Received 29 September 2023; revised 12 January 2024 and 12 February 2024; accepted 19 February 2024. Date of publication 5 March 2024; date of current version 3 April 2024. Recommended by Senior Editor Prof. M. Alizadeh.

Digital Object Identifier 10.1109/OJCSYS.2024.3372929

Nonstationary Stochastic Bandits: UCB Policies and Minimax Regret

LAI WEI [1] AND VAIBHAV SRIVASTAVA [2] (Senior Member, IEEE)

[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109 USA
[2]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA

CORRESPONDING AUTHOR: LAI WEI (e-mail: weilaitim@gmail.com)

This work was supported in part by USA National Science Foundation under Grant IIS-1734272 and in part by ONR under Grant N00014-22-1-2813.

ABSTRACT We study the nonstationary stochastic Multi-Armed Bandit (MAB) problem in which the distributions of rewards associated with arms are assumed to be time-varying and the total variation in the expected rewards is subject to a variation budget. The regret of a policy is defined by the difference in the expected cumulative reward obtained using the policy and using an oracle that selects the arm with the maximum mean reward at each time. We characterize the performance of the proposed policies in terms of the worst-case regret, which is the supremum of the regret over the set of reward distribution sequences satisfying the variation budget. We design Upper-Confidence Bound (UCB)-based policies with three different approaches, namely, periodic resetting, sliding observation window, and discount factor, and show that they are order-optimal with respect to the minimax regret, i.e., the minimum worst-case regret achieved by any policy. We also relax the sub-Gaussian assumption on reward distributions and develop robust versions of the proposed policies that can handle heavy-tailed reward distributions and maintain their performance guarantees.

INDEX TERMS Heavy-tailed distributions, minimax regret, nonstationary multiarmed bandit, upper-confidence bound, variation budget.

I. INTRODUCTION

Uncertainty and nonstationarity of the environment are two of the major barriers to decision-making problems across scientific disciplines, including engineering, economics, social science, neuroscience, and ecology. An efficient strategy in such environments requires balancing several tradeoffs, including *exploration-versus-exploitation*, i.e., choosing between the most informative and the empirically most rewarding alternatives, and *remembering-versus-forgetting*, i.e., using more but possibly outdated information or using less but recent information.

The stochastic MAB problem is a canonical formulation of the exploration-versus-exploitation tradeoff. In an MAB problem, an agent selects one from K options at each time and receives a reward associated with it. The reward sequence at each option is assumed to be an unknown i.i.d random process. The MAB formulation has been applied in many scientific and technological areas. For example, it is used for opportunistic spectrum access in communication networks, wherein the arm models the availability of a channel [1], [2]. In MAB formulation of online learning for demand response [3], [4], an aggregator calls upon a subset of users (arms) who have an unknown response to the request to reduce their loads. MAB formulation has also been used in robotic foraging and surveillance [5], [6], [7], [8] and acoustic relay positioning for underwater communication [9], wherein the information gain at different sites is modeled as rewards from arms. Besides, contextual bandits are widely used in recommender systems [10], [11], wherein the acceptance of a recommendation corresponds to the rewards from an arm. The stationarity assumption in classic MAB problems limits their utility in these applications since channel usage, the robot's working environment, and individual preferences are inherently uncertain and evolving. In this paper, we relax this assumption and study nonstationary stochastic MAB problems.

Robbins [12] formulated the objective of the stochastic MAB problem as minimizing the regret, that is, the loss in expected cumulative rewards caused by failing to select the best

arm every time. In their seminal work, Lai and Robbins [13], followed by Burnetas and Katehakis [14], established a logarithm *problem-dependent* asymptotic lower bound on the regret achieved by any policy, which has a leading constant determined by the underlying reward distributions. A general method of constructing UCB rules for parametric families of reward distributions is also presented in [13], and the associated policy is shown to attain the logarithm lower bound. Several subsequent UCB-based algorithms [15], [16] with efficient finite time performance have been proposed.

The adversarial MAB [17] is a paradigmatic nonstationary problem. In this model, the bounded reward sequence at each arm is arbitrary without any probabilistic model. The performance of a policy is evaluated using the *weak regret*, which is the difference in the cumulated reward of a policy compared to the best single action policy. A $\Omega(\sqrt{KT})$[1] lower bound on the weak regret and a near-optimal policy Exp3 is also presented in [17]. While being able to capture nonstationarity, the generality of the reward model in adversarial MAB makes the investigation of globally optimal policies very challenging.

The nonstationary stochastic MAB can be viewed as a compromise between stationary stochastic MAB and adversarial MAB. It maintains the stochastic nature of the reward sequence while allowing some degree of nonstationarity in reward distributions. Instead of the weak regret analyzed in adversarial MAB, a strong notion of regret defined regarding the best arm at each time step is studied in these problems. A broadly studied nonstationary problem is *piecewise stationary* MAB, wherein the reward distributions are piecewise stationary. To deal with the remembering-versus-forgetting tradeoff, it is proposed to use a discount factor in the computation of UCB in [19]. Garivier and Moulines [20] present and analyze Discounted UCB (D-UCB) and sliding window UCB (SW-UCB), in which they compute the UCB using discounted sampling history and recent sampling history, respectively. They pointed out that if the number of change points N_T is available, both algorithms can be tuned to achieve regret close to the $\Omega(\sqrt{KN_TT})$ regret lower bound. These approaches have been employed to design payment routing policies to maximize the transaction success rate [21]. In our earlier work [22], the near-optimal regret is achieved using deterministic sequencing of explore and exploit with limited memory. Other works address the change of reward distributions adaptively with change point detection [23], [24], [25], [26], [27], which has been extended to handle nonstationary representations in linear bandits [28].

In another line of nonstationary bandit research, the expected rewards are assumed to vary according to stochastic processes such as Brownian motion [29] and stochastic linear dynamical system [30]. A more general nonstationary problem is studied in [31], wherein the cumulative maximum variation in mean rewards is subject to a variation budget V_T. Additionally, the authors in [31] establish a $\Omega((KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$ minimax regret lower bound and propose the Rexp3 policy. In their subsequent work [32], they tune Exp3.S policy from [17] to achieve near-optimal worst-case regret. Discounted Thomson Sampling (DTS) [33] has also been shown to have good experimental performance within this general framework. However, we are not aware of any analytic regret bounds for the DTS algorithm.

Most nonstationary MAB algorithms require information about the nonstationarity of the environment to tune their parameters such as the sliding window size or discounting factor. This is also the case for this work since we assume the variation density V_T/T in the general nonstationary environment is known. More recently, parameter-free algorithms that adaptively tune parameters by actively detecting the nonstationarity of the environment have been proposed. In [34], ADSWITCH randomly assigns the change detection tasks at different arms to update the estimate of N_T. Once a change in mean reward is detected, ADSWITCH initializes a new episode and runs the algorithm with the new estimate of N_T. In [35], a similar approach is applied to the contextual bandit problem in the general nonstationary setting with a variation budget. Cheung et al. [36] adopt an alternative approach in which several copies of SW-UCB with different window sizes are computed, and a master bandit algorithm is used to manage these copies. Though the parameter-free problem is appealing, it remains unknown whether the minimax regret lower bound $\Omega((KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$ retains its tightness without the information on environment nonstationarity.

In this paper, we adopt the variation budget formulation of nonstationary stochastic MAB problem [31] and design UCB-based policies that achieve efficient performance. Besides the commonly studied environments with sub-Gaussian rewards, we extend our algorithms to handle environments with heavy-tailed rewards. Such heavy-tailed rewards are common in many domains such as social networks [37] and financial markets [38]. Our UCB-based policies relax the assumption of the bounded reward in the Exp3-type policies in the literature and incur smaller variance in the cumulative reward [17]. Especially in the heavy-tailed reward environment, Exp3-type policies, in general, fail to maintain consistent performance. We show that by using a robust mean estimator, the UCB-based policies for light-tailed rewards can be modified to handle heavy-tailed rewards.

Our algorithms are inspired by the ideas of periodic resetting [31], sliding window, and discounting factor [20] but feature some key differences in algorithm design and analysis. Our algorithms leverage the MOSS algorithm [39] that has been designed for stationary environments. For stationary environments, in contrast to algorithms such as UCB1, the upper bound on the minimax regret for the MOSS algorithm does not feature any extraneous logarithmic term compared with the associated lower bound. In this work, we establish similar results for nonstationary environments. The major contributions of this work are the following:

[1]We use Bachmann–Landau asymptotic notation [18] $\Omega(\cdot)$ and $\mathcal{O}(\cdot)$ to highlight the dominating term of the lower and upper bounds, respectively.

- We establish that for a known variation density V_T/T, the order-optimal $\mathcal{O}((KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$ minimax regret for the nonstationary stochastic MAB problems can be achieved by UCB policies. The algorithms are designed by extending MOSS [39] to Resetting MOSS (R-MOSS) and Sliding Window MOSS (SW-MOSS).
- We show that D-UCB, which is more memory-efficient than R-MOSS and SW-MOSS, can be tuned to achieve a near-optimal worst-case regret $\mathcal{O}(\ln(T)(KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$. Based on the regret analysis of D-UCB, we qualitatively explain why the factor $\ln(T)$ cannot be removed.
- We relax the bounded support or sub-Gaussian assumption on the rewards required in existing works [20], [31], [32] and design robust UCB policies that can handle heavy-tailed reward distributions. We demonstrate that our robust policies achieve the same order-optimal minimax regret as the light-tailed reward setup.
- We numerically compare the proposed algorithms with several state-of-the-art algorithms. For light-tailed reward distributions and a certain class of environments, we show that R-MOSS and SW-MOSS perform the best. For heavy-tailed rewards, we show that our robust policies yield regrets with smaller mean and variance.

The remainder of the paper is organized as follows. We formulate nonstationary stochastic MAB with variation budget in Section II and review some preliminaries in Section III. In Section IV, we present and analyze three UCB policies: R-MOSS, SW-MOSS, and D-UCB. We present and analyze algorithms for nonstationary heavy-tailed bandit in Section V. We complement the theoretical results with numerical illustrations in Section VI and conclude this work in Section VII.

II. PROBLEM FORMULATION

We consider a nonstationary stochastic MAB problem with K arms and a horizon length T. Let $\mathcal{K} := \{1, \ldots, K\}$ be the set of arms and $\mathcal{T} := \{1, \ldots, T\}$ be the sequence of time slots. The reward sequence $\{X_t^k\}_{t\in\mathcal{T}}$ for each arm $k \in \mathcal{K}$ is composed of independent samples from potentially time-varying probability distribution function sequence $f_{\mathcal{T}}^k := \{f_t^k(x)\}_{t\in\mathcal{T}}$. The set of reward distribution sequences at all arms $\mathcal{F}_T^{\mathcal{K}} = \{f_{\mathcal{T}}^k \mid k \in \mathcal{K}\}$ is referred to as *environment*. Let $\mu_t^k = \mathbb{E}[X_t^k]$. We define the *total variation* of $\mathcal{F}_T^{\mathcal{K}}$ as

$$v\left(\mathcal{F}_T^{\mathcal{K}}\right) := \sum_{t=1}^{T-1} \max_{k\in\mathcal{K}} \left|\mu_{t+1}^k - \mu_t^k\right|, \tag{1}$$

which captures the non-stationarity of the environment. We focus on the class of nonstationary environments with total variation subjecting to a *variation budget* $V_T \geq 0$:

$$\mathcal{E}(V_T, T, K) := \left\{\mathcal{F}_T^{\mathcal{K}} \mid v\left(\mathcal{F}_T^{\mathcal{K}}\right) \leq V_T\right\}.$$

At each time slot $t \in \mathcal{T}$, a decision-making agent selects an arm $\varphi_t \in \mathcal{K}$ and receives an associated random reward $X_t^{\varphi_t}$. The objective is to maximize the expected value of the *cumulative reward* $S_T := \sum_{t=1}^T X_t^{\varphi_t}$. We assume that φ_t is selected based upon past observations $\{X_s^{\varphi_s}, \varphi_s\}_{s=1}^{t-1}$ following some policy ρ. Specifically, ρ determines the conditional distribution

$$\mathbb{P}^\rho\left(\varphi_t = k \mid \{X_s^{\varphi_s}, \varphi_s\}_{s=1}^{t-1}\right)$$

at each time $t \in \{1, \ldots, T-1\}$. If $\mathbb{P}^\rho(\cdot)$ takes binary values, ρ is called deterministic; otherwise, it is called stochastic.

Let the expected reward from the best arm at time t be $\mu_t^* = \max_{k\in\mathcal{K}} \mu_t^k$. Then, the objective of maximizing the expected cumulative reward is equivalent to minimizing the regret defined by

$$R_T^\rho := \sum_{t=1}^T \mu_t^* - \mathbb{E}^\rho[S_T] = \mathbb{E}^\rho\left[\sum_{t=1}^T \mu_t^* - \mu_t^{\varphi_t}\right],$$

where the expectation is with respect to different realizations of φ_t that depend on obtained rewards through policy ρ.

Note that the performance of a policy ρ differs with different $\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)$. For a fixed variation budget V_T and a policy ρ, the *worst-case regret* is the regret with respect to the worst possible choice of environment, i.e.,

$$R_{\text{worst}}^\rho(V_T, T, K) = \sup_{\mathcal{F}_T^{\mathcal{K}}\in\mathcal{E}(V_T,T,K)} R_T^\rho.$$

In this paper, we aim to design policies to minimize the worst-case regret. The optimal worst-case regret achieved by any policy is called the *minimax regret* and is defined by

$$\inf_\rho \sup_{\mathcal{F}_T^{\mathcal{K}}\in\mathcal{E}(V_T,T,K)} R_T^\rho.$$

We will study the nonstationary bandit problem under the following two setups concerning reward distributions. In the first setup, we follow [20], [32] to assume rewards to be sub-Gaussian with bounded expected value.

Assumption 1 (Sub-Gaussian reward): For any $k \in \mathcal{K}$ and any $t \in \mathcal{T}$, distribution $f_t^k(x)$ is $1/2$ sub-Gaussian, i.e.,

$$\forall\lambda \in \mathbb{R} : \mathbb{E}\left[\exp(\lambda(X_t^k - \mu_t^k))\right] \leq \exp\left(\frac{\lambda^2}{8}\right).$$

Moreover, for any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, $\mathbb{E}[X_t^k] \in [a, a+b]$, where $a \in \mathbb{R}$ and $b > 0$.

The second setup adheres to the heavy-tailed reward assumption in [40]. It is a relaxation of Assumption 1.[2]

Assumption 2 (Heavy-tailed reward): For any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, $\mathbb{E}[(X_t^k)^2] \leq 1$.

III. PRELIMINARIES

In this section, we review existing minimax regret lower bounds and minimax policies from the literature. These results apply to both sub-Gaussian and heavy-tailed rewards. The discussion is made first for the stationary environment ($V_T = 0$). Then, we show how the minimax regret lower bound for

[2] A zero-mean random variable X is sub-Gaussian iff there exists a constant $\theta > 0$ such that $\mathbb{E}[X^{2n}] \leq \theta^{2n}(2n)!/2^n n!$ for all $n \in \mathbb{N}$ [41, Th. 2.6]. Assumption 2 only require $\mathbb{E}[X^{2n}]$ is bounded when $n = 1$.

stationary stochastic bandit can be extended to establish the minimax regret lower bound for the nonstationary setup where $V_T > 0$. Furthermore, we review two UCB policies for the stationary stochastic MAB problem: UCB1 and MOSS. In the later sections, they are extended to design a variety of policies to match with the minimax regret lower bound for $V_T > 0$.

A. LOWER BOUND ON MINIMAX REGRET WHEN $V_T = 0$

In the setting of $V_T = 0$, for each arm $k \in \mathcal{K}$, μ_t^k is identical for all $t \in \mathcal{T}$. In stationary stochastic MAB problems, the rewards from each arm $k \in \mathcal{K}$ are independent and identically distributed, so they belong to the environment set $\mathcal{E}(0, T, K)$. According to [42], if $V_T = 0$, the minimax regret is no smaller than $1/20\sqrt{KT}$. This result is closely related to the standard logarithmic lower bound on the regret for stationary stochastic MAB problems as discussed below. Consider a scenario in which there is a unique best arm and all other arms have identical mean rewards such that the gap between optimal and suboptimal mean rewards is Δ. From [43], for such a stationary stochastic MAB problem

$$R_T^\rho \geq C_1 \frac{K}{\Delta} \ln\left(\frac{T\Delta^2}{K}\right) + C_2 \frac{K}{\Delta}, \tag{2}$$

for any policy ρ, where C_1 and C_2 are some positive constants. It needs to be noted that for $\Delta = \sqrt{K/T}$, the above lower bound becomes $C_2\sqrt{KT}$, which matches with the lower bound $1/20\sqrt{KT}$.

B. LOWER BOUND ON MINIMAX REGRET WHEN $V_T > 0$

In the setting of $V_T > 0$, we recall here the minimax regret lower bound for nonstationary stochastic MAB problems.

Lemma 1 (Minimax Lower Bound. $V_T > 0$ [31]): For the nonstationary MAB problem with K arms, time horizon T and variation budget $V_T \in [1/K, T/K]$,

$$\inf_{\rho} \sup_{\mathcal{F}_T^K \in \mathcal{E}(V_T, T, K)} R_T^\rho \geq C(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}},$$

for some constant $C > 0$.

In [31], the lower bound is derived assuming bounded rewards, so it holds under both Assumption 1 andAssumption 2. To understand this lower bound, consider the following nonstationary environment. We partition $\mathcal{T}$ into epochs of length $\tau = \left\lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}} \right\rceil$. In each epoch, the reward distribution sequences are stationary and all the arms have identical mean rewards except for the unique best arm. Let the gap in the mean be $\Delta = \sqrt{K/\tau}$. The index of the best arm switches at the end of each epoch following some unknown rule. So, the total variation is no greater than $\Delta T/\tau$, which satisfies the variation budget V_T. Besides, for any policy ρ, we know from (2) that worst-case regret in each epoch is no less than $C_2\sqrt{K\tau}$. Summing up the regret over all the epochs, minimax regret is lower bounded by $T/\tau \times C_2\sqrt{K\tau}$, which is consistent with Lemma 1.

C. UCB POLICIES IN STATIONARY ENVIRONMENTS

The family of UCB policies uses the principle of optimism in the face of uncertainty. In these policies, at each time slot, a UCB index which is a statistical index composed of both mean reward estimate and the associated uncertainty measure is computed at each arm, and the arm with the maximum UCB is picked. Within the family of UCB policies, two state-of-the-art algorithms for stationary stochastic MAB problems are UCB1 [15] and MOSS [39]. Let $n_k(t)$ be the number of times arm k is sampled until time $t-1$, and $\hat{\mu}_{k,n_k(t)}$ be the associated empirical mean. Then, UCB1 computes the UCB index for each arm k at time t as

$$g_{k,t}^{\text{UCB1}} = \hat{\mu}_{k,n_k(t)} + \sqrt{\frac{2\ln t}{n_k(t)}}.$$

It has been proved in [15] that, for the stationary stochastic MAB problem, UCB1 satisfies

$$R_T^{\text{UCB1}} \leq 8 \sum_{k:\Delta_k > 0} \frac{\ln T}{\Delta_k} + \left(1 + \frac{\pi^2}{3}\right) \sum_{k=1}^{K} \Delta_k,$$

where Δ_k is the difference in the mean rewards from arm k and the best arm. In [39], a simple variant of this result is given by selecting values for Δ_k to maximize the upper bound, resulting in

$$\sup_{\mathcal{F}_T^K \in \mathcal{E}(0,T,K)} R_T^{\text{UCB1}} \leq 10\sqrt{(K-1)T(\ln T)}.$$

Comparing this result with the lower bound on the minimax regret discussed in Section III-A, there exists an extra factor $\sqrt{\ln T}$. This issue has been resolved by the MOSS algorithm. With prior knowledge of horizon length T, the UCB index for MOSS is expressed as

$$g_{k,t}^{\text{MOSS}} = \hat{\mu}_{k,n_k(t)} + \sqrt{\frac{\max\left(\ln\left(\frac{T}{Kn_k(t)}\right), 0\right)}{n_k(t)}}. \tag{3}$$

We now recall the worst-case regret upper bound for MOSS.

Lemma 2 (Worst-case regret upper bound for MOSS [39]): For the stationary stochastic MAB problem ($V_T = 0$), the worst-case regret of the MOSS algorithm satisfies

$$\sup_{\mathcal{F}_T^K \in \mathcal{E}(0,T,K)} R_T^{\text{MOSS}} \leq 49\sqrt{KT}.$$

IV. UCB ALGORITHMS FOR SUB-GAUSSIAN NONSTATIONARY STOCHASTIC MAB PROBLEMS

In this section, we extend UCB1 and MOSS to design nonstationary UCB algorithms for scenarios with $V_T > 0$. Three different techniques are employed, namely periodic resetting, sliding observation window, and discount factor, to deal with the remembering-forgetting tradeoff. The proposed algorithms are analyzed to provide guarantees on the worst-case regret. We show their performances match closely with the lower bound in Lemma 1.

The following notations are used in later discussions. Let $|\cdot|$ represent the cardinality of a set when applied to a set,

Algorithm 1: *R-MOSS.*

Input : $V_T \in \mathbb{R}_{\geq 0}$ and $T \in \mathbb{N}$

Set : resetting period $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$

Output : sequence of arm selection

while $t \leq T$ **do**
  if $\mod(t,\tau) \leq K$ **then**
    Pick arm $\varphi_t = \mod(t,\tau)$; *% restart MOSS*
    Set $n_{\varphi_t}(t) = 1$ and $s_{\varphi_t}(t) = X_t^{\varphi_t}$;
  else
    Pick arm $\varphi_t \in \arg\max_{k\in\mathcal{K}} g_{k,t}^{\text{MOSS}}$ (defined in (3));
    Set $n_{\varphi_t}(t) = n_k(t) + 1$, $s_k(t) = s_k(t) + X_t^{\varphi_t}$

and denote the absolute value when applied to a real number. Let $\tau \in \{1,\dots,T\}$ be a design parameter and set $N = \lceil T/\tau \rceil$. We denote the indicator function as $\mathbf{1}\{\cdot\}$. Let $\{\mathcal{T}_1,\dots,\mathcal{T}_N\}$ be a partition $\mathcal{T}$, where each epoch $\mathcal{T}_i$ has length τ except possibly $\mathcal{T}_N$, i.e., for $i \in \{1,\dots,N\}$,

$$\mathcal{T}_i = \{1 + (i-1)\tau\ ,\dots,\ \min(i\tau, T)\}. \tag{4}$$

Let the maximum mean reward within $\mathcal{T}_i$ be achieved at time $\tau_i \in \mathcal{T}_i$ by arm κ_i, i.e., $\mu_{\tau_i}^{\kappa_i} = \max_{t\in\mathcal{T}_i} \mu_t^*$. We define the variation within $\mathcal{T}_i$ as

$$v_i := \sum_{t\in\mathcal{T}_i} \sup_{k\in\mathcal{K}} \left|\mu_{t+1}^k - \mu_t^k\right|, \tag{5}$$

where we trivially assign $\mu_{T+1}^k = \mu_T^k$ for all $k \in \mathcal{K}$.

A. RESETTING MOSS ALGORITHM

Periodic resetting is an effective technique to preserve the freshness and authenticity of the information history. It has been employed in [31] and [44] to design policies for nonstationary stochastic MAB problems. We extend this approach to MOSS and propose a nonstationary policy Resetting MOSS (R-MOSS). In R-MOSS, after every $\tau = \left\lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}} \right\rceil$ time slots, the sampling history is erased and MOSS is restarted. The pseudo-code is provided in Algorithm 1. In the following theorem, we show R-MOSS enjoys order-optimal worst-case regret matching with the lower bound in Lemma 1.

Theorem 1: For the nonstationary MAB problem with K arms, horizon T, and variation budget $V_T > 0$, if Assumption 1 is true, the worst case regret of R-MOSS satisfies,

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^{\text{R-MOSS}} = \mathcal{O}\left((KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

Sketch of the proof: Note that one run of MOSS takes place in each epoch. For epoch $\mathcal{T}_i$, define the set of *bad arms* for R-MOSS by

$$\mathcal{B}_i^{\text{R}} := \left\{k \in \mathcal{K} \mid \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^k \geq 2v_i\right\}. \tag{6}$$

Notice that for any $t_1, t_2 \in \mathcal{T}_i$,

$$\left|\mu_{t_1}^k - \mu_{t_2}^k\right| \leq v_i, \quad \forall k \in \mathcal{K}. \tag{7}$$

Therefore, for any $t \in \mathcal{T}_i$, we have

$$\mu_t^* - \mu_t^{\varphi_t} \leq \mu_{\tau_i}^{\kappa_i} - \mu_t^{\varphi_t} \leq \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} + v_i.$$

Then, the regret from $\mathcal{T}_i$ can be bounded as the following,

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in\mathcal{T}_i} \mu_t^* - \mu_t^{\varphi_t}\right] &\leq |\mathcal{T}_i|\, v_i + \mathbb{E}\left[\sum_{t\in\mathcal{T}_i} \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t}\right] \\ &\leq 3\,|\mathcal{T}_i|\, v_i + S_i, \end{aligned} \tag{8}$$

where $S_i = \mathbb{E}\left[\sum_{t\in\mathcal{T}_i} \sum_{k\in\mathcal{B}_i^{\text{R}}} \mathbf{1}\{\varphi_t = k\}(\mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} - 2v_i)\right]$.

Now, we have decoupled the problem, enabling us to generalize the analysis of MOSS in stationary environment [39] to bound S_i. We will only specify the generalization steps and skip the details for brevity.

First notice inequality (7) indicates that for any $k \in \mathcal{B}_i^{\text{R}}$ and any $t \in \mathcal{T}_i$,

$$\mu_t^{\kappa_i} \geq \mu_{\tau_i}^{\kappa_i} - v_i \text{ and } \mu_t^k \leq \mu_{\tau_i}^k + v_i.$$

So, at any $t \in \mathcal{T}_i$, $\hat{\mu}_{\kappa_i, n_{\kappa_i}(t)}$ concentrate around a value no smaller than $\mu_{\tau_i}^{\kappa_i} - v_i$, and $\hat{\mu}_{k,n_k(t)}$ concentrate around a value no greater than $\mu_{\tau_i}^k + v_i$ for any $k \in B_i^{\text{R}}$. Also $\mu_{\tau_i}^{\kappa_i} - v_i \geq \mu_{\tau_i}^k + v_i$ due to the definition in (6).

In the analysis of MOSS in stationary environment [39], the UCB of each suboptimal arm is compared with the best arm and each selection of suboptimal arm k contributes Δ_k to the regret. Here, we can apply a similar analysis by comparing the UCB of each arm $k \in B_i^{\text{R}}$ with κ_i and each selection of arm $k \in B_i^{\text{R}}$ contributes $(\mu_{\tau_i}^{\kappa_i} - v_i) - (\mu_{\tau_i}^k + v_i)$ in S_i. Accordingly, we borrow the upper bound in Lemma 2 to get $S_i \leq 49\sqrt{K|\mathcal{T}_i|}$.

Substituting the upper bound on S_i into (8) and summarizing over all the epochs, we conclude that

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T,T,K)} R_T^{\text{R-MOSS}} \leq 3\tau V_T + \sum_{i=1}^{N} 49\sqrt{K\tau},$$

which implies the theorem. ■

B. SLIDING WINDOW MOSS ALGORITHM

We have shown that periodic resetting coarsely adapts the stationary policy to a nonstationary setting. However, it is inefficient to entirely remove the sampling history at the restarting points and the regret accumulates quickly close to these points. To address the MAB problem with piece-wise stationary mean rewards, a sliding observation window is used to erase outdated information smoothly and more efficiently utilize the information history [20]. We show that a similar approach can also deal with the general nonstationary environment with a variation budget. In contrast to [20], we integrate the sliding window technique with MOSS instead of UCB1 and achieve the order-optimal worst-case regret.

Let the sliding observation window at time t be $\mathcal{W}_t := \{\min(1, t-\tau), \dots, t-1\}$, where $\tau = \left\lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}} \right\rceil$. Then,

Algorithm 2: SW-MOSS.

Input : $V_T \in \mathbb{R}_{>0}$, $T \in \mathbb{N}$ and $\eta > 1/2$

Set : sliding window size $\tau = \left\lceil K^{\frac{1}{3}} \left(T/V_T\right)^{\frac{2}{3}} \right\rceil$

Output : sequence of arm selection

Pick each arm once;

while $t \le T$ **do**

Set $\mathcal{W}_t = \{\min(1, t-\tau), \ldots, t-1\}$;

Compute statistics within sliding window $\mathcal{W}_t$:

$$n_k(t) = \sum_{s \in \mathcal{W}_t} \mathbf{1}\{\varphi_s = k\},$$

$$\hat{\mu}^k_{n_k(t)} = \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} X_s \mathbf{1}\{\varphi_s = k\};$$

Pick arm

$$\varphi_t = \arg\max_{k \in \mathcal{K}} \hat{\mu}^k_{n_k(t)} + \sqrt{\eta \frac{\max\left(\ln\left(\frac{\tau}{K n_k(t)}\right), 0\right)}{n_k(t)}};$$

the associated mean estimator is given by

$$\hat{\mu}^k_{n_k(t)} = \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} X_s \mathbf{1}\{\varphi_s = k\}, \ n_k(t) = \sum_{s \in \mathcal{W}_t} \mathbf{1}\{\varphi_s = k\}.$$

In SW-MOSS, the UCB for each arm $k \in \mathcal{K}$ is define as

$$g^k_t = \hat{\mu}^k_{n_k(t)} + c_{n_k(k)}, \ c_{n_k(t)} = \sqrt{\eta \frac{\max\left(\ln\left(\frac{\tau}{K n_k(t)}\right), 0\right)}{n_k(t)}},$$

where $\eta > 1/2$ is a tunable parameter. With these notations, SW-MOSS is defined in Algorithm 2.

At time t, for each arm $k \in \mathcal{K}$, we define

$$M^k_t := \frac{1}{n_k(t)} \sum_{s \in \mathcal{W}_t} \mu^k_s \mathbf{1}\{\varphi_s = k\}.$$

The following lemma presents concentration bounds for the sliding window empirical mean $\hat{\mu}^k_{n_k(t)}$, a crucial property employed in the regret analysis of SW-MOSS.

Lemma 3: For any arm $k \in \mathcal{K}$ and any time $t \in \mathcal{T}$, if $\eta > 1/2$, for any $x > 0$ and $l \ge 1$, the probability of either event $A = \{\hat{\mu}^k_{n_k(t)} + c_{n_k(t)} \le M^k_t - x, n_k(t) \ge l\}$ or event $B = \{\hat{\mu}^k_{n_k(t)} - c_{n_k(t)} \ge M^k_t + x, n_k(t) \ge l\}$ is no greater than

$$\frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau x^2} \exp\left(-x^2 \, l/\eta\right).$$

We defer the proof to Appendix A. Leveraging Lemma 3, we provide an upper bound on the worst-case regret for SW-MOSS in the following Theorem 2, showing SW-MOSS also enjoys order-optimal worst-case regret.

Theorem 2: For the nonstationary MAB problem with K arms, horizon T, and variation budget $V_T > 0$, if Assumption 1 is true, the worst-case regret of SW-MOSS satisfies

$$\sup_{\mathcal{F}^{\mathcal{K}}_T \in \mathcal{E}(V_T, T, K)} R^{\text{SW-MOSS}}_T = O\left((K V_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

Proof: The proof consists of the following five steps.

Step 1: Recall that v_i defined in (5) is the variation within $\mathcal{T}_i$ and $\mu^{\kappa_i}_{\tau_i} = \max_{t \in \mathcal{T}_i} \mu^*_t$. Here, we trivially assign $\mathcal{T}_0 = \emptyset$ and $v_0 = 0$. Then, for each $i \in \{1, \ldots, N\}$, let

$$\Delta^k_i := \mu^{\kappa_i}_{\tau_i} - \mu^k_{\tau_i} - 2v_{i-1} - 2v_i, \quad \forall k \in \mathcal{K}.$$

Define the set of bad arms for SW-MOSS in $\mathcal{T}_i$ as

$$\mathcal{B}^{\text{SW}}_i := \left\{k \in \mathcal{K} \mid \Delta^k_i \ge \epsilon\right\},$$

where we assign $\epsilon = 4\sqrt{e\eta K/\tau}$.

Step 2: We decouple the regret in this step. For any $t \in \mathcal{T}_i$, since $|\mu^k_t - \mu^k_{\tau_i}| \le v_i$ for any $k \in \mathcal{K}$, it satisfies that

$$\begin{aligned}\mu^*_t - \mu^{\varphi_t}_t &\le \mu^{\kappa_i}_{\tau_i} - \mu^{\varphi_t}_t \le \mu^{\kappa_i}_{\tau_i} - \mu^{\varphi_t}_{\tau_i} + v_i \\ &\le \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i\right\}(\Delta^{\varphi_t}_i - \epsilon) + 2v_{i-1} + 3v_i + \epsilon.\end{aligned}$$

Then we get the following inequalities,

$$\begin{aligned}&\sum_{t \in \mathcal{T}} \mu^*_t - \mu^{\varphi_t}_t \\ &\le \sum_{i=1}^{N} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i\right\}(\Delta^{\varphi_t}_i - \epsilon) + 2v_{i-1} + 3v_i + \epsilon \\ &\le 5\tau V_T + T\epsilon + \sum_{i=1}^{N} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i\right\}(\Delta^{\varphi_t}_i - \epsilon). \quad (9)\end{aligned}$$

To bound the regret $\mathbb{E}[\sum_{t \in \mathcal{T}} \mu^*_t - \mu^{\varphi_t}_t]$, we only need to provide an upper bound on $\mathbb{E}[(9)]$. To continue, we take a decomposition inspired by the analysis of MOSS in [39],

$$\begin{aligned}&\sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i\right\}(\Delta^{\varphi_t}_i - \epsilon) \\ &\le \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i, g^{\kappa_i}_t > M^{\kappa_i}_t - \frac{\Delta^{\varphi_t}_i}{4}\right\} \Delta^{\varphi_t}_i \quad (10) \\ &+ \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i, g^{\kappa_i}_t \le M^{\kappa_i}_t - \frac{\Delta^{\varphi_t}_i}{4}\right\}(\Delta^{\varphi_t}_i - \epsilon), \quad (11)\end{aligned}$$

where summands (10) describes the regret generated when arm κ_i is fairly estimated and summand (11) quantifies the regret incurred by underestimating arm κ_i.

Step 3: In this step, we bound $\mathbb{E}[(10)]$. Since $g^{\varphi_t}_t \ge g^{\kappa_i}_t$, we upper bound (10) as

$$\begin{aligned}(10) &\le \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t \in \mathcal{B}^{\text{SW}}_i, g^{\varphi_t}_t > M^{\kappa_i}_t - \frac{\Delta^{\varphi_t}_i}{4}\right\} \Delta^{\varphi_t}_i \\ &= \sum_{k \in \mathcal{B}^{\text{SW}}_i} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g^k_t > M^{\kappa_i}_t - \frac{\Delta^k_i}{4}\right\} \Delta^k_i. \quad (12)\end{aligned}$$

Notice that for any $t \in \mathcal{T}_{i-1} \cup \mathcal{T}_i$,

$$\left|\mu^k_t - \mu^k_{\tau_i}\right| \le v_{i-1} + v_i, \quad \forall k \in \mathcal{K}.$$

It indicates that an arm $k \in \mathcal{B}_i^{\text{SW}}$ is at least Δ_i^k worse in mean reward than arm κ_i at any time slot $t \in \mathcal{T}_{i-1} \cup \mathcal{T}_i$. Since $\mathcal{W}_t \subset \mathcal{T}_{i-1} \cup \mathcal{T}_i$, for any $t \in \mathcal{T}_i$,

$$M_t^{\kappa_i} - M_t^k \geq \Delta_i^k \geq \epsilon, \quad \forall k \in \mathcal{B}_i^{\text{SW}}.$$

It follows that

$$(12) \leq \sum_{k \in \mathcal{B}_i^{\text{SW}}} \sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^k + \frac{3\Delta_i^k}{4}\right\} \Delta_i^k. \quad (13)$$

The summation of indicator functions in (13) can be further bounded as below.

Let t_s^{ik} be the s-th time slot when arm k is selected within $\mathcal{T}_i$. Then, for any $k \in \mathcal{B}_i^{\text{SW}}$,

$$\begin{aligned}
&\sum_{t \in \mathcal{T}_i} \mathbf{1}\left\{\varphi_t = k, g_t^k > M_t^k + \frac{3\Delta_i^k}{4}\right\} \\
&= \sum_{s \geq 1} \mathbf{1}\left\{g_{t_s^{ik}}^k > M_{t_s^{ik}}^k + \frac{3\Delta_i^k}{4}\right\} \\
&\leq l_i^k + \sum_{s \geq l_i^k + 1} \mathbf{1}\left\{g_{t_s^{ik}}^k > M_{t_s^{ik}}^k + \frac{3\Delta_i^k}{4}\right\}, \quad (14)
\end{aligned}$$

where we set $l_i^k = \left\lceil \eta(\frac{4}{\Delta_i^k})^2 \ln(\frac{\tau}{\eta K}(\frac{\Delta_i^k}{4})^2) \right\rceil$. Since $\Delta_i^k \geq \epsilon$, for $k \in \mathcal{B}_i^{\text{SW}}$, we have

$$l_i^k \geq \left\lceil \eta \left(4/\Delta_i^k\right)^2 \ln\left(\frac{\tau}{\eta K}(\epsilon/4)^2\right) \right\rceil \geq \eta \left(4/\Delta_i^k\right)^2,$$

where the second inequality follows by substituting $\epsilon = 4\sqrt{e\eta K/\tau}$. Additionally, since $t_1^{ik}, \ldots, t_{s-1}^{ik} \in \mathcal{W}_{t_s^{ik}}$, we get $n_k(t_s^{ik}) \geq s - 1$. Furthermore, since c_m is monotonically decreasing with m,

$$c_{n_k(t_s^k)} \leq c_{l_i^k} \leq \sqrt{\frac{\eta}{l_i^k} \ln\left(\frac{\tau}{\eta K}\left(\frac{\Delta_i^k}{4}\right)^2\right)} \leq \frac{\Delta_i^k}{4},$$

for $s \geq l_i^k + 1$. Therefore,

$$(14) \leq l_i^k + \sum_{s \geq l_i^k + 1} \mathbf{1}\left\{g_{t_s^{ik}}^k - 2c_{n_k(t_s^{ik})} > M_{t_s^{ik}}^k + \frac{\Delta_i^k}{4}\right\}. \quad (15)$$

By applying Lemma 3, considering $n_k(t_s^{ik}) \geq s - 1$, the expected value of the second term in (15) satisfies

$$\begin{aligned}
&\sum_{s \geq l_i^k + 1} \mathbb{P}\left\{g_{t_s^{ik}}^k - 2c_{n_k(t_s^{ik})} > M_{t_s^{ik}}^k + \frac{\Delta_i^k}{4}\right\} \\
&\leq \sum_{s \geq l_i^k} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau} \left(\frac{4}{\Delta_i^k}\right)^2 \exp\left(-\frac{s}{\eta}\left(\frac{\Delta_i^k}{4}\right)^2\right) \\
&\leq \int_{l_i^k - 1}^{+\infty} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau} \left(\frac{4}{\Delta_i^k}\right)^2 \exp\left(-\frac{y}{\eta}\left(\frac{\Delta_i^k}{4}\right)^2\right) dy \\
&\leq \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \left(\frac{4}{\Delta_i^k}\right)^4. \quad (16)
\end{aligned}$$

Let $h(x) = 16\eta/x \ln(\tau x^2/16\eta K)$ which achieves maximum at $4e\sqrt{\eta K/\tau}$. Combining (16), (15), (14), (13), and (12), we obtain

$$\begin{aligned}
\mathbb{E}[(10)] &\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\left(\Delta_i^k\right)^3} + l_i^k \Delta_i^k \\
&\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\left(\Delta_i^k\right)^3} + h(\Delta_i^k) + \Delta_i^k \\
&\leq \sum_{k \in \mathcal{B}_i} \frac{(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{\eta K}{\tau} \frac{256}{\epsilon^3} + h\left(4e\sqrt{\eta K/\tau}\right) + b \\
&\leq \left(\frac{2.6\eta}{\ln(2\eta)} + 3\sqrt{\eta}\right)\sqrt{K\tau} + Kb.
\end{aligned}$$

Step 4: In this step, we establish a bound on $\mathbb{E}[(11)]$. When event $\{\varphi_t \in \mathcal{B}_i^{\text{SW}}, g_t^{\kappa_i} \leq M_t^{\kappa_i} - \Delta_i^{\varphi_t}/4\}$ happens, we know

$$\Delta_i^{\varphi_t} \leq 4M_t^{\kappa_i} - 4g_t^{\kappa_i} \text{ and } g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\epsilon}{4}.$$

Thus, we have

$$\begin{aligned}
&\mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\text{SW}}, g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\Delta_i^{\varphi_t}}{4}\right\}\left(\Delta_i^{\varphi_t} - \epsilon\right) \\
&\leq \mathbf{1}\left\{g_t^{\kappa_i} \leq M_t^{\kappa_i} - \frac{\epsilon}{4}\right\} \times \left(4M_t^{\kappa_i} - 4g_t^{\kappa_i} - \epsilon\right) := Y
\end{aligned}$$

Since Y is a nonnegative random variable, its expectation can be computed involving only its cumulative density function:

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^{+\infty} \mathbb{P}(Y > x)\, dx \\
&\leq \int_0^{+\infty} \mathbb{P}\left(4M_t^{\kappa_i} - 4g_t^{\kappa_i} - \epsilon \geq x\right) dx \\
&= \int_\epsilon^{+\infty} \mathbb{P}\left(4M_t^{\kappa_i} - 4g_t^{\kappa_i} > x\right) dx \\
&\leq \int_\epsilon^{+\infty} \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau x^2} dx = \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{K}{\tau\epsilon}.
\end{aligned}$$

Hence, $\mathbb{E}[(11)] \leq 16(2\eta)^{\frac{3}{2}} K|\mathcal{T}_i|/(\ln(2\eta)\tau\epsilon)$.

Step 5: For any nonstationary environment subject variation budget V_T, $R_T^{\text{SW-MOSS}} \leq \mathbb{E}[(9)]$. Furthermore, with bounds on $\mathbb{E}[(10)]$ and $\mathbb{E}[(11)]$ from previous steps,

$$\mathbb{E}[(9)] \leq 5\tau V_T + T\epsilon + N\left(\frac{2.6\eta}{\ln(2\eta)} + 3\sqrt{\eta}\right)\sqrt{K\tau}$$

Algorithm 3: *D-UCB.*

Input : $V_T \in \mathbb{R}_{>0}$, $T \in \mathbb{N}$ and $\xi > \frac{1}{2}$

Set : $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$

Output : sequence of arm selection

for $t \in \{1, \dots, K\}$ **do**

  Pick arm $\varphi_t = t$ and set $n^t \leftarrow \gamma^{K-t}$ and $\hat{\mu}^t \leftarrow X_t^t$;

while $t \leq T$ **do**

  Pick arm $\varphi_t = \arg\max_{k \in \mathcal{K}} \hat{\mu}^k + 2\sqrt{\frac{\xi \ln(\tau)}{n^k}}$;

  For each arm $k \in \mathcal{K}$, set $n^k \leftarrow \gamma n^k$;

  Set $n^{\varphi_t} \leftarrow n^{\varphi_t} + 1$ & $\hat{\mu}^{\varphi_t} \leftarrow \hat{\mu}^{\varphi_t} + \frac{1}{n^{\varphi_t}}(X_t^{\varphi_t} - \bar{X}^{\varphi_t})$;

$$+ NKb + \frac{16(2\eta)^{\frac{3}{2}}}{\ln(2\eta)} \frac{KT}{\tau\epsilon} \leq C(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}$$

for some constant C, which concludes the proof. ■

One limitation of the sliding window method is the necessity to store the entire sampling history within the observation window. Since window size is selected to be $\tau = \lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}} \rceil$, large memory is needed for large horizon length T. The next policy resolves this problem.

C. DISCOUNTED UCB ALGORITHM

The discount factor method normally requires less memory, and it is widely used in estimators to forget old information and pay more attention to recent information. In [20], such an estimation is used together with UCB1 to solve the piecewise stationary MAB problem, and the policy designed is called Discounted UCB (D-UCB). Here, we tune D-UCB to work in the nonstationary environment with variation budget V_T.[3] Specifically, the mean estimator used is the discounted empirical average given by

$$\hat{\mu}_{\gamma,t}^k = \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} X_s,$$

$$n_{\gamma,t}^k = \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\},$$

where $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$ is the discount factor. Besides, the UCB is designed as $g_t^k = \hat{\mu}_t^k + 2c_t^k$, where $c_{\gamma,t}^k = \sqrt{\xi \ln(\tau)/n_{\gamma,t}^k}$ for some constant $\xi > 1/2$. The pseudo-code for D-UCB is reproduced in Algorithm 3. It can be noticed that the memory size is only related to the number of arms, so D-UCB requires a smaller memory.

To analyze D-UCB, we recall an exisiting concentration inequality for the discounted empirical average. Let

$$M_{\gamma,t}^k := \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \mu_s^k.$$

The following fact is a corollary of [20, Th. 18].

Fact 1 (A Hoeffding-type inequality[4]): For any $t \in \mathcal{T}$ and for any $k \in \mathcal{K}$, the probability of event $A = \{\hat{\mu}_{\gamma,t}^k - M_{\gamma,t}^k \geq \delta/\sqrt{n_{\gamma,t}^k}\}$ is no greater than

$$\lceil \log_{1+\lambda}(\tau) \rceil \exp\left(-2\delta^2\left(1 - \lambda^2/16\right)\right) \tag{17}$$

for any $\delta > 0$ and $\lambda > 0$. The probability of event $B = \{\hat{\mu}_{\gamma,t}^k - M_{\gamma,t}^k \leq -\delta/\sqrt{n_{\gamma,t}^k}\}$ is also upper bounded by (17).

In the following, we provide an upper bound on the worst-case regret for D-UCB. In comparison with the regret lower bound in Lemma 1, there exists an extra factor $\ln(T)$.

Theorem 3: For the nonstationary MAB problem with K arms, horizon T, and variation budget $V_T > 0$, if Assumption 1 is true, then by setting $\gamma = 1 - K^{-\frac{1}{3}}(T/V_T)^{-\frac{2}{3}}$, the worst-case regret of D-UCB satisfies

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)} R_T^{\text{D-UCB}} = \mathcal{O}\left(\ln(T)(KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

Proof: We establish the theorem in four steps.

Step 1: Let $\tau = \lceil K^{\frac{1}{3}}(T/V_T)^{\frac{2}{3}} \rceil$ and recall $\mathcal{T}_i$ defined in (4). Also recall that the maximum mean reward within $\mathcal{T}_i$ is achieved at time τ_i by arm κ_i. Let $\tau' = \log_\gamma((1-\gamma)\xi \ln(\tau)/b^2)$ and take $t - \tau'$ as a dividing point, then for any $t \in \mathcal{T}_i$,

$$\begin{aligned}\left|\mu_{\tau_i}^k - M_{\gamma,t}^k\right| &\leq \frac{1}{n_{\gamma,t}^k} \sum_{s=1}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right| \\ &\leq \frac{1}{n_{\gamma,t}^k} \sum_{s \leq t-\tau'} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right| \quad (18) \\ &\quad + \frac{1}{n_{\gamma,t}^k} \sum_{s \geq t-\tau'}^{t-1} \gamma^{t-s} \mathbf{1}\{\varphi_s = k\} \left|\mu_{\tau_i}^k - \mu_s^k\right|. \quad (19)\end{aligned}$$

Since $\mu_t^k \in [a, a+b]$ for all $t \in \mathcal{T}$, we have (18) $\leq b$. Also,

$$(18) \leq \frac{1}{n_{\gamma,t}^k} \sum_{s \leq t-\tau'} b\gamma^{t-s} \leq \frac{b\gamma^{\tau'}}{(1-\gamma)n_{\gamma,t}^k} = \frac{\xi \ln(\tau)}{b n_{\gamma,t}^k}.$$

Accordingly, we get

$$(18) \leq \min\left(b, \frac{\xi \ln(\tau)}{b n_{\gamma,t}^k}\right) \leq \sqrt{\frac{\xi \ln(\tau)}{n_{\gamma,t}^k}} = c_{\gamma,t}^k.$$

[3] D-UCB has been extended to D-LinUCB [45] to deal with nonstationary linear bandit problem. However, the notion of variation budget used in [45] is slightly different from this work.

[4] The proof is the same as [20, Th. 18]. The only change required is the MGF for bounded variables is replaced by that of sub-Gaussian variables.

Furthermore, for any $t \in \mathcal{T}_i$,

$$(19) \leq \max_{s\in[t-\tau',t-1]} \left|\mu_{\tau_i}^k - \mu_s^k\right| \leq \sum_{j=i-n'}^{i} v_j,$$

where $n' = \lceil \tau'/\tau \rceil$ and v_j is the variation within $\mathcal{T}_j$. So we conclude that for any $t \in \mathcal{T}_i$,

$$\left|\mu_{\tau_i}^k - M_{\gamma,t}^k\right| \leq c_{\gamma,t}^k + \sum_{j=i-n'}^{i} v_j, \quad \forall k \in \mathcal{K}. \tag{20}$$

This intermediate result will be used in Step 3 of the proof.

Step 2: Within partition $\mathcal{T}_i$, let

$$\hat{\Delta}_i^k = \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^k - 2\sum_{j=i-n'}^{i} v_j,$$

and define a subset of bad arms as $\mathcal{B}_i^{\mathrm{D}} = \{k \in \mathcal{K} \mid \hat{\Delta}_i^k \geq \epsilon'\}$, where we set $\epsilon' = 4\sqrt{\xi\gamma^{1-\tau}K\ln(\tau)/\tau}$. Since $|\mu_t^k - \mu_{\tau_i}^k| \leq v_i$ for any $t \in \mathcal{T}_i$ and for any $k \in \mathcal{K}$

$$\begin{aligned}
\sum_{t\in\mathcal{T}} \mu_t^* - \mu_t^{\varphi_t} &\leq \sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i} \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^{\varphi_t} + v_i \\
&\leq \tau V_T + \sum_{i=1}^{N}\sum_{t\in\mathcal{T}_i}\left[\mathbf{1}\left\{\varphi_t \in \mathcal{B}_i^{\mathrm{D}}\right\}\hat{\Delta}_i^{\varphi_t} + 2\sum_{j=i-n'}^{i} v_j + \epsilon'\right] \\
&\leq (2n'+3)\tau V_T + N\epsilon'\tau + \sum_{i=1}^{N}\sum_{k\in\mathcal{B}_i^{\mathrm{D}}}\hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i}\mathbf{1}\{\varphi_t = k\}.
\end{aligned} \tag{21}$$

We will upper bound regret $\mathbb{E}[\sum_{t\in\mathcal{T}} \mu_t^* - \mu_t^{\varphi_t}]$ by upper bounding $\mathbb{E}[(21)]$ in the following steps.

Step 3: In this step, we follow from (21) to provide a upper bound on $\mathbb{E}[\hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i}\mathbf{1}\{\varphi_t = k\}]$ for an arm $k \in \mathcal{B}_i^{\mathrm{D}}$. Let $t_i^k(l)$ be the l-th time slot arm k is selected within $\mathcal{T}_i$. From arm selection policy, we get $g_t^{\varphi_t} \geq g_t^{\kappa_i}$, which result in

$$\sum_{t\in\mathcal{T}_i}\mathbf{1}\{\varphi_t = k\} \leq l_i^k + \sum_{t\in\mathcal{T}_i}\mathbf{1}\left\{g_t^k \geq g_t^{\kappa_i}, t > t_i^k(l_i^k)\right\}, \tag{22}$$

where we pick $l_i^k = \lceil 16\xi\gamma^{1-\tau}\ln(\tau)/(\hat{\Delta}_i^k)^2\rceil$. Note that $g_t^k \geq g_t^{\kappa_i}$ is true means at least one of the following inequalities holds,

$$\hat{\mu}_{\gamma,t}^k \geq M_{\gamma,t}^k + c_{\gamma,t}^k, \tag{23}$$

$$\hat{\mu}_{\gamma,t}^{\kappa_i} \leq M_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^{\kappa_i}, \tag{24}$$

$$M_{\gamma,t}^{\kappa_i} + c_{\gamma,t}^{\kappa_i} < M_{\gamma,t}^k + 3c_{\gamma,t}^k. \tag{25}$$

For any $t \in \mathcal{T}_i$, since every sample before t within $\mathcal{T}_i$ has a weight greater than $\gamma^{\tau-1}$, if $t > t_i^k(l_i^k)$,

$$c_{\gamma,t}^k = \sqrt{\frac{\xi\ln(\tau)}{n_{\gamma,t}^k}} \leq \sqrt{\frac{\xi\ln(\tau)}{\gamma^{\tau-1}l_i^k}} \leq \frac{\hat{\Delta}_i^k}{4}.$$

Combining it with (20) yields

$$\begin{aligned}
M_{\gamma,t}^{\kappa_i} - M_{\gamma,t}^k &\geq \mu_{\tau_i}^{\kappa_i} - \mu_{\tau_i}^k - c_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^k - 2\sum_{j=i-n'}^{i} v_j \\
&\geq \hat{\Delta}_i^k - c_{\gamma,t}^{\kappa_i} - c_{\gamma,t}^k \geq 3c_{\gamma,t}^k - c_{\gamma,t}^{\kappa_i},
\end{aligned}$$

which indicates (25) is false. As $\xi > 1/2$, we select $\lambda = 4\sqrt{1-1/(2\xi)}$ and apply Fact 1 to get

$$\mathbb{P}((23) \text{ is true}) \leq \left\lceil \log_{1+\lambda}(\tau)\right\rceil \tau^{-2\xi(1-\frac{\lambda^2}{16})} \leq \frac{\left\lceil \log_{1+\lambda}(\tau)\right\rceil}{\tau}.$$

The probability of (24) to be true shares the same bound as above. Then, it follows from (22) that

$$\begin{aligned}
&\mathbb{E}\left[\hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i}\mathbf{1}\{\varphi_t = k\}\right] \\
&\leq \hat{\Delta}_i^k l_i^k + \hat{\Delta}_i^k \sum_{t\in\mathcal{T}_i}\mathbb{P}\left((23) \text{ or } (24) \text{ is true}\right) \\
&\leq \frac{16\xi\gamma^{1-\tau}\ln(\tau)}{\hat{\Delta}_i^k} + \hat{\Delta}_i^k + 2\hat{\Delta}_i^k\left\lceil \log_{1+\lambda}(\tau)\right\rceil \\
&\leq \frac{16\xi\gamma^{1-\tau}\ln(\tau)}{\epsilon'} + b + 2b\left\lceil \log_{1+\lambda}(\tau)\right\rceil,
\end{aligned} \tag{26}$$

where we use $\epsilon' \leq \hat{\Delta}_i^k \leq b$ in the last step.

Step 4: For any nonstationary environment subject variation budget V_T, $R_T^{\text{D-UCB}} \leq \mathbb{E}[(21)]$. With (21) and (26), plugging in the value of ϵ', a straightforward calculation leads to

$$\begin{aligned}
\mathbb{E}[(21)] &\leq (2n'+3)\tau V_T + 8N\sqrt{\xi\gamma^{1-\tau}K\tau\ln(\tau)} \\
&\quad + 2Nb + 2Nb\log_{1+\lambda}(\tau),
\end{aligned}$$

where the dominating term is $(2n'+3)\tau V_T$. Considering

$$\tau' = \frac{\ln\left((1-\gamma)\xi\ln(\tau)/b^2\right)}{\ln\gamma} \leq \frac{-\ln\left((1-\gamma)\xi\ln(\tau)/b^2\right)}{1-\gamma},$$

there exists some constant C' such that $n' \leq C'\ln(T)$. Accordingly, we get $R_T^{\text{D-UCB}} = O(\ln(T)(KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$. ■

Remark 1: Although D-UCB requires less memory, it is suboptimal with an extra factor $\ln(T)$ in the worst-case regret upper bound as shown in Theorem 3. This is because the discount factor method does not entirely cut off the outdated sampling history like periodic resetting or sliding window techniques, resulting in a bias term (18) that needs to be addressed in the UCB index design.

V. UCB POLICIES FOR HEAVY-TAILED NONSTATIONARY STOCHASTIC MAB PROBLEMS

In this section, we relax the sub-Gaussain assumption to study the nonstationary stochastic MAB problem with heavy-tailed rewards defined in Assumption 2. We first recall a minimax policy for the stationary heavy-tailed MAB problem called Robust MOSS [46]. We then extend it to the nonstationary setting and design two robust UCB algorithms. The worst-case regret upper bounds are also presented.

A. BACKGROUND ON ROBUST MOSS FOR THE STATIONARY HEAVY-TAILED MAB PROBLEM

In [46], Robust MOSS is designed to address stationary heavy-tailed MAB problems, in which for a reward X from an arbitrary arm, $\mathbb{E}[|X|^{1+\epsilon}]$ is bounded for some $\epsilon \in (0, 1]$. Such stochastic observations with finite moments are common in many applications including social networks [37] and financial markets [38]. For simplicity, as stated in Assumption 2, we restrict our discussion to $\epsilon = 1$.

Instead of the empirical mean, Robust MOSS uses the saturated empirical mean, which truncates outliers in reward values, ensuring robustness in a heavy-tailed environment. Let $n_k(t)$ be the number of times that the arm k has been selected until time $t-1$. Pick $a > 1$ and let $h(m) = a^{\lfloor \log_a(m) \rfloor + 1}$. Let the saturation limit at time t be defined by $B_{n_k(t)} := \sqrt{h(n_k(t))/\ln_+(\frac{T}{Kh(n_k(t))})}$, where $\ln_+(x) := \max(\ln x, 1)$. Then, the saturated empirical mean estimator is defined by

$$\bar{\mu}_{n_k(t)} := \frac{1}{n_k(t)} \sum_{s=1}^{t-1} \mathbf{1}\{\varphi_s = k\} \operatorname{sat}(X_s, B_{n_k(t)}), \tag{27}$$

where $\operatorname{sat}(X_s, B_m) := \operatorname{sign}(X_s) \min\{|X_s|, B_m\}$. The Robust MOSS algorithm initializes by selecting each arm once and subsequently, at each time t, selects the arm that maximizes the following upper confidence bound

$$g^k_{n_k(t)} = \bar{\mu}^k_{n_k(t)} + (1+\zeta) c_{n_k(t)},$$

where $c_{n_k(t)} = \sqrt{\ln_+(\frac{T}{Kn_k(t)})/n_k(t)}$, ζ is an positive constant such that $\psi(2\zeta/a) \geq 2a/\zeta$ and $\psi(x) = (1+1/x)\ln(1+x) - 1$. Note that for $x \in (0, \infty)$, function $\psi(x)$ is monotonically increasing in x.

B. RESETTING ROBUST MOSS ALGORITHM

Similarly to R-MOSS, Resetting Robust MOSS (R-RMOSS) restarts Robust MOSS after every τ time slots. For a stationary heavy-tailed MAB problem, the worst-case regret of Robust MOSS belongs to $\mathcal{O}(\sqrt{KT})$ [46]. This result, coupled with an analysis similar to the one for R-MOSS in Theorem 1, leads to the following Theorem 4. Putting it together with Lemma 1, we show the worst-case regret for R-RMOSS in the heavy-tailed nonstationary stochastic MAB problems is order-optimal. For brevity, we skip the proof.

Theorem 4: For the nonstationary MAB problem with K arms, horizon T, and variation budget $V_T > 0$, ifAssumption 2 is true, the worst-case regret of R-RMOSS satisfies

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)} R_T^{\text{R-RMOSS}} = \mathcal{O}\left((KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

C. SLIDING WINDOW ROBUST MOSS ALGORITHM

In Sliding Window Robust MOSS (SW-RMOSS), $n_k(t)$ and $\bar{\mu}_{n_k(t)}$ are computed from the sampling history within $\mathcal{W}_t$, and $c_{n_k(t)} = \sqrt{\ln_+(\frac{\tau}{Kn_k(t)})/n_k(t)}$. To analyze SW-RMOSS, we need a similar property as Lemma 3 to bound the probability of an arm being under or over-estimated, and it is presented in the following Lemma 4.

Lemma 4: For any arm $k \in \{1, \ldots, K\}$ and any $t \in \{K+1, \ldots, T\}$, if $\psi(2\zeta/a) \geq 2a/\zeta$, the probability of either event $A = \{g^k_t \leq M^k_t - x, n_k(t) \geq l\}$ or event $B = \{g^k_t - 2c_{n_k(t)} \geq M^k_t + x, n_k(t) \geq l\}$, for any $x > 0$ and any $l \geq 1$, is no greater than

$$\frac{2a}{\beta^2 \ln(a)} \frac{K}{\tau x^2} (\beta x \sqrt{h(l)/a} + 1) \exp\left(-\beta x \sqrt{h(l)/a}\right),$$

where $\beta = \psi(2\zeta/a)/(2a)$.

The proof is deferred to the Appendix B. In the following Theorem 5, we utilize Lemma 4 to show SW-MOSS also enjoys order-optimal worst-case regret in the heavy-tailed nonstationary reward setup. Since the analysis is similar to Theorem 2, we provide a proof sketch.

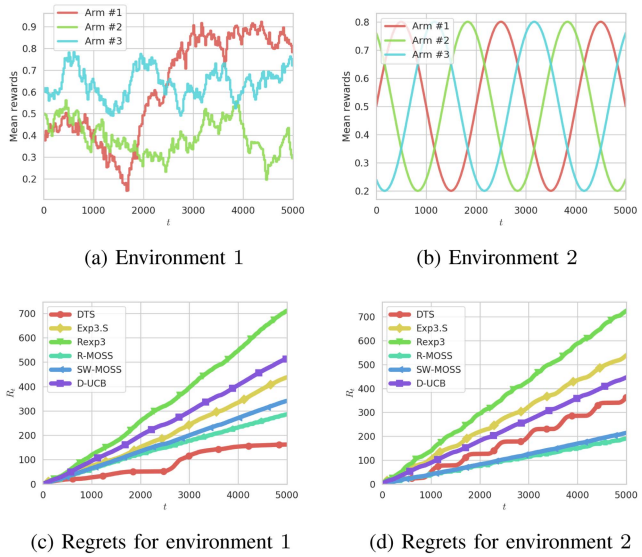
Theorem 5: For the nonstationary MAB problem with K arms, horizon T, and variation budget $V_T > 0$, if Assumption 2 is true, the worst-case regret of SW-RMOSS satisfies

$$\sup_{\mathcal{F}_T^{\mathcal{K}} \in \mathcal{E}(V_T, T, K)} R_T^{\text{SW-RMOSS}} = \mathcal{O}\left((KV_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right).$$

Sketch of the proof: The procedure is similar to the proof of Theorem 2. The key difference is due to the nuance between the concentration properties on the mean estimators. Neglecting the leading constants, the probability upper bound in Lemma 3 has a factor $\exp(-x^2\, l/\eta)$ comparing with $(\beta x\sqrt{h(l)/a} + 1)\exp(-\beta x\sqrt{h(l)/a})$ in Lemma 4. Since both factors are no greater than 1, by simply replacing η with $(1+\zeta)^2$ and taking similar calculations in every step except inequality (16), comparable bounds that only differ in leading constants can be obtained. Applying Lemma 4, we revise the computation of (16) as the following,

$$\begin{aligned}
&\sum_{s \geq l_i^k + 1} \mathbb{P}\left(g^k_{t_s} - 2c_{n_k(t_s)} > M^k_{t_s} + \frac{\Delta^k_i}{4}\right) \\
&\leq \sum_{s \geq l_i^k} C' \left(\frac{\beta \Delta^k_i}{4}\sqrt{\frac{h(l)}{a}} + 1\right) \exp\left(-\frac{\beta \Delta^k_i}{4}\sqrt{\frac{h(l)}{a}}\right) \\
&\leq \int_{l_i^k - 1}^{+\infty} C' \left(\frac{\beta \Delta^k_i}{4}\sqrt{\frac{y}{a}} + 1\right) \exp\left(-\frac{\beta \Delta^k_i}{4}\sqrt{\frac{y}{a}}\right) dy \\
&\leq \frac{6a}{\beta^2} \frac{2a}{\beta^2 \ln(a)} \frac{K}{\tau} \left(\frac{4}{\Delta^k_i}\right)^4.
\end{aligned} \tag{28}$$

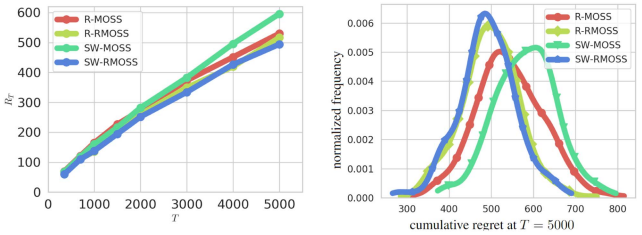
where $C' = 2aK(4/\Delta^k_i)^2/(\beta^2 \ln(a)\tau)$.The second inequality is due to the fact that $(x+1)\exp(-x)$ is monotonically decreasing in x for $x \in [0, \infty)$ and $h(l) > l$. In the last inequality, we change the lower limits of the integration from $l^k_i - 1$ to 0 since $l^k_i \geq 1$ and plug in the value of C'. Compared



(a) Environment 1 (b) Environment 2

(c) Regrets for environment 1 (d) Regrets for environment 2

FIGURE 1. Comparison of different policies.



(a) Mean regret for different T (b) Cumulative regret Histogram

FIGURE 2. Performances with heavy-tailed rewards.

with (16), this upper bound only varies in constant multiplier. So the worst-case regret for SW-RMOSS is $\mathcal{O}((KV_T)^{\frac{1}{3}}T^{\frac{2}{3}})$.■

Remark 2: The benefit of the discount factor method is that it is memory-friendly. This advantage is lost if a truncated empirical mean is used. As $n^k_{\gamma,t}$ could both increase and decrease with time, the truncated point could both grow and decline, so all sampling history needs to be recorded. It remains an open problem how to effectively use the discount factor in a nonstationary heavy-tailed MAB problem.

VI. NUMERICAL EXPERIMENTS

We complement the theoretical results in the previous section with two Monte Carlo experiments simulated in SMPyBandits [47], which is a bandit simulation framework. For the light-tailed setting, we compare R-MOSS, SW-MOSS, and D-UCB in this paper with other state-of-art policies. For the heavy-tailed setting, we test the robustness of R-RMOSS and SW-RMOSS against heavy-tailed nonstationary rewards. Each result in this section is derived by running designated policies 500 times. Parameter selections for compared policies are strictly coherent with the referred literature.

A. BERNOULLI NONSTATIONARY EXPERIMENT

To evaluate the performance of different policies, we consider two nonstationary environments as shown in Fig. 1(a) and (b), which both have 3 arms with nonstationary Bernoulli reward. The success probability sequence at each arm is a Brownian motion in Environment 1 and a sinusoidal function of time t in Environment 2. The variation budgets V_T are 8.09 and 3, respectively.

The growths of regret in Fig. 1(c) and (d) show that UCB-based policies (R-MOSS, SW-MOSS, and D-UCB) maintain their superior performance against adversarial bandit-based policies (Rexp3 [31] and Exp3.S [32]) for stochastic bandits even in nonstationary settings, especially for R-MOSS and SW-MOSS. Besides, DTS [33] outperforms other policies when the best arm does not switch. In contrast, each switch of the best arm seems to incur a larger regret accumulation for DTS, which results in a larger regret compared with SW-MOSS and R-MOSS.

B. HEAVY-TAILED NONSTATIONARY EXPERIMENT

Again we consider the 3-armed bandit problem with sinusoidal mean rewards. In particular, for each arm $k \in \{1, 2, 3\}$, $\mu^k_t = 0.3\sin(5\pi t/T + 2k\pi/3) \forall t \in \{1, \ldots, T\}$. Thus, the variation budget is 3 for any horizon length T. Besides, the mean reward is contaminated by additive sampling noise ν, where $|\nu|$ is a generalized Pareto random variable and the sign of ν has an equal probability to be "+" and "−". So the probability distribution for X^k_t is

$$f^k_t(x) = \frac{1}{2\sigma}\left(1 + \frac{\xi\left|x - \mu^k_t\right|}{\sigma}\right)^{-\frac{1}{\xi}-1} \quad \text{for } x \in (-\infty, +\infty).$$

We select $\xi = 0.4$ and $\sigma = 0.23$ such that Assumption 2 is satisfied. We select $a = 1.1$ and $\zeta = 2.2$ for both R-RMOSS and SW-RMOSS such that condition $\psi(2\zeta/a) \geq 2a/\zeta$ in Theorems 4, and 5 is met.

Fig. 2(a) shows that RMOSS-based policies achieve sublinear mean regret with respect to different horizon T, and they slightly outperform MOSS-based policies in heavy-tailed settings. By comparing the histogram of cumulative regret $\sum_{t\in\mathcal{T}} \mu^*_t - \mu^{\varphi_t}_t$ for different policies in Fig. 2(b), both R-RMOSS and SW-RMOSS have better consistency and less probability of a particular realization of the regret deviating significantly from the mean value.

VII. CONCLUSION

We studied the general nonstationary stochastic MAB problem with a variation budget and proposed three UCB-based policies. When the reward distributions are sub-Gaussian, our analysis showed that the proposed R-MOSS and SW-MOSS achieved the worst-case regret within a constant factor of the minimax regret lower bound. Additionally, D-UCB after tuning could achieve near-optimal worst-case regret with an additional $\ln(T)$ factor. Furthermore, we relaxed the sub-Gaussian assumption to study the heavy-tailed nonstationary MAB problem. We showed that the order-optimal worst-case

regret can be maintained by extending R-MOSS and SW-MOSS to their robust versions.

There are several possible avenues for future research. In this paper, we relied on passive methods to balance the remembering-versus-forgetting tradeoff. The general idea is to keep taking in new information and removing outdated information. Parameter-free active approaches that adaptively detect and react to environmental changes are promising alternatives and may result in better experimental performance. Also, extensions from the single decision-maker to distributed multiple decision-makers are of interest.

APPENDIX

The proofs in this section are developed upon two important concentration inequalities that have been introduced and utilized in [46], [48], [49], [50]. To make the paper self-contained, both are presented here with detailed proofs. We first introduce some measure-theoretic probability concepts that are necessary for rigorous analysis. Let $\{X_i\}_{i=1}^n$ be a sequence of random variables adapted to the filtration $\mathbb{F} = \{\mathcal{F}_i\}_{i=1}^n$, i.e., $X_1, \ldots, X_i$ is $\mathcal{F}_i$-measurable. The sequence $\{X_i\}_{i=1}^n$ is an $\mathbb{F}$-adapted submartingale if $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \geq X_{i-1}$ for all i. The subsequent result, stemming from Doob's optional stopping theorem [49], will serve as a key result to derive the concentration inequalities.

Lemma 5 (Maximal inequality [49, Th. 3.10]): Let $\{M_i\}_{i=1}^n$ be a submartingale with $M_i \geq 0$ almost surely for all $i \in \{1, \ldots, n\}$. Then for any $\delta > 0$,

$$\mathbb{P}\left(\max_{i\in\{1,\ldots,n\}} M_i \geq \delta\right) \leq \frac{\mathbb{E}[M_n]}{\delta}.$$

Let $d_i = X_i - \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$ and let $M_m = \exp(\lambda \sum_{i=1}^m d_i)$. Then by Jensen's inequality and the convexity of $\exp(\lambda x)$,

$$\begin{aligned}\mathbb{E}[M_m \mid \mathcal{F}_{m-1}] &= M_{m-1}\mathbb{E}[\exp(\lambda d_m) \mid \mathcal{F}_{m-1}] \\ &\geq M_{m-1}\exp(\lambda \mathbb{E}[d_m] \mid \mathcal{F}_{m-1}) = M_{m-1},\end{aligned}$$

which means $\{M_i\}_{i=1}^n$ is an $\mathbb{F}$-adapted submartingale. Besides, with $M_i \geq 0$, we can apply Lemma 5 to get the following result, which is the Azuma-Hoeffding inequality [51] generalized to sub-Gaussian random variables.

Lemma 6: Let $\{X_i\}_{i=1}^n$ be a sequence of random variables adapted to the filtration $\mathbb{F} = \{\mathcal{F}_i\}_{i=1}^n$. Define $d_i := X_i - \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$. If $\mathbb{P}(X_i \mid \mathcal{F}_{i-1})$ is σ sub-Gaussian for all i, then for any $\delta > 0$,

$$\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : \sum_{i=1}^m d_i \geq \delta\right) \leq \exp\left(-\frac{\delta^2}{2n\sigma^2}\right)$$

$$\text{and } \mathbb{P}\left(\exists m \in \{1,\ldots,n\} : \sum_{i=1}^m d_i \leq -\delta\right) \leq \exp\left(-\frac{\delta^2}{2n\sigma^2}\right).$$

Proof: Since $\mathbb{P}(X_i \mid \mathcal{F}_{i-1})$ is σ sub-Gaussian for all i,

$$\begin{aligned}\mathbb{E}[M_m] &= \mathbb{E}[M_{m-1}]\mathbb{E}[\exp(\lambda d_m) \mid \mathcal{F}_{m-1}] \\ &= \mathbb{E}[M_{m-1}]\exp\left(\frac{1}{2}\lambda^2\sigma^2\right).\end{aligned}$$

Applying the above equality iteratively from n to 1,

$$\mathbb{E}[M_n] = \exp\left(\frac{n}{2}\lambda^2\sigma^2\right).$$

Then, using Lemma 5, we have that for any $\lambda > 0$,

$$\begin{aligned}&\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : \sum_{i=1}^m d_i \geq \delta\right) \\ &= \mathbb{P}\left(\max_{m\in\{1,\ldots,n\}} M_m \geq \exp(\lambda\delta)\right) \\ &\leq \frac{\mathbb{E}[M_n]}{\exp(\lambda\delta)} = \exp\left(\frac{n}{2}\lambda^2\sigma^2 - \lambda\delta\right).\end{aligned}$$

Taking $\lambda = \delta/(n\sigma^2)$ to minimize the last term, the probability is no greater than $\exp(\delta^2/(2n\sigma^2))$. The lower tail probability bound regarding $\sum_{i=1}^m d_i \leq -\delta$ can be proved similarly by reversing the sign of d_i. ■

The following is the martingale version of Bennett's inequality [52], whose proof procedure is similar to Lemma 6.

Lemma 7: Let $\{X_i\}_{i=1}^n$ be a sequence of bounded random variables in $[-B, B]$ for some $B \geq 0$ adapted to the filtration $\mathbb{F} = \{\mathcal{F}_i\}_{i=1}^n$. Let $d_i = X_i - \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$ and let $S_m = \sum_{i=1}^m d_i$ for all m. Suppose that $\mathbf{Var}[X_i|\mathcal{F}_{i-1}] \leq v$. Then, for any $\delta \geq 0$

$$\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : S_m \geq \delta\right) \leq \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

$$\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : S_m \leq -\delta\right) \leq \exp\left(-\frac{\delta}{B}\psi\left(\frac{B\delta}{nv}\right)\right),$$

where $\psi(x) = (1 + 1/x)\ln(1 + x) - 1$.

Proof: Let $M_m = \exp(\lambda S_m/B)$ for each $m \in \{1, \ldots, n\}$. Since $(e^x - x - 1)/x^2$ is a non-decreasing function of $x \in \mathbb{R}$, for any random variable $X \in [-1, 1]$, by comparing the function value at $x = X$ and $x = 1$, we have

$$\exp(\lambda X) - \lambda X - 1 \leq X^2(e^\lambda - \lambda - 1).$$

If $\mathbb{E}[X] = 0$, we take expectations on both sides of the inequality and rearrange the equation to get

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &\leq 1 + \lambda\mathbb{E}[X] + \mathbb{E}X^2(e^\lambda - \lambda - 1) \\ &= 1 + \mathbf{Var}[X](e^\lambda - \lambda - 1).\end{aligned}$$

Since $d_m/B \in [-1, 1]$, we substitute d_m/B into X to get

$$\begin{aligned}\mathbb{E}[M_m|\mathcal{F}_{m-1}] &= M_{m-1}\mathbb{E}[\exp(\lambda d_m/B) \mid \mathcal{F}_{m-1}] \\ &\leq M_{m-1}\left(1 + \frac{v(e^\lambda - \lambda - 1)}{B^2}\right),\end{aligned}$$

where the second inequality is due to $\mathbf{Var}[d_m/B \mid \mathcal{F}_{m-1}] \leq v/B^2$ and $\mathbb{E}[d_m \mid \mathcal{F}_{m-1}] = 0$. Thus, we have

$$\mathbb{E}[M_m] \leq \mathbb{E}[M_{m-1}]\left(1 + \frac{v(e^\lambda - \lambda - 1)}{B^2}\right).$$

Applying the above inequality iteratively from n to 1, we have

$$\ln\mathbb{E}[M_n] \leq n\ln\left(1 + \frac{v(e^\lambda - \lambda - 1)}{B^2}\right)$$

$$\leq nv(e^{\lambda} - \lambda - 1)/B^2,$$

where the second inequality is due to the fact $\ln(1+x) \leq x$. Then, using Lemma 5, we have that for any $\lambda > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\exists m \in \{1,\ldots,n\} : S_m \geq \delta\right) \\
&= \mathbb{P}\left(\max_{m\in\{1,\ldots,n\}} M_m \geq \exp(\lambda\delta/B)\right) \\
&\leq \frac{\mathbb{E}[M_n]}{\exp(\lambda\delta/B)} \leq \exp\left(\frac{nv(e^{\lambda}-\lambda-1)}{B^2} - \frac{\lambda\delta}{B}\right).
\end{aligned}$$

Taking $\lambda = \ln(1+\frac{B\delta}{nv})$ to minimize the last term, the probability is no greater than $\exp(-\frac{\delta}{B}\psi(\frac{B\delta}{nv}))$. The lower tail bound corresponding to $S_m \leq -\delta$ can be proved similarly by reversing the sign of d_i. ■

A PROOF OF LEMMA 3

For any $t \in \mathcal{T}$, let u_i^{kt} be the i-th time slot the arm k is selected within $\mathcal{W}_t$ and let $d_i^{kt} = X^k_{u_i^{kt}} - \mu^k_{u_i^{kt}}$. We have

$$\mathbb{P}(A) \leq \mathbb{P}\left(\exists m \in \{l,\ldots,\tau\} : \frac{1}{m}\sum_{i=1}^{m} d_i^{kt} \leq -x - c_m\right),$$

Let $a = \sqrt{2\eta}$ such that $a > 1$. We now apply a peeling argument [53, Sec 2.2] with geometric grid $a^s l < m \leq a^{s+1} l$ over $\{l,\ldots,\tau\}$. Since c_m is monotonically decreasing in m,

$$\begin{aligned}
&\mathbb{P}\left(\exists m \in \{l,\ldots,\tau\} : \frac{1}{m}\sum_{i=1}^{m} d_i^{kt} \leq -x - c_m\right) \\
&\leq \sum_{s\geq 0} \mathbb{P}\left(\exists m \in [a^s l, a^{s+1} l) : \sum_{i=1}^{m} d_i^{kt} \leq -a^s l\left(x + c_{a^{s+1}l}\right)\right).
\end{aligned}$$

According to Lemma 6, the above summand is no greater than

$$\begin{aligned}
&\sum_{s\geq 0} \mathbb{P}\left(\exists m \in [1, a^{s+1}l) : \sum_{i=1}^{m} d_i^{kt} \leq -a^s l\left(x + c_{a^{s+1}l}\right)\right) \\
&\leq \sum_{s\geq 0} \exp\left(-2\frac{a^{2s} l^2}{\lfloor a^{s+1} l\rfloor}\left(x^2 + c^2_{a^{s+1}l}\right)\right) \\
&\leq \sum_{s\geq 0} \exp\left(-2a^{s-1} l x^2 - \frac{2\eta}{a^2}\ln\left(\frac{\tau}{K a^{s+1} l}\right)\right) \\
&= \sum_{s\geq 1} \frac{K l a^s}{\tau} \exp\left(-2a^{s-2} l x^2\right).
\end{aligned}$$

Let $b = 2x^2\, l/a^2$. It follows that

$$\begin{aligned}
&\sum_{s\geq 1} \frac{K l a^s}{\tau} \exp\left(-b a^s\right) \leq \frac{Kl}{\tau}\int_0^{+\infty} a^{y+1} \exp\left(-b a^y\right) dy \\
&= \frac{Kla}{\tau \ln(a)} \int_1^{+\infty} \exp(-bz) dz = \frac{Klae^{-b}}{\tau b \ln(a)},
\end{aligned}$$

where we apply change of variable $z = a^y$. We conclude the bound for the probability of event A. By using the upper tail bound, a similar result exists for event B.

B PROOF OF LEMMA 4

The saturated empirical mean is a key component of the UCB index of SW-RMOSS. Thus, the following properties for truncated random variables are used in the proof.

Lemma 8: Let X be a random variable with expected value μ and $\mathbb{E}[X^2] \leq 1$. Let $d := \text{sat}(X,B) - \mathbb{E}[\text{sat}(X,B)]$. Then for any $B > 0$, it satisfies (i) $|d| \leq 2B$ (ii) $\mathbb{E}[d^2] \leq 1$ (iii) $|\mathbb{E}[\text{sat}(X,B)] - \mu| \leq 1/B$.

Proof: Property i) follows immediately from the definition of d and property ii) follows from

$$\mathbb{E}[d^2] \leq \mathbb{E}\left[\text{sat}^2(X,B)\right] \leq \mathbb{E}\left[X^2\right].$$

For iii), since $\mu = \mathbb{E}[X(\mathbf{1}\{|X| \leq B\} + \mathbf{1}\{|X| > B\})]$,

$$\begin{aligned}
|\mathbb{E}[\text{sat}(X,B)] - \mu| &\leq \mathbb{E}\left[(|X| - B)\,\mathbf{1}\{|X| > B\}\right] \\
&\leq \mathbb{E}\left[|X|\,\mathbf{1}\{|X| > B\}\right] \leq \mathbb{E}\left[X^2/B\right].
\end{aligned}$$

■

Proof of Lemma 4: Recall that u_i^{kt} is the i-th time slot when arm k is selected within $\mathcal{W}_t$. Since c_m is a monotonically decreasing in m, $1/B_m = c_{h(m)} \leq c_m$ due to $h(m) \geq m$. Then, it follows from property iii) in Lemma 8 that

$$\begin{aligned}
\mathbb{P}(A) &\leq \mathbb{P}\left(\exists m \in \{l,\ldots,\tau\} : \bar{\mu}_m^k \leq \sum_{i=1}^{m} \frac{\mu^k_{u_i^{kt}}}{m} - (1+\zeta)c_m - x\right) \\
&\leq \mathbb{P}\left(\exists m \in \{l,\ldots,\tau\} : \sum_{i=1}^{m} \frac{\bar{d}_{im}^{kt}}{m} \leq \frac{1}{B_m} - (1+\zeta)c_m - x\right) \\
&\leq \mathbb{P}\left(\exists m \in \{l,\ldots,\tau\} : \frac{1}{m}\sum_{i=1}^{m} \bar{d}_{im}^{kt} \leq -x - \zeta c_m\right), \quad (29)
\end{aligned}$$

where $\bar{d}_{im}^{kt} = \text{sat}(X^k_{u_i^{kt}}, B_m) - \mathbb{E}[\text{sat}(X^k_{u_i^{kt}}, B_m)]$. Recall we select $a > 1$. Again, we apply a peeling argument with geometric grid $a^s \leq m < a^{s+1}$ over time interval $\{l,\ldots,\tau\}$. Let $s_0 = \lfloor \log_a(l) \rfloor$. Since c_m is monotonically decreasing with m, we have

$$(29) \leq \sum_{s\geq s_0} \mathbb{P}\left(\exists m \in [a^s, a^{s+1}) : \sum_{i=1}^{m} \bar{d}_{im}^{kt} \leq -a^s\left(x + \zeta c_{a^{s+1}}\right)\right).$$

For all $m \in [a^s, a^{s+1})$, since $B_m = B_{a^s}$, from Lemma 8 we know $|\bar{d}_{im}^{kt}| \leq 2B_{a^s}$ and $\mathbf{Var}[\bar{d}_{im}^{kt}] \leq 1$. Continuing from the previous step, we apply Lemma 7 to get

$$\begin{aligned}
(29) &\leq \sum_{s\geq s_0} \exp\left(-\frac{a^s\left(x + \zeta c_{a^{s+1}}\right)}{2B_{a^s}} \psi\left(\frac{2B_{a^s}}{a}\left(x + \zeta c_{a^{s+1}}\right)\right)\right) \\
&\quad \left(\text{since } \psi(x) \text{ is monotonically increasing}\right) \\
&\leq \sum_{s\geq s_0} \exp\left(-\frac{a^s\left(x + \zeta c_{a^{s+1}}\right)}{2B_{a^s}} \psi\left(\frac{2\zeta}{a} B_{a^s} c_{a^{s+1}}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\left(\text{substituting} c_{a^{s+1}},\, B_{a^s} \text{and using} h(a^s) = a^{s+1}\right) \\
&= \sum_{s \geq s_0+1} \exp\left(-a^s \left(\frac{x}{B_{a^{s-1}}} + \zeta c_{a^s}^2\right) \frac{\psi\left(2\zeta/a\right)}{2a}\right) \\
&\quad (\text{since } \zeta\psi(2\zeta/a) \geq 2a) \\
&\leq \frac{K}{\tau} \sum_{s \geq s_0+1} a^s \exp\left(-a^s \frac{x}{B_{a^{s-1}}} \frac{\psi\left(2\zeta/a\right)}{2a}\right). \qquad (30)
\end{aligned}
$$

Let $b = x\psi(2\zeta/a)/(2a)$. Since $\ln_+(x) \geq 1$ for all $x > 0$,

$$
\begin{aligned}
(30) &\leq \frac{K}{\tau} \sum_{s \geq s_0+1} a^s \exp\left(-b\sqrt{a^s}\right) \\
&\leq \frac{K}{\tau} \int_{s_0+1}^{+\infty} a^y \exp\left(-b\sqrt{a^{y-1}}\right) dy \\
&= \frac{K}{\tau} a \int_{s_0}^{+\infty} a^y \exp\left(-b\sqrt{a^y}\right) dy \\
&= \frac{K}{\tau} \frac{2a}{\ln(a) b^2} \int_{b\sqrt{a^{s_0}}}^{+\infty} z \exp(-z)\, dz \,(\text{where } z = b\sqrt{a^y}) \\
&\leq \frac{K}{\tau} \frac{2a}{\ln(a) b^2} \left(b\sqrt{a^{s_0}} + 1\right) \exp\left(-b\sqrt{a^{s_0}}\right),
\end{aligned}
$$

which concludes the proof.

REFERENCES

[1] A. B. H. Alaya-Feki, E. Moulines, and A. LeCornec, "Dynamic spectrum access with non-stationary multi-armed bandit," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, 2008, pp. 416–420.

[2] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.

[3] Y. Li, Q. Hu, and N. Li, "A reliability-aware multi-armed bandit approach to learn and select users in demand response," *Automatica*, vol. 119, 2020, Art. no. 109015.

[4] D. Kalathil and R. Rajagopal, "Online learning for demand response," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, 2015, pp. 218–222.

[5] J. R. Krebs, A. Kacelnik, and P. Taylor, "Test of optimal sampling by foraging great tits," *Nature*, vol. 275, no. 5675, pp. 27–31, 1978.

[6] V. Srivastava, P. Reverdy, and N. E. Leonard, "On optimal foraging and multi-armed bandits," in *Proc. IEEE Annu. Allerton Conf. Commun., Control, Comput.*, 2013, pp. 494–499.

[7] V. Srivastava, P. Reverdy, and N. E. Leonard, "Surveillance in an abruptly changing world via multiarmed bandits," in *Proc. IEEE Conf. Decis. Control*, 2014, pp. 692–697.

[8] C. Baykal, G. Rosman, S. Claici, and D. Rus, "Persistent surveillance of events with unknown, time-varying statistics," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 2682–2689.

[9] M. Y. Cheung, J. Leighton, and F. S. Hover, "Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 3368–3373.

[10] D. Agarwal et al., "Online models for content optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 17–24.

[11] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 661–670.

[12] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[13] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[14] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Adv. Appl. Math.*, vol. 17, no. 2, pp. 122–142, 1996.

[15] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[16] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. Annu. Conf. Learn. Theory*, 2011, pp. 359–376.

[17] P. Auer, Y. F. N. Cesa-Bianchi, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[18] D. E. Knuth, "Big omicron and big omega and big theta," *ACM Sigact News*, vol. 8, no. 2, pp. 18–24, 1976.

[19] L. Kocsis and C. Szepesvári, "Discounted UCB," in *Proc. 2nd PASCAL Challenges Workshop*, 2006, pp. 784–791.

[20] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 174–188.

[21] A. Chaudhary, A. Rai, and A. Gupta, "Maximizing success rate of payment routing using non-stationary bandits," 2023, *arXiv:2308.01028*.

[22] L. Wei and V. Srivastava, "On abruptly-changing and slowly-varying multiarmed bandit problems," in *Proc. Amer. Control Conf.*, Milwaukee, WI, USA, 2018, pp. 6291–6296.

[23] C. Hartland, N. Baskiotis, S. Gelly, M. Sebag, and O. Teytaud, "Change point detection and meta-bandits for online learning in dynamic environments," in *Proc. Conf. Francophone Sur L'Apprentissage Automatique*, Jul. 2007, pp. 237–250.

[24] F. Liu, J. Lee, and N. Shroff, "A change-detection based framework for piecewise-stationary multi-armed bandit problem," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3651–3658.

[25] L. Besson and E. Kaufmann, "The generalized likelihood ratio test meets klUCB: An improved algorithm for piece-wise non-stationary bandits," 2019, *arXiv:1902.01575*.

[26] Y. Cao, Z. Wen, B. Kveton, and Y. Xie, "Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2019, pp. 418–427.

[27] J. Mellor and J. Shapiro, "Thompson sampling in switching environments with bayesian online change detection," in *Proc. Artif. Intell. Statist.*, 2013, pp. 442–450.

[28] Y. Qin, T. Menara, S. Oymak, S. Ching, and F. Pasqualetti, "Non-stationary representation learning in sequential linear bandits," *IEEE Open J. Control Syst.*, vol. 1, pp. 41–56, 2022.

[29] A. Slivkins and E. Upfal, "Adapting to a changing environment: The Brownian restless bandits," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 343–354.

[30] J. Gornet, M. Hosseinzadeh, and B. Sinopoli, "Stochastic multi-armed bandits with non-stationary rewards generated by a linear dynamical system," in *Proc. IEEE Conf. Decis. Control*, 2022, pp. 1460–1465.

[31] O. Besbes and Y. Gur, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 199–207.

[32] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards," *Stochastic Syst.*, vol. 9, no. 4, pp. 319–337, 2019.

[33] V. Raj and S. Kalyani, "Taming non-stationary bandits: A Bayesian approach," 2017, *arXiv:1707.09727*.

[34] P. Auer, P. Gajane, and R. Ortner, "Adaptively tracking the best bandit arm with an unknown number of distribution changes," in *Proc. Annu. Conf. Learn. Theory*, 2019, pp. 138–158.

[35] Y. Chen, C.-W. Lee, H. Luo, and C.-Y. Wei, "A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free," in *Proc. Annu. Conf. Learn. Theory*, 2019, pp. 696–726.

[36] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Hedging the drift: Learning to optimize under nonstationarity," *Manage. Sci.*, vol. 68, no. 3, pp. 1696–1713, 2022.

[37] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Modern Phys.*, vol. 74, no. 1, 2002, Art. no. 47.

[38] M. Vidyasagar, "Law of large numbers, heavy-tailed distributions, and the recent financial crisis," in *Perspectives in Mathematical System Theory, Control, and Signal Processing*. Berlin, Germany: Springer, 2010, pp. 285–295.

[39] J. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. Annu. Conf. Learn. Theory*, 2009, pp. 217–226.
[40] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7711–7717, Nov. 2013.
[41] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[42] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proc. IEEE Annu. Found. Comput. Sci.*, 1995, pp. 322–331.
[43] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, pp. 623–648, 2004.
[44] P. Zhao, L. Zhang, Y. Jiang, and Z.-H. Zhou, "A simple approach for non-stationary linear bandits," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 746–755.
[45] Y. Russac, C. Vernade, and O. Cappé, "Weighted linear bandits for non-stationary environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12017–12026.
[46] L. Wei and V. Srivastava, "Minimax policy for heavy-tailed bandits," *IEEE Control Syst. Lett.*, vol. 5, no. 4, pp. 1423–1428, 2021.
[47] L. Besson, "SMPyBandits: An open-source research framework for single and multi-player multi-arms bandits (MAB) algorithms in python," 2018. [Online]. Available: https://github.com/SMPyBandits/SMPyBandits
[48] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.
[49] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.
[50] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'UCB: An optimal exploration algorithm for multi-armed bandits," in *Proc. Conf. Learn. Theory*, 2014, pp. 423–439.
[51] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. J., Second Ser.*, vol. 19, no. 3, pp. 357–367, 1967.
[52] G. Bennett, "Probability inequalities for the sum of independent random variables," *J. Amer. Stat. Assoc.*, vol. 57, no. 297, pp. 33–45, 1962.
[53] S. Bubeck, "Bandits games and clustering foundations," Theses, Université des Sciences et Technologie de Lille - Lille I, 2010. [Online]. Available: https://tel.archives-ouvertes.fr/tel-00845565

LAI WEI received the B.E. degree in mechanical engineering from the Harbin Institute of Technology, Harbin, China, in 2012, the M.S. degree in mechanical engineering from the Beijing University of Aeronautics & Astronautics, Beijing, China, in 2015, the M.S. degree in computer science, and the Ph.D. degree in electrical and computer engineering from Michigan State University, East Lansing, MI, USA. He is currently a Postdoc Research Fellow with the Life Sciences Institute, University of Michigan, Ann Arbor, MI. His research interests include control and learning in robotic systems, decision-making theory, and causal inference.

VAIBHAV SRIVASTAVA (Senior Member, IEEE) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 2007, the M.S. degree in mechanical engineering, the M.A. degree in statistics, and the Ph.D. degree in mechanical engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2011, 2012, and 2012, respectively. During 2013–2016, he was a Lecturer and Associate Research Scholar with the Mechanical and Aerospace Engineering Department, Princeton University, Princeton, NJ, USA. He is currently an Associate Professor of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA. He is also affiliated with Mechanical Engineering, Cognitive Science Program, and Connected and Autonomous Networked Vehicles for Active Safety (CANVAS). His research interests include cyber physical human systems with emphasis on mixed human-robot systems, networked multi-agent systems, aerial robotics, and connected and autonomous vehicles.