson et al., 2022), we set $\gamma_{m,\text{GLR-UCB}} = \sqrt{mK \log T/T}$, where m is the number of alarms. We utilize the threshold function $\beta(n, \delta) = \log\left(n^{3/2}/\delta\right)$ and set $\delta = 1/\sqrt{T}$. In our setup, for both the diminishing versions of M-UCB and CUSUM-UCB, we follow the parameter selection approach described earlier, except for the choice of the exploration rate. In this context, we opt for $\alpha = 1$. For passive methods, including DUCB (Garivier and Moulines, 2011) and DTS (Qi et al., 2023), we use a discounting factor $\gamma = 0.75$. MASTER, on the other hand, follows the theoretical settings outlined in Wei and Luo (2021) and is categorized as an active method.

To evaluate the scalability of our methods, we conducted three sets of scaling experiments. For scaling in t (Figure 4a), scaling in M (Figure 4b), and scaling in T (Figure 4c), The parameters of DUCB, DTS and MASTER follow the theoretical settings outlined in Garivier and Moulines (2011), Qi et al. (2023) and Wei and Luo (2021), respectively.

Table 1: Parameter Selection for Active and Passive Methods

Method	Parameters	References
M-UCB	- Window size $w = 200$ (default), $w = 50$ for $M = 100$ (Figure 4b). - Threshold: $b_{\text{M-UCB}} = \sqrt{w/2 \log(2KT^2)}$. - Exploration rate: $\gamma_{\text{M-UCB}} = \sqrt{MK \log T/T}$.	Cao et al. (2019),
CUSUM-UCB	- Fixed $\epsilon = 0.1$. - Threshold: $b_{\text{CUSUM-UCB}} = \log(T/M - 1)$. - Exploration rate: $\gamma_{\text{CUSUM-UCB}} = \sqrt{MK \log T/T}$. - Change detection based on averaging first $H = 100$ samples.	Liu et al. (2018),
GLR-UCB	- Exploration rate: $\gamma_{m,\text{GLR-UCB}} = \sqrt{mK \log T/T}$. - Threshold function: $\beta(n, \delta) = \log\left(n^{3/2}/\delta\right)$, $\delta = 1/\sqrt{T}$.	Besson et al. (2022)
MASTER	- All parameters follow theoretical settings.	Wei and Luo (2021)
DUCB (Passive)	- Discounting factor $\gamma = 0.75$.(Figure 4a and 4d) - Discounting factor follows theoretical setting. (Figure 4b and 4c)	Garivier and Moulines (2011)
DTS (Passive)	- Discounting factor $\gamma = 0.75$. (Figure 4a and 4d) - Discounting factor follows theoretical setting. (Figure 4b and 4c)	Qi et al. (2023)