On Abruptly-Changing and Slowly-Varying Multiarmed Bandit Problems

Lai Wei Vaibhav Srivastava

***Abstract*— We study the non-stationary stochastic multi-armed bandit (MAB) problem and propose two generic algorithms, namely, Limited Memory Deterministic Sequencing of Exploration and Exploitation (LM-DSEE) and Sliding-Window Upper Confidence Bound# (SW-UCB#). We rigorously analyze these algorithms in abruptly-changing and slowly-varying environments and characterize their performance. We show that the expected cumulative regret for these algorithms in either of the environments is upper bounded by sublinear functions of time, i.e., the time average of the regret asymptotically converges to zero. We complement our analysis with numerical illustrations.**

I. Introduction

Decision-making in uncertain and non-stationary environments is one of the most fundamental problems across scientific disciplines, including economics, social science, neuroscience and ecology, and often requires balancing several decision-making tradeoffs, such as speed-versus-accuracy, robustness-versus-efficiency, and explore-versus-exploit. The MAB problem is a prototypical example of the explore-versus-exploit tradeoff: choosing between the most informative and seemingly the most rewarding alternative.

In an MAB problem, a decision-maker sequentially allocates a single resource by repeatedly choosing one among a set of competing alternative arms (options). These problems have been applied in several interesting areas such as robotic foraging and surveillance [1]–[3], acoustic relay positioning for underwater communication [4], and channel allocation in communication networks [5]. In a standard MAB problem, a stationary environment is considered, however, many application areas are inherently non-stationary. In this paper, we seek to address this gap and study the MAB problem in two classes of non-stationary environments: (i) abruptly-changing environment and (ii) slowly-varying environment.

The performance of a sequential allocation policy for the MAB problem is characterized in terms of the expected cumulative regret which is defined as the cumulative sum of the difference between the maximum mean reward and the mean reward at the arm selected by the policy at each time. An algorithm for the MAB problem is said to be efficient if it achieves a sublinear expected cumulative regret, i.e., the time average of the regret asymptotically converges to zero.

Some classes of non-stationary MAB have been studied in the literature. In [6], authors study a non-stochastic MAB problem in which the rewards are deterministic and non-stationary. They study a weaker notion of the regret, wherein the policy generated by the algorithm is compared against the best policy within the policies that select the same arm at each time. In a recent work [7], the algorithms developed in [6] are adapted to handle a class of non-stationary environments and upper bounds on the standard notion of the regret are derived. In [8], authors study a class of non-stationary MAB problems in which the mean rewards at arms may switch abruptly at unknown times to unknown values. They design an upper confidence bound (UCB) based algorithm that relies on estimates of the mean rewards from a recent time-window of observations. In [9], authors study the MAB problem in a piecewise-stationary environment. They use active detection algorithms to determine the change-points and restart the UCB algorithm.

An application area of interest for the MAB problem is robotic search and surveillance in which a robot is routed to collect stochastic rewards [10], [11]. These rewards may correspond to, for example, likelihood of an anomaly at a spatial location, concentration of a certain type of algae in the ocean, etc. MAB algorithms have been extended to these problems by introducing block-allocation strategies that seek to balance the explore-exploit tradeoff using sufficiently small travel time [12], [13]. In [3], authors extended the algorithm in [8] to develop block-allocation strategies for the MAB problem with abruptly-changing reward.

While the above algorithms balance the explore-exploit tradeoff while ensuring sufficiently small travel time, they are reactive in the sense that they select only one arm at a time, i.e., they only provide information about the next location to be visited by the robot. Certain motion constraints on the robots such as non-holonomicity may make such movements energetically demanding. Therefore, we seek algorithms that have a deterministic and predictable structure which can be leveraged to design trajectories for the robot that can be efficiently traversed even under motion constraints. Towards this end, we focus on DSEE algorithms [14]–[16].

In this paper, we study the MAB problem in abruptly-changing and slowly-varying environments, and develop upper confidence bound type and DSEE type algorithms for these environments. Our assumptions on the environment are similar to those in [7] and [8], but we focus on alternative algorithms which include algorithms with deterministic structure as discussed above. In particular, we extend the DSEE algorithm to non-stationary environments and develop the LM-DSEE algorithm. We also extend the SW-UCB algorithm, developed and analyzed for abruptly-changing environments in [8], to the SW-UCB# algorithm for non-stationary environments.

The major contributions of this paper are threefold. First, in Section III, we develop two novel algorithms: the LM-

This work has been supported by NSF Award IIS-1734272.

L. Wei and V. Srivastava are with the Department of Electrical and Computer Engineering. Michigan State University, East Lansing, MI 48823 USA. e-mail: weilai1@msu.edu; e-mail: vaibhav@egr.msu.edu

DSEE and the SW-UCB# for the non-stationary MAB problem. Second, in Sections IV and V, we analyze the LM-DSEE and the SW-UCB# algorithms for abruptly-changing and slowly-varying environments and establish upper bounds on the expected cumulative regret. Third, in Section VI, we illustrate our analysis using numerical examples.

II. Background & Problem Description

In this section, we recall the stationary stochastic MAB problem, and introduce the stochastic MAB problem in abruptly-changing and slowly-varying environments.

A. The stationary stochastic MAB problem

Consider an N-armed bandit problem, i.e., an MAB problem with N arms. The reward associated with arm $j \in \{1, \ldots, N\}$ is a random variable with bounded support $[0, 1]$ and an unknown stationary mean $\mu_j \in [0, 1]$. Let the decision-making agent choose arm j_t at time $t \in \{1, \ldots, T\}$ and receive a reward r_t associated with the arm. The decision-maker's objective is to choose a sequence of arms $\{j_t\}_{t\in\{1,\ldots,T\}}$ that maximizes the expected cumulative reward $\sum_{t=1}^{T} \mu_{j_t}$, where T is the horizon length of the sequential allocation process.

For an MAB problem, the expected *regret* at time t is defined by $\mu_{j^*} - \mu_{j_t}$, where $\mu_{j^*} = \max\{\mu_j \mid j \in \{1, \ldots, N\}\}$. The objective of the decision-maker can be equivalently defined as minimizing the expected cumulative regret defined by $R(T) = \sum_{t=1}^{T} \mathbb{E}[\mu_{j^*} - \mu_{j_t}] = \sum_{j=1}^{N} \Delta_j \mathbb{E}[n_j(T)]$, where $n_j(T)$ is the cumulative number of times a suboptimal arm j has been chosen until time T and $\Delta_j = \mu_{j^*} - \mu_j$ is the expected regret due to picking arm j instead of arm j^*.

B. Algorithms for the stationary stochastic MAB problem

We recall two state-of-the-art algorithms for the stationary stochastic MAB problem relevant to this paper: (i) the UCB algorithm, and (ii) the DSEE algorithm.

The UCB algorithm maintains a statistical estimate of the mean rewards associated with each arm. It initializes by selecting each arm once and subsequently selects the arm j_t at time t defined by

$$j_t \in \arg\max \Big\{ \bar{r}_j(t-1) + \sqrt{\frac{2\ln(t-1)}{n_j(t-1)}} \Big| j \in \{1, \ldots, N\} \Big\},$$

where $\bar{r}_j(t-1)$ is the statistical mean of the rewards received at arm j until time t. Auer [17] showed that the UCB algorithm achieves expected cumulative regret that is within a constant factor of the optimal.

The DSEE algorithm divides the set of natural numbers $\mathbb{N}$ into interleaving epochs of exploration and exploitation [14]. In the exploration epoch, each arm is played in a round-robin fashion, while in the exploitation epoch, only the arm with the maximum statistical mean reward is played. For an appropriately defined $w \in \mathbb{R}_{>0}$, the DSEE algorithm at time t exploits if number of exploration steps until time $t-1$ are greater than or equal to $N\lceil w \log t \rceil$, otherwise it starts a new exploration epoch. In [14], Vakili *et al.* derived bounds on the regret of the DSEE algorithm.

C. The non-stationary stochastic MAB problem

The non-stationary stochastic MAB problem is the stochastic MAB problem in which the mean reward at each arm is changing with time. Let the mean reward associated with arm j at time t be $\mu_j(t) \in [0, 1]$. The decision-maker's objective is to choose a sequence of arms $\{j_t\}_{t\in\{1,\ldots,T\}}$ that maximizes the expected cumulative reward $\sum_{t=1}^{T} \mu_{j_t}(t)$, where T is the horizon length of the sequential allocation process. We will characterize the performance of algorithms for these problems using the notion of the expected cumulative regret defined by

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \mathbb{E}[\mu_{j_t^*}(t) - \mu_{j_t}(t)] \\ &= \sum_{t=1}^{T} \mu_{j_t^*}(t) - \mathbb{E}\Big[\sum_{j=1}^{N}\sum_{t=1}^{T} \mathbf{1}_{\{j_t=j\}} \mu_j(t)\Big], \end{aligned}$$

where $\mu_{j_t^*}(t) = \max_{j\in\{1,\ldots,N\}} \mu_j(t)$, $\mathbf{1}_{\{.\}}$ is the indicator function and the expectation is computed over different realizations of j_t. For brevity, in the following, we will refer to $R(T)$ simply as the regret.

In this paper, we study the above MAB problem for two classes of non-stationary environments:

Abruptly-changing environment: In an abruptly-changing environment, the mean rewards from arms switch to unknown values at unknown time-instants. We refer to these time-instants as *breakpoints*. We assume that the number of breakpoints until time T is $\Upsilon_T \in O(T^\nu)$, where $\nu \in [0, 1)$ and is known a priori.

Slowly-varying environment: In a slowly-varying environment, the change in the mean reward at each arm between any two subsequent time-instants is small and is upper bounded by $\epsilon_T \in O(T^{-\kappa})$, where $\kappa \in \mathbb{R}_{>0}$ and is known a priori. Here, lower values of κ correspond to higher changes in the mean reward at subsequent time-instants. We refer to ϵ_T as the non-stationarity parameter.

III. Algorithms for Non-Stationary Stochastic MAB Problem

In this section, we present two algorithms for the non-stationary stochastic MAB problem: the Limited-Memory DSEE (LM-DSEE) algorithm and the Sliding-Window UCB# (SW-UCB#) algorithm. These algorithms are generic and require some parameters to be tuned based on environment characteristics.

A. The LM-DSEE algorithm

The LM-DSEE algorithm comprises interleaving epochs of exploration and exploitation. In the k-th exploration epoch, each arm is sampled $L(k) = \lceil \gamma \ln(k^\rho l b) \rceil$ number of times. In the k-th exploitation epoch, the arm with the highest sample mean in the k-th exploration epoch is sampled $\lceil a k^\rho l \rceil - NL(k)$ times. Here, parameters ρ, γ, a, b and l are tuned based on the environment characteristics (see Algorithm 1 for details). In the following, we will set a and b to unity for the purposes of analysis. Parameters a and b

Algorithm 1: *The LM-DSEE Algorithm*

For abruptly-changing environment

Input : $\nu \in [0,1), \Delta_{\min} \in (0,1)$, $T \in \mathbb{N}, a \in \mathbb{R}_{>0}, b \in (0,1]$;

Set : $\gamma \geq \frac{2}{\Delta_{\min}^2}$, $l \in \{\frac{N}{a}\lceil \gamma \ln lb \rceil, \ldots, +\infty\}$, and $\rho = \frac{1-\nu}{1+\nu}$;

For slowly-varying environment

Input : $\kappa \in \mathbb{R}_{>0}$, $\kappa_{\max} \in (0, \frac{4}{3}), T \in \mathbb{N}, a \in \mathbb{R}_{>0}$, $b \in (0,1]$;

Set : $\tilde{\kappa} \leftarrow \min\{\kappa, \kappa_{\max}\}, \rho \leftarrow \frac{3\tilde{\kappa}}{4-3\tilde{\kappa}}$, $l \in \{\frac{N}{a}\lceil l^{\frac{2}{3}} \ln lb \rceil, \ldots, +\infty\}$, and $\gamma = 2(k^\rho l)^{\frac{2}{3}}$;

Output : sequence of arm selection;

% Initialization:

Set batch index $k \leftarrow 1$ and $t \leftarrow 1$;

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

  % Exploration

  for $j \in \{1, \ldots, N\}$ **do**

    Pick arm j, $L(k) \leftarrow \lceil \gamma \ln(k^\rho lb) \rceil$ times ;

    collect rewards $\{\hat{r}_j^i(k)\}_{i \in \{1,\ldots,L(k)\}}$;

    compute sample mean $\bar{r}_j^{\text{epch}}(k) \leftarrow \frac{1}{L(k)} \sum_{i=1}^{L(k)} \hat{r}_j^i(k)$;

  % Exploitation

  Select the best arm $j_k^{\text{epch}} = \arg\max_{j \in \{1,\ldots,N\}} \bar{r}_j^{\text{epch}}(k)$;

  Pick arm j_k^{epch}, $\lceil ak^\rho l \rceil - NL(k)$ times ;

  Update batch index $k \leftarrow k+1$ and $t \leftarrow t + \lceil ak^\rho l \rceil$

do not influence the order of regret bounds derived below, but they can be tuned to enhance the transient performance.

The LM-DSEE algorithm is similar in spirit to the DSEE algorithms [14], [15], wherein the length of exploitation epoch increases exponentially with the epoch number and all the data collected in the previous exploration epochs is used to estimate the mean rewards. However, in a non-stationary environment, using all the rewards from the previous exploration epochs may lead to a heavily biased estimate of the mean rewards. Furthermore, an exponentially increasing exploitation epoch length may lead to excessive exploitation based on an outdated estimate of the mean rewards. To address these issues, we modify the DSEE algorithm by using only the rewards from the current exploration epoch to estimate the mean rewards, and we increase the length of exploitation epoch using a power law.

B. The SW-UCB# algorithm

The SW-UCB# algorithm is an adaptation of the SW-UCB algorithm proposed and studied in [8]. The SW-UCB# algorithm, at time t, maintains an estimate of the mean reward $\bar{r}_j(t,\alpha)$ at each arm j, using only the rewards collected within a sliding-window of observations. Let the width of the sliding-window at time $t \in \{1, \ldots, T\}$ be $\tau(t,\alpha) = \min\{\lceil \lambda t^\alpha \rceil, t\}$, where parameters $\alpha \in (0,1]$ and $\lambda \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$ are tuned based on environment characteristics. Let $n_j(t,\alpha) = \sum_{s=t-\tau(t,\alpha)+1}^{t} \mathbf{1}_{\{j_s=j\}}$ be the number of times arm j has been selected in the time-window

Algorithm 2: *The SW-UCB# Algorithm*

For abruptly-changing environment

Input : $\nu \in [0,1)$, $\Delta_{\min} \in (0,1)$, $\lambda \in \mathbb{R}_{>0}$ & $T \in \mathbb{N}$;

Set : $\alpha = \frac{1-\nu}{2}$

For slowly-varying environment

Input : $\kappa \in \mathbb{R}_{>0}$, $\lambda \in \mathbb{R}_{>0}$ & $T \in \mathbb{N}$;

Set : $\alpha = \min\{1, \frac{3\kappa}{4}\}$

Output : sequence of arm selection;

% Initialization:

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

  if $t \in \{1, \ldots, N\}$ **then**

    Pick arm $j_t = t$;

  else

    Pick arm j_t defined in (1) ;

at time t, then

$$\bar{r}_j(t,\alpha) = \frac{1}{n_j(t,\alpha)} \sum_{s=t-\tau(t,\alpha)+1}^{t} r_j(s) \mathbf{1}_{\{j_s=j\}}.$$

Based on the above estimate, the SW-UCB algorithm at each time selects the arm

$$j_t = \arg\max\{\bar{r}_j(t-1,\alpha) + c_j(t-1,\alpha) \mid j \in \{1, \ldots, N\}\}, \tag{1}$$

where $c_j(t,\alpha) = \sqrt{\frac{(1+\alpha)\ln t}{n_j(t,\alpha)}}$. The details of the algorithm are presented in Algorithm 2.

In contrast to the SW-UCB algorithm [8], the SW-UCB# algorithm employs a time-varying width of the sliding-window. The tuning of the fixed window width in [8] requires a priori knowledge of the time horizon T which is no longer needed for the SW-UCB# algorithm.

IV. Analysis of the LM-DSEE algorithm

In this section, we analyze the performance of the LM-DSEE algorithm (Algorithm 1) in abruptly-changing and slowly-varying environments. Here, we only present the sketch of the proofs. For detailed proofs, see [18].

A. LM-DSEE in the abruptly-changing environment

Before we analyze the LM-DSEE algorithm in the abruptly-changing environment, we introduce the following notation. Let

$$\Delta_j = \max\{\mu_{j_t^*}(t) - \mu_j(t) \mid t \in \{1, \ldots, T\}\},$$
$$\Delta_{\max} = \max\{\Delta_j \mid j \in \{1, \ldots, N\}\},$$
$$\text{and } \Delta_{\min} = \min\{\mu_{j_t^*}(t) - \mu_j(t) \mid t \in \{1, \ldots, T\}, j \in \{1, \ldots, N\} \setminus \{j_t^*\}\}.$$

*Theorem 1 (**LM-DSEE in abruptly-changing environment**):* For the abruptly-changing environment with the number of breakpoints $\Upsilon_T \in O(T^\nu)$, $\nu \in [0,1)$ and the LM-DSEE algorithm, the expected cumulative regret

$$R^{\text{LM-DSEE}}(T) \in O(T^{\frac{1+\nu}{2}} \ln T).$$

Proof: Let K be the index of the epoch containing the time-instant T, then the length of each epoch is at most $\lceil K^\rho l \rceil$. Since breakpoints are located in at most Υ_T epochs, we can upper bound the regret from epochs containing breakpoints by $R_b \leq \Upsilon_T \lceil K^\rho l \rceil \Delta_{\max}$.

In the epochs containing no breakpoint, let R_e and R_i denote, respectively, the regret from exploration and exploitation epochs. Note that in such epochs the mean reward from each arm does not change. Then, the regret in exploration epochs R_e satisfies,

$$R_e \leq \sum_{k=1}^{K}\sum_{j=1}^{N} \lceil \gamma \ln(k^\rho l) \rceil \Delta_j \leq K \lceil \gamma \ln(K^\rho l) \rceil \sum_{j=1}^{N} \Delta_j.$$

In exploitation epochs, regret is incurred if the best arm is not selected, and consequently R_i satisfies

$$R_i \leq \sum_{k=1}^{K}\sum_{j=1}^{N} \big[\lceil k^\rho l \rceil - NL(k)\big] \mathbb{P}(j_k^{\text{epch}} = j \neq j^*_{\text{no-break}}(k)) \Delta_j, \tag{2}$$

where $j^*_{\text{no-break}}(k)$ is the best arm and j_k^{epch} is selected arm in the k-th exploitation epoch. It follows from the Chernoff-Hoeffding inequality [19, Theorem 1] that $\mathbb{P}(j_k^{\text{epch}} = j \neq j^*_{\text{no-break}}(k)) \leq 2(k^\rho l)^{-1}$. Substituting it into (2), we have $R_i \leq 2K \sum_{j=1}^{N} \Delta_j$. Furthermore, it can be seen that $K \in O(T^{\frac{1}{1+\rho}})$. Therefore, it follows that

$$\begin{aligned} R^{\text{LM-DSEE}}(T) &= R_b + R_e + R_i \\ &\leq \Upsilon_T K^\rho l \Delta_{\max} + K(\lceil \gamma \ln(K^\rho l) \rceil + 2) \textstyle\sum_{j=1}^{N} \Delta_j. \end{aligned}$$

Thus, the regret $R^{\text{LM-DSEE}}(T) \in O(T^{\frac{1+\nu}{2}} \ln T)$, and this establishes the theorem. ∎

B. LM-DSEE in the slowly-varying environment

*Theorem 2 (**LM-DSEE in slowly-varying environment**):* For the slowly-varying environment with the non-stationarity parameter $\epsilon_T = O(T^{-\kappa})$, $\kappa \in \mathbb{R}_{>0}$, and the LM-DSEE algorithm, the expected cumulative regret

$$R^{\text{LM-DSEE}}(T) \in O(T^{\frac{3+2\rho}{3+3\rho}} \ln T),$$

where $\rho = \frac{3\tilde{\kappa}}{4-3\tilde{\kappa}}$, $\tilde{\kappa} = \min\{\kappa, \kappa_{\max}\}$, and $\kappa_{\max} \in (0, \frac{4}{3})$.

Proof: Similar to the proof of Theorem 1, we divide the regret into R_e and R_i, the regret in the exploration epoch and the exploitation epoch, respectively. It follows that

$$\begin{aligned} R_e &\leq \sum_{j=1}^{N}\sum_{k=1}^{K} \lceil \gamma \ln(k^\rho l) \rceil \Delta_j \\ &\leq \Big[\frac{2l^{\frac{2}{3}}}{\frac{2}{3}\rho + 1} (K+1)^{\frac{2}{3}\rho+1} \ln(K^\rho l) + K \Big] \sum_{j=1}^{N} \Delta_j. \end{aligned}$$

Also, for the regret in the exploitation epoch, we have

$$R_i \leq \sum_{j=1}^{N}\sum_{k=1}^{K} \sum_{t \in \text{epoch } k} \mathbb{P}(j_k^{\text{epch}} = j \neq j_t^*) \big(\mu_{j_t^*}(t) - \mu_{j_k^{\text{epch}}}(t)\big). \tag{3}$$

In the context of slowly-varying environment, when the best arm switches, there may exist a period around the switching instant during which the difference in the mean rewards between the best arm and the next-best arm is extremely small. To handle such a situation, we define

$$J(t) = \{j \in \{1, \ldots, N\} \mid \mu_{j_t^*}(t) - \mu_j(t) \leq \sigma\},$$

where we set $\sigma = (k^\rho l)^{-\frac{1}{3}} + 2\varrho$ and $\varrho = \epsilon_T k^\rho l$, which is the maximum change in the mean reward at any arm in the k-th epoch. Then, it follows that

$$\begin{aligned} \mathbb{P}(j_k^{\text{epch}} = j \neq j_t^*) &= \mathbb{P}(j_k^{\text{epch}} = j \neq j_t^*, j \in J(t)) \\ &\quad + \mathbb{P}(j_k^{\text{epch}} = j \neq j_t^*, j \notin J(t)). \end{aligned}$$

Substituting it into (3), we obtain

$$R_i \leq \sum_{j=1}^{N}\sum_{k=1}^{K} \sum_{t \in \text{epoch } k} \big[\mathbb{P}\big(j_k^{\text{epch}} = j, j \notin J(t)\big)\Delta_j + \sigma\big].$$

Denote $\chi_{j,k}$ as the set of time indices at which the arm j is sampled in the k-th exploration epoch. Define

$$M_j(k) \triangleq \frac{1}{|\chi_{j,k}|} \sum_{t \in \chi_{j,k}} \mu_j(t).$$

Then, it can be shown that $\mu_{j_t^*}(t) - \mu_j(t) > \sigma$, for all $t \in$ epoch k, implies $M_{j_t^*}(k) - M_j(k) > \sigma - 2\varrho$. It follows from Chernoff-Hoeffding inequality [19, Theorem 1] that $\mathbb{P}(j_k^{\text{epch}} = j, j \notin J(t)) \leq 2(k^\rho l)^{-1}$ and consequently,

$$\begin{aligned} R_i &\leq \sum_{j=1}^{N}\sum_{k=1}^{K} \big[2(k^\rho l)^{-1}\Delta_j + (k^\rho l)^{-\frac{1}{3}} + 2\varrho\big]\big[\lceil k^\rho l \rceil - NL(k)\big] \\ &\leq \frac{Nl^{\frac{2}{3}}}{\frac{2}{3}\rho + 1}(K+1)^{\frac{2}{3}\rho+1} + 2K\sum_{j=1}^{N}\Delta_j + \frac{2Nl^2\epsilon_T}{1+2\rho}(K+1)^{1+2\rho}. \end{aligned}$$

Using the fact that $K \in O(T^{\frac{1}{1+\rho}})$, we have $R_i \in O(T^{\frac{3+2\rho}{3+3\rho}})$ and $R_e \in O(T^{\frac{3+2\rho}{3+3\rho}} \ln T)$. Thus, $R^{\text{LM-DSEE}}(T) \in O(T^{\frac{3+2\rho}{3+3\rho}} \ln T)$, and this concludes the proof. ∎

V. Analysis of the SW-UCB# algorithm

In this section, we analyze the performance of the SW-UCB# algorithm (Algorithm 2) in abruptly-changing and slowly-varying environments. Here, we only present the sketch of the proofs. For detailed proofs, see [18].

A. SW-UCB# in the abruptly-changing environment

*Theorem 3 (**SW-UCB# in abruptly-changing environment**):* For the abruptly-changing environment with the number of breakpoints $\Upsilon_T = O(T^\nu)$, $\nu \in [0, 1)$ and the SW-UCB# algorithm, the expected cumulative regret

$$R^{\text{SW-UCB\#}}(T) \in O(T^{\frac{1+\nu}{2}} \ln T).$$

Proof: We define $\mathcal{T}$ such that for all $t \in \mathcal{T}$, t is either a breakpoint or there exists a breakpoint in its sliding-window of observations $\{t - \tau(t-1, \alpha), \ldots, t-1\}$. For $t \in \mathcal{T}$, the estimate of the mean rewards may be significantly biased. It can be shown that $|\mathcal{T}| \leq \Upsilon_T \lceil \lambda(T-1)^\alpha \rceil$, and consequently, the regret can be upper bounded as follows:

$$R(T) \leq \sum_{j=1}^{N} \mathbb{E}[\tilde{N}_j(T)]\Delta_j + \Upsilon_T[\lambda(T-1)^\alpha + 1]\Delta_{\max}, \tag{4}$$

where $\tilde{N}_j(T) := \sum_{t=1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*,\, t\notin\mathcal{T}\}}$ satisfies

$$\begin{aligned}\tilde{N}_j(T) \leq 1 &+ \sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)<A(t-1)\}} \\ &+ \sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*,\, t\notin\mathcal{T},\, n_j(t-1,\alpha)\geq A(t-1)\}},\end{aligned} \tag{5}$$

where $A(t) = \frac{4(1+\alpha)\ln t}{\Delta_{\min}^2}$.

We first bound the second term on the right hand side of inequality (5). Let $G \in \mathbb{N}$ be such that

$$\lceil\lambda(1-\alpha)(G-1)\rceil^{\frac{1}{1-\alpha}} < T \leq \lceil\lambda(1-\alpha)G\rceil^{\frac{1}{1-\alpha}}. \tag{6}$$

Then, consider the following partition of time indices

$$\Big\{\big\{1+\lfloor\lceil\lambda(1-\alpha)(g-1)\rceil^{\frac{1}{1-\alpha}}\rfloor, \ldots, \lfloor\lceil\lambda(1-\alpha)g\rceil^{\frac{1}{1-\alpha}}\rfloor\big\}\Big\}_{g\in\{1,\ldots,G\}}. \tag{7}$$

In the g-th epoch in the partition, suppose there exist a time-instant t such that $j_t = j \neq j^*(t)$ and $n_j(t-1,\alpha) < A(t-1)$. Let the last time-instant satisfying these conditions in the g-th epoch be $t_j(g) = \max\{t \in g\text{-th epoch} | j_t = j \neq j_t^* \text{ and } n_j(t-1,\alpha) < A(t-1)\}$. It can be proved that

$$t_j(g) - \tau\big(t_j(g)-1,\alpha\big) \leq 2 + \lfloor\lambda(1-\alpha)(g-1)^{\frac{1}{1-\alpha}}\rfloor,$$

i.e., the first time-instant in the sliding-window at $t_j(g)$ is located at or to the left of the second time-instant of the g-th epoch in the partition (7). Therefore, it follows that

$$\begin{aligned}&\sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j^*, n_j(t-1,\alpha)<A(t-1)\}} \\ &\leq \sum_{g=1}^{G} \big[A(t_j(g)-1)+2\big] \leq G\Big(2 + \frac{4(1+\alpha)\ln T}{\Delta_{\min}^2}\Big).\end{aligned} \tag{8}$$

Next, we upper-bound expectation of the last term in (5). Note that when $t \notin \mathcal{T}$, for each $j \in \{1,\ldots,N\}$, $\mu_j(s)$ is a constant for all $s \in \{t-\tau(t-1,\alpha),\ldots,t\}$, and the problem reduces to the stationary MAB. Accordingly, we have

$$\mathbb{E}\Big[\sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j^*,\, t\notin\mathcal{T},\, n_j(t-1,\alpha)\geq A(t-1)\}}\Big] \leq \frac{(\lambda+1)^2\pi^2}{3}. \tag{9}$$

Therefore, it follows from (4), (5), (8), and (9) that

$$\begin{aligned}R(T) \leq &\sum_{j=1}^{N} \Big(G\big(2 + \frac{4(1+\alpha)\ln T}{\Delta_{\min}^2}\big) + 1 + \frac{(\lambda+1)^2\pi^2}{3}\big)\Delta_j \\ &+ \Upsilon_T\lceil\lambda(T-1)^{\alpha}\rceil\Big)\Delta_{\max}.\end{aligned}$$

From (6), we have $G = O(T^{1-\alpha})$, and this yields $R^{\text{SW-UCB\#}}(T) \in O(T^{\frac{1+\nu}{2}}\ln T)$. ∎

B. The SW-UCB# in the slowly-varying environment

*Theorem 4 (**SW-UCB# in slowly-varying environment**):* For slowly-varying environment with the non-stationarity parameter $\epsilon_T = O(T^{-\kappa})$, $\kappa \in \mathbb{R}_{>0}$, and the SW-UCB algorithm, the expected cumulative regret

$$R^{\text{SW-UCB\#}}(T) \in O(T^{1-\frac{\alpha}{3}}\ln T),$$

where $\alpha = \min\{1, \frac{3\kappa}{4}\}$.

Proof: We start by noting that the number of times arm j is selected when it is suboptimal satisfies

$$\begin{aligned}\hat{n}_j(T) \leq 1 &+ \sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)<A(t-1)\}} \\ &+ \sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)\geq A(t-1), j\notin J(t)\}} \\ &+ \sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)\geq A(t-1), j\in J(t)\}},\end{aligned} \tag{10}$$

where $\sigma = t^{-\frac{\alpha}{3}} + 2\lceil\lambda(t-1)^{\alpha}\rceil\epsilon_T$, $A(t) = 4t^{\frac{2\alpha}{3}}(1+\alpha)\ln t$, $n_j(t-1,\alpha)$ is defined in Algorithm 2, and $J(t) = \{j \in \{1,\ldots,N\} \mid \mu_{j_t^*}(t) - \mu_j(t) \leq \sigma\}$.

We first focus on the second term on the right hand side of (10), and bound it similarly to the proof of Theorem 3:

$$\begin{aligned}&\sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)<A(t-1)\}} \leq \sum_{g=1}^{G}[A(t_j(g)-1)+2] \\ &\leq \frac{12(1+\alpha)}{3-\alpha}\lambda^{\frac{2\alpha}{3-3\alpha}}[(1-\alpha)(G+1)]^{\frac{3-\alpha}{3-3\alpha}}\ln T + 2G,\end{aligned} \tag{11}$$

where G is defined in (6).

We now analyze the third term in (10). Let $M_j(t) = \frac{1}{n_j(t,\alpha)}\sum_{s=t-\lceil\lambda t^{\alpha}\rceil+1}^{t} \mu_j(s)\mathbf{1}_{\{j_s=j\}}$. Since the change in the mean reward for any arm within time-window $\{t-\tau(t-1,\alpha),\ldots,t\}$ is less than $\lceil\lambda(t-1)^{\alpha}\rceil\epsilon_T$ and $\mu_{j_t^*}(t) - \mu_j(t) > \sigma$, $M_{j_t^*}(t-1) - M_j(t-1) > t^{-\frac{\alpha}{3}}$. Consequently, the expected value of the third term in (10) satisfies

$$\mathbb{E}\left[\sum_{t=N+1}^{T} \mathbf{1}_{\{j_t=j\neq j_t^*, n_j(t-1,\alpha)\geq A(t-1), j\notin J(t)\}}\right] \leq \frac{(\lambda+1)^2\pi^2}{3}. \tag{12}$$

Therefore, from (10), (11), and (12), we have

$$\begin{aligned}R(T) \leq \sum_{j=1}^{N} \Delta_j\Big\{&\frac{12(1+\alpha)}{3-\alpha}\lambda^{\frac{2\alpha}{3-\alpha}}[(1-\alpha)(G+1)]^{\frac{3-\alpha}{3-3\alpha}}\ln T \\ &+ 2G + \frac{(\lambda+1)^2\pi^2}{3} + 1\Big\} + \sum_{t=1}^{T}\sigma.\end{aligned}$$

Since $G \in O(T^{1-\alpha})$, we have $R^{\text{SW-UCB\#}}(T) \in O(T^{1-\frac{\alpha}{3}}\ln T)$. ∎

VI. Numerical Illustration

In this section, we present simulation results for the SW-UCB# and LM-DSEE algorithms in both abruptly-changing and slowly-varying environments. For all the simulations, we consider a 10-armed bandit in which the reward at each arm is generated using Beta distribution. For the abruptly-changing environment, the breakpoints are introduced at time-instants where the next element of the sequence $\{\lfloor t^{\nu}\rfloor\}_{t\in\{1,\ldots,T\}}$ is different from the current element. At each breakpoint, the mean rewards at each arm are randomly selected from the set $\{0.05, 0.12, 0.19, 0.26, 0.33, 0.39, 0.46, 0.53, 0.6, 0.9\}$. In the

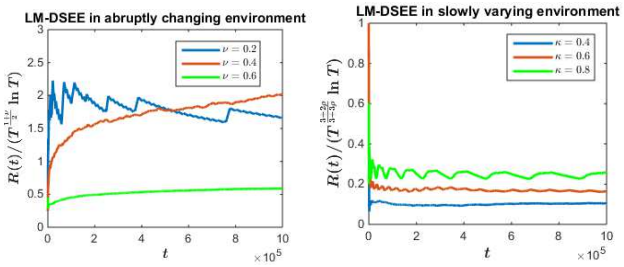

Fig. 1. The performance of the LM-DSEE algorithm in abruptly-changing and slowly-varying environments.

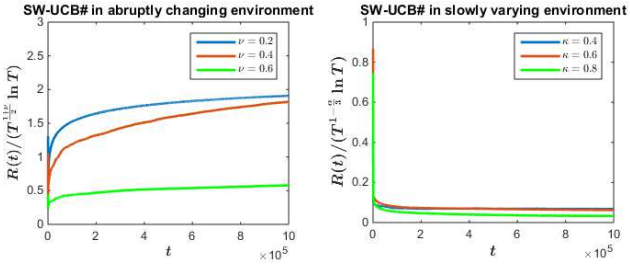

Fig. 2. The performance of the SW-UCB# algorithm in abruptly-changing and slowly-varying environments.

slowly-varying environment, the change in the mean reward at an arm is uniformly randomly sampled from the set $[-2T^{-\kappa}, 2T^{-\kappa}]$. For Algorithm 1, we select (a, b) equal to $(1, 0.25)$ and $(20, 1)$ for abruptly-changing and slowly-varying environments, respectively. For Algorithm 2, we select $\lambda = 12.3$ and $\lambda = 4.3$ for abruptly-changing and slowly-varying environments, respectively. The parameters ν and κ that describe characteristics of non-stationarity are varied to evaluate the performance of algorithms. Figs. 1 and 2 show that both SW-UCB# and LM-DSEE are effective in non-stationary environments.

In can be seen in Figs. 1 and 2 that for both algorithms in either of the environments, as expected, the ratio of the empirical regret to the order of the regret established in Sections IV and V is upper bounded by a constant. The regret for the SW-UCB# is relatively smoother than the regret for the LM-DSEE algorithm. The saw-tooth behavior of the regret for LM-DSEE is attributed to the fixed exploration-exploitation structure, wherein the regret certainly increases during the exploration epochs.

While both the algorithms incur the same order of regret, compared with LM-DSEE, SW-UCB# has a better leading constant. This illustrates the cost of constraining the algorithm to have a deterministic structure. On the other hand, this deterministic structure can be very useful, for example, in the context of planning trajectories for a mobile robot performing search using an MAB framework.

VII. Conclusions and Future Directions

We studied the stochastic MAB problem in two classes of non-stationary environments and designed two novel algorithms, LM-DSEE and SW-UCB# for these problems. We analyzed these algorithms for abruptly-changing and slowly-varying environments, and characterized their performance in terms of expected cumulative regret. In particular, we showed that these algorithms incur sublinear expected cumulative regret, i.e., the time average of the regret asymptotically converges to zero.

There are several possible avenues for future research. In this paper, we focused on a single decision-maker. Extensions of this work to multiple decision-makers is of significant interest. Implementation of these algorithms for robotic search and surveillance is an exciting direction to pursue. Finally, extension of the methodology developed in this paper to other classes on MAB problems such as the Markovian MAB problem and the restless MAB problem is also of interest.

References

[1] 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.

[2] V. Srivastava, P. Reverdy, and N. E. Leonard, "On optimal foraging and multi-armed bandits," in *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, 2013, pp. 494–499.

[3] ——, "Surveillance in an abruptly changing world via multiarmed bandits," in *IEEE Conference on Decision and Control*, 2014, pp. 692–697.

[4] 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 International Conference on Intelligent Robots and Systems*, Tokyo, Japan, November 2013, pp. 3368–3373.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[7] O. Besbes, Y. Gur, and A. Zeevi, "Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards," *arXiv preprint arXiv:1405.3316*, 2014.

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

[9] F. Liu, J. Lee, and N. Shroff, "A change-detection based framework for piecewise-stationary multi-armed bandit problem," *arXiv preprint arXiv:1711.03539*, 2017.

[10] R. Dimitrova, I. Gavran, R. Majumdar, V. S. Prabhu, and S. E. Z. Soudjani, "The robot routing problem for collecting aggregate stochastic rewards," *arXiv preprint arXiv:1704.05303*, 2017.

[11] V. Srivastava, F. Pasqualetti, and F. Bullo, "Stochastic surveillance strategies for spatial quickest detection," *The International Journal of Robotics Research*, vol. 32, no. 12, pp. 1438–1458, 2013.

[12] R. Agrawal, M. V. Hedge, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost," *IEEE Transactions on Automatic Control*, vol. 33, no. 10, pp. 899–906, 1988.

[13] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[14] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[15] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multi-player multi-armed bandits," *IEEE Transactions on Control of Network Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[16] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

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

[18] L. Wei and V. Srivastava, "On abruptly-changing and slowly-varying multiarmed bandit problems," *arXiv preprint arXiv:1802.08380*, 2018.

[19] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.