Learning in Restless Bandits Under Exogenous Global Markov Process

Tomer Gafni, Michal Yemini, *Member, IEEE*, and Kobi Cohen, *Senior Member, IEEE*

***Abstract*—We consider an extension to the restless multi-armed bandit (RMAB) problem with unknown arm dynamics, where an unknown exogenous global Markov process governs the rewards distribution of each arm. Under each global state, the rewards process of each arm evolves according to an unknown Markovian rule, which is non-identical among different arms. At each time, a player chooses an arm out of N arms to play, and receives a random reward from a finite set of reward states. The arms are restless, that is, their local state evolves regardless of the player's actions. Motivated by recent studies on related RMAB settings, the regret is defined as the reward loss with respect to a player that knows the dynamics of the problem, and plays at each time t the arm that maximizes the expected immediate value. The objective is to develop an arm-selection policy that minimizes the regret. To that end, we develop the Learning under Exogenous Markov Process (LEMP) algorithm. We analyze LEMP theoretically and establish a finite-sample bound on the regret. We show that LEMP achieves a logarithmic regret order with time. We further analyze LEMP numerically and present simulation results that support the theoretical findings and demonstrate that LEMP significantly outperforms alternative algorithms.**

***Index Terms*—Markov processes, restless multi-armed bandit, sequential learning, sequential decision making.**

I. Introduction

THE multi-armed bandit (MAB) problem is a popular model for sequential decision making with unknown information: A player chooses actions repeatedly among N different arms. After each action it receives a random reward having an unknown probability distribution that depends on the chosen arm. The objective is to maximize the expected total reward over a finite horizon of T periods. Restless multi-armed bandit (RMAB) problems are generalizations of the MAB problem. Differing from the classic MAB, where the states of passive arms remain frozen, in the RMAB setting, the state of each arm (active or passive) can change. In this paper we consider an extension to the RMAB problem, in which we assume that an exogenous (global) Markov process governs the distribution of the restless arms, and thus the reward depends on both the state of the global process, and the local state of the chosen (active) arm.

A. Applications

RMAB problems have attracted much attention for their wide application in diverse areas such as manufacturing systems [2], economic systems [3], biomedical engineering [4], wireless communication systems [5], [6] and communication network [7], [8]. A particularly relevant application captured by the extended RMAB model considered in this paper is the Dynamic Spectrum Access (DSA) paradigm [9], [10], [11], where primary users (licensed) occupy the spectrum occasionally, and a secondary user is allowed to transmit over a single channel when the channel is free. This behavior is commonly modeled by a Gilbert-Elliot model that comprises a Markov chain with two binary states. This model is captured as our exogenous (global) process, where global state 1 denotes a transmitting primary user and global state 0 denotes a vacant channel, i.e., an inactive primary user. The statistical model for the arms establishes the relationship between a physical channel and its finite-state Markov model for a packet transmission system. We adopt the view of previous studies, forming a finite-state Markov channel model to reflect the fading channel effect [12]. The received SNR values are partitioned into a finite number of states according to a criterion based on the average duration of each state. This model is very useful and enables one to avoid slow bit-level simulations and focus on the overall system design. The differences in time scales of fading and primary user switching can be manifested in our model by increasing the probability of the global state to remain in its state, compared to the probability of a local state to remain in a certain state.

Other potential applications of the model include important tasks in federated learning [6], recommendation systems [13], [14], and dynamic pricing [15]. Federated learning [6] is an emerging machine learning paradigm for training models across multiple edge devices holding local datasets, without explicitly exchanging the data. An important task in federated learning is to design user scheduling algorithms that determine which subset of users transmit at each round to save the communication resources, where the overall completion time of a task depends on the local processing time (including sensing, computation, transmission) of each selected user [6]. This application can be modeled by our new RMAB framework by modeling the random

Manuscript received 9 February 2022; revised 31 August 2022; accepted 15 November 2022. Date of publication 25 November 2022; date of current version 7 December 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yao Xie. This work was supported in part by Israel Science Foundation under Grant 2640/20 and in part by the Israeli National Cyber Bureau via the Cyber Security Research Center at Ben-Gurion University of the Negev. An early version of this paper that introduces the algorithm, and preliminary simulation results was accepted for presentation in the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022 [DOI: 10.1109/ICASSP43922.2022.9746078]. *(Corresponding author: Tomer Gafni.)*

Tomer Gafni and Kobi Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel (e-mail: gafnito@post.bgu.ac.il; yakovsec@bgu.ac.il).

Michal Yemini is with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel (e-mail: michal.yemini@biu.ac.il).

Digital Object Identifier 10.1109/TSP.2022.3224790

processing time of local users as Markov processes (e.g., due to Markovian channel fading, queue length of data samples, etc.). The global feature of the objective function in the federating learning task can be modeled by the global Markov process (e.g., the source state that generates the data samples, the location of the server that affects the communication channels, etc.).

In recommendation system tasks [13], [14], a well-known problem is the task of learning the "preference" or "rating" that a new user in the system would give to an item. The agent's goal is then to maximize the aggregated reward, which is associated with e.g., clicks, likes, shares, sells, etc. This problem is called the cold-start problem, and MAB-based formulations have been used to model it [13], [14]. Here, the arms represent items (or categories of items) and the reward distribution of each arm represents the user preference regarding this item. If the user interacts with the item, a positive feedback (i.e., a high reward) is given, if not, a negative feedback (e.g., a zero reward) is given. Furthermore, it is well known that often there are some global "trends" that affect the preferences of the user. Rather than assuming deterministic feature signals in MAB-based formulations [14], using the new RMAB framework considered in this paper, global trends can be modeled by an exogenous global Markov process (drawn from an unknown Markovian process and needs to be learned). Our new model allows different distributions when arms are active (selected) or passive (not selected), which is well suited to capture different preference distributions when items are presented to the user or not.

Finally, in dynamic pricing tasks [15], a key problem for businesses is to learn online the market demands to set prices for products or services dynamically. Consider the pricing decision for a manager at an online retailer. At the time of pricing, the manager is unlikely to have complete information about each product's demand curve. In these markets, a manager must consider an automated pricing policy to set real-time retail prices with incomplete demand information. For example, a consumer would buy a product only if its preference is higher than the price. Using our new RMAB model, the local arms represent the consumer's preferences which evolve as a Markov process. However, global economic changes (e.g., inflation) affect the consumer's preferences regarding various purchases. These global changes are captured by the global Markov process.

B. Performance Measure of RMAB Under Exogenous Global Markov Process

Computing the optimal policy for RMABs is P-SPACE hard even when the Markovian model is known [16], therefore, alternative tractable policies and objective functions have been proposed. Nevertheless, always playing the arm with the highest expected reward is optimal in the classic MAB under i.i.d. or rested Markovian rewards, up to an additional constant term [17]. Thus, a commonly used approach in classic RMAB (i.e., without exogenous process) with unknown dynamics settings to measure the algorithm performance in a tractable way defines the regret as the reward loss of the algorithm with respect to a genie that always plays the arm with the highest expected reward, also known as *weak* regret [18], [19], [20].

However, in our setting, due to the exogenous process, each global state is associated with different "best" arm (i.e., the arm with the highest expected reward). To accommodate the effect of the exogenous global Markovian state, we extend the definition of regret, and measure the performance of the algorithm by the reward loss of the algorithm with respect to a genie that plays in each time step the arm with the highest expected reward given the global state. Furthermore, since the next global state is unknown before choosing the arm for the next time step, we adopt a myopic performance measure, as considered also in [21], [22]. That is, the objective in this paper is to select the arm that has the highest immediate *expected* value at each time slot under unknown arm dynamics. The expected value, and thus also the arm selection, depend on both the transition probabilities of the global exogenous Markov process and the mean reward of the arms, which depends on the global state. Consequently, we define the regret as the reward loss of an algorithm with respect to a genie that knows the transition probabilities of the global process and the expected rewards of the local arms. Thus, we note that the regret is not defined with respect to the best arm on average (that would result in a weak regret), but with respect to a strategy tracking the best arm at each step, which is stronger. This notion of regret was also considered in Section 8 of [23] and in [24], for the non-stochastic bandit problem.

C. Main Results

Due to the restless nature of both active and passive arms, learning the Markovian reward statistics requires that arms will be played in a consecutive manner for a period of time (i.e., phase) [18], [19], [20]. Thus we divide the time horizon into two phases, an exploration phase and exploitation phase. The goal of the exploration phase is to identify the best arm for each global state before entering the exploitation phase.

Upper Confidence Bound (UCB)-based policies, that are used to identifying the best arm, require parameter tuning depending on the unobserved hardness of the task [25], [26], [27]. The hardness parameter is a characteristic of the hardness of the problem, in the sense that it determines the order of magnitude of the sample complexity required to find the best arm with a required probability. In the classic MAB formulation, the hardness of the task is characterized by $H_i = \frac{1}{(\mu^* - \mu^i)^2}$, where μ^*, μ^i are the means of the best arm and arm i, respectively. However, since the hardness parameter is unknown, existing algorithms use an upper bound on $\max_i H_i$ (e.g., [19]), which increases the order of magnitude of exploration phases, and consequently the regret. Considering the above, we summarize our main results and contributions.

1) An Extended Model for RMAB: RMAB problems have been investigated under various models of observation distributions in past and recent years. The extended model considered in this paper is capable of capturing more complex scenarios and requires an adaptation of the regret measure as discussed above. Handling this extension in the RMAB setting leads to different algorithm design and analysis as compared to existing methods.

2) Algorithm Development: We develop a novel algorithm, dubbed Learning under Exogenous Markov Process (LEMP), that estimates online the appropriate hardness parameter from past observations (Section III-A). Based on these past observations, the LEMP algorithm generates adaptive sizes of exploration phases, designed to explore each arm in each global state with the appropriate number of samples. Thus, LEMP avoids oversampling bad arms, and at the same time identifies the best arms with sufficient high probability. To ensure the consistency of the restless arms' mean estimation, LEMP performs regenerative sampling cycles (Section III-B). In the exploitation phases, LEMP dynamically chooses the best estimated arm, based on the evaluation of the global state (Section III-C). The rules that decide when to enter each phase are adaptive in the sense that they are updated dynamically and controlled by the current sample means and the estimated global transition probabilities in a closed-loop manner (Section III-D). Interestingly, the size of the exploitation phases is deterministic and the size of the exploration phases is random.

3) Performance Analysis: We provide a rigorous theoretical analysis of LEMP algorithm. Specifically, we establish a finite sample upper bound on the expected regret, and show that its order is logarithmic with time. We also characterize the appropriate hardness parameter for our model (the $\overline{D}_i$ parameter defined in (3)), and we demonstrate that estimating the hardness parameter indeed results in a scaled regret proportional to the hardness of the problem. The result in Theorem 1 also clarifies the impact of different system parameters (rewards, mean hitting times of the states, eigenvalues of the transition probability matrices, etc.) on the regret. We provide numerical simulations that support the theoretical results presented in this paper.

D. Related Work

The extended RMAB model considered here is a generalization of the classic MAB problem [17], [28], [29], [30], [31]. RMAB problems have been extensively studied under both the non-Bayesian [18], [19], [20], [21], [32], [33], [34], [35], [36], and Bayesian [37], [38], [39], [40], [11], [41], [42], [43], [44], [45] settings. Under the non-Bayesian setting, special cases of Markovian dynamics have been studied in [18], [21]. There are a number of studies that focused on special classes of RMABs. In particular, the optimality of the myopic policy was shown under positively correlated two-state Markovian arms [41], [42], [43], [46] under the model where a player receives a unit reward for each arm that was observed in a good state. In [44], [47], the indexability of a special classes of RMAB has been established. In [33], the traditional restless bandit is extended by relaxing the restriction of a risk-neutral target function, and a general risk measure is introduced to construct a performance criterion for each arm. Our work is also related to models of partially observed Markov decision process (POMDP) [48], with the goal of balancing between increasing the immediate reward and the benefits of improving the learning accuracy of the unknown states. In [7], the offloading policy design in a large-scale asynchronous MEC system with random task arrivals, distinct workloads, and diverse deadlines is formulated as an RMAB problem, and the authors in [49] considered a tracking problem with independent objects and used an approximated Gittins index approach for finding policies.

The setting in this paper is also related to the non-stationary bandit problems, where distributions of rewards may change in time [24], [50], [51], [52], [53]. However, the distribution that governs the non-stationary models in these studies differs from our settings, and leads to a different problem structure. Finally, [54], [55] and recently [56] considered the setting of global Markov process that governs the reward distribution. However, they addressed the linear/affine model, and not the RMAB formulation that is explored in this paper. This new setting leads to fundamentally different algorithm design and regret analysis, mainly due to the restless nature of both active and passive arms in our model, that requires that arms will be played in a consecutive manner for a period of time.

II. System Model and Problem Formulation

We consider a set of N arms, indexed by $\{1,\ldots,N\} \triangleq \mathcal{N}$, and a global system state process $\{s_t\}_{t=1,2,\ldots}$, which is governed by a finite space, irreducible, and aperiodic discrete time Markov chain $\mathcal{S}$ with unknown transition matrix P_S. We denote the transition probability between states $\tilde{s}$ and $\check{s}$ in $\mathcal{S}$ by $p_{\tilde{s}\check{s}}$, and we denote by π_s the stationary distribution of states $s \in \mathcal{S}$. For each global state $s \in \mathcal{S}$, the i^{th} arm is modeled as a finite space, irreducible, and aperiodic discrete time Markov chain $\mathcal{X}^i_s$ with unknown transition matrix $P_{\mathcal{X}^i_s}$. We denote the transition probability between states x and y in $\mathcal{X}^i_s$ by $p^{s,i}_{xy}$. We assume that $\mathcal{X}^i_{\tilde{s}} \bigcap \mathcal{X}^i_{\check{s}} = \emptyset$ for all $i, \tilde{s}, \check{s}$ (i.e., we can recover the global state in each time slot[1]). We also define the stationary distribution of state x in arm i at global state s to be $\pi^i_s(x)$. An illustration for the model with $|\mathcal{S}| = 2, N = 2, |\mathcal{X}^i_s| = 2, \forall s, i$ is given in Fig. 1.

At each time t, the player chooses one arm to play. When played, each arm offers a certain positive reward that defines the current state of the arm, $x^i_{s_t}$. The player receives the reward of the chosen arm, and infers the current global state s_t. Then, the global state transitions to a new state, which is unknown to the player before choosing the next arm to play. We assume that the arms are mutually independent and restless, i.e., the local states of the arms continue to evolve regardless of the player's actions according to the unknown Markovian rule $P_{\mathcal{X}^i_s}$. The unknown stationary reward mean of arm i at global state s, μ^i_s, is given by:

$$\mu^i_s = \sum_{x \in \mathcal{X}^i_s} x \pi^i_s(x).$$

We further define the expected value of arm i in global state s to be

$$V^i_s \triangleq \sum_{\check{s} \in \mathcal{S}} p_{s\check{s}} \mu^i_{\check{s}}. \qquad (1)$$

Let σ be a permutation of $\{1,\ldots,N\}$ such that

$$V^{\sigma(1)}_s \geq V^{\sigma(2)}_s \geq \cdots \geq V^{\sigma(N)}_s.$$

Let $V^i_{s_t}(t)$ denote the value of arm i at time t, let i^*_t be the arm with the highest expected value at time t, i.e., $i^*_t \triangleq \arg\max_i V^i_{s_t}(t)$,

[1] If this assumption does not hold, the player may recover the global state using techniques for the specific application task, e.g., spectrum sensing methods in the DSA paradigm [57].

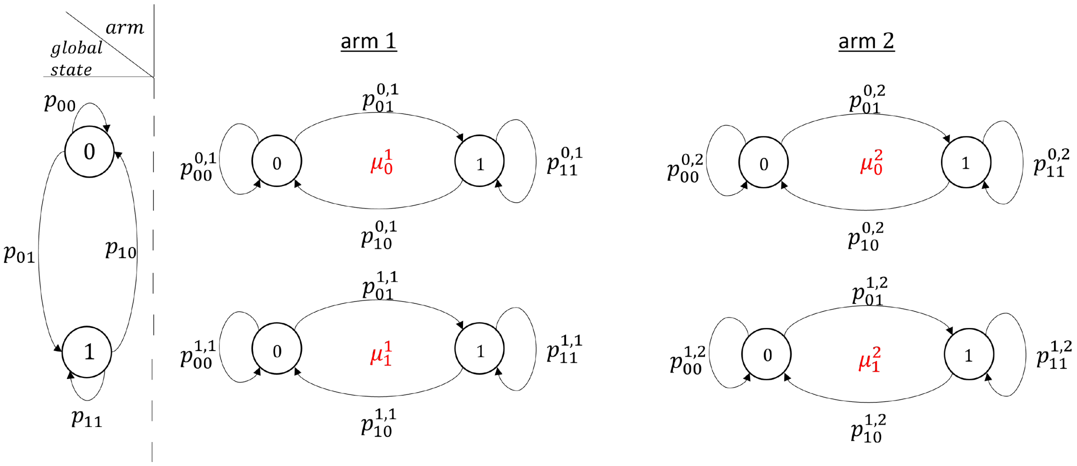

Fig. 1. An illustration of the system model with $|\mathcal{S}| = 2, N = 2, |\mathcal{X}^i_s| = 2, \forall s, i$.

and let $\phi(t) \in \{1, 2, \ldots, N\}$ be a selection rule indicating which arm is chosen to be played at time t, which is a mapping from the observed history of the process to $\mathcal{N}$. Denote

$$V_{\mathrm{e}}(n) = \{i : V^i_{s_n}(n) < V^{\sigma(1)}_{s_n}(n)\}.$$

The expected regret of policy ϕ is defined as:

$$\mathbb{E}_\phi[r(t)] = \mathbb{E}_\phi\left[\sum_{n=1}^{t} \sum_{i \in V_{\mathrm{e}}(n)} \left(x^{i^*_n}_{s_n}(n) - x^i_{s_n}(n)\right) \mathbb{1}_{\{\phi(n)=i\}}\right], \quad (2)$$

where hereafter $\mathbb{1}_{\{A\}}$ denotes an indicator of an event A. The objective is to find a policy that minimizes the growth rate of the regret with time (this notion of regret is similar to the "regret against arbitrary strategies" introduced in Section 8 of [23] and in [24] for the non-stochastic bandit problem). We note that, in this paper, the regret is not defined with respect to the best arm on average (e.g., as in RMAB models in [18], [19], [20]), but with respect to the best arm at each step according to the instantaneous global state, which is a stronger regret. A list of notations used in the paper is summarized in Table 1.

TABLE I
NOTATIONS

Notation	Description
N	number of arms
$\mathcal{S}$	global process
s_t	global state at time t
$\mathcal{X}^i_s$	local process of arm i in state s
$x^i_{s_t}$	reward (local state) of arm i at global state s_t
π_s	stationary distribution of state $s \in \mathcal{S}$
$\pi^i_s(x)$	stationary distribution of local state x in arm i at global state s
$p_{\tilde{s}\check{s}}$	transition probability between states $\tilde{s}$ and $\check{s}$ in $\mathcal{S}$
μ^i_s	stationary reward mean of arm i in state s
V^i_s	expected value of arm i in state s (Eq. (1))
$\hat{V}^i_s(t)$	estimated expected value of arm i in state s at time t (Eq. (7))
$\overline{D}^i_s$	exploration rate of arm i in state s (Eq. (3))
$\hat{D}^i_s(t)$	estimated exploration rate of arm i in state s at time t (Eq. (6))
$T^i_s(t)$	the number of samples from arm i in global state s in sub-block SB2 up to time t
$t^i_s(n)$	time index of the nth play on arm i in global state s in sub-block SB2
$N_s(t)$	number of occurrences of the state s until time t
$N_{s\tilde{s}}(t)$	number of transitions from s to $\tilde{s}$ up to time t.
I_L, I_G	local and global (respectively) minimal rate function (Eq. (8), (9))
L	exploration coefficient (Eq. (12))

III. The Learning Under Exogenous Markov Process (LEMP) Algorithm

The LEMP algorithm divides the time horizon into two types of phases, namely exploration and exploitation. In order to ensure sufficient small regret in exploitation phases (i.e., to reduce the probability for choosing sub-optimal arms in exploitation), our strategy estimates the required exploration rate of each arm, and updates the arm selection dynamically with time, controlled by the random sample means and transition probability estimates in a closed loop manner.

A. Design Principles of LEMP

For sufficient small regret during exploitation phases, we should take a sufficiently large number of samples in the exploration phases. From (1) we observe that we should estimate accurately two terms: the mean reward of each arm i in each global state s, μ^i_s, and the transition probabilities of the global Markov chain $\mathcal{S}$, $p_{\tilde{s}\check{s}}$.

In the analysis, we show that in each global state s, we must explore a suboptimal arm i with a *local exploration rate* of at least $\overline{D}^i_s \log(t)$ times for being able to distinguish it from $i^*_s \triangleq \arg\max_i V^i_s$ (i.e., the arm that maximizes the expected value in state s) with a sufficiently high accuracy, where

$$\overline{D}^i_s \triangleq \frac{4L}{(V^*_s - V^i_s)^2}, \quad (3)$$

where $V^*_s \triangleq \max_i V^i_s$, and L is the exploration coefficient that depends on the system parameters, defined in (12). The $\overline{D}^i_s$ parameter is a type of hardness parameter [25], appropriate for the setting considered in this paper, in the sense that it determines

Algorithm 1: LEMP Algorithm.

```
initialize: $t = 0, N_s = 0, n_I = 0, n_O^i = 1, T_s^i = 0, r_s^i = 0, \forall i = 1 \dots N$
for i = 1:N do
    play arm $i$; observe global state $s$ (denote the previous global state by $\tilde{s}$) and local state as $x$ and set $\gamma^i(n_O^i) = x$
    $t := t + 1; T_s^i := T_s^i + 1; n_O^i := n_O^i + 1; r_s^i := r_s^i + r_{x_s}$
    $N_s = N_s + 1;$
while (1) do
    for $i = 1 : N$ do
        set $\hat{\mu}_s^i(t), \hat{p}_{\tilde{s}s}(t), \widehat{D}_s^i(t)$ according to (4), (5), (6), respectively;
    while condition (10) holds for some arm $i$ (or condition (11) holds) do
        play arm $i$ (or arm $i_M$); observe global state $s$ and local state as $x$;
        while $x \neq \gamma^i(n_O^i - 1)$ (SB1) do
        $t := t + 1 \; N_s = N_s + 1$
        play arm $i$; observe global state $s$ and local state as $x$
      $t := t + 1; T_s^i := T_s^i + 1; r_s^i := r_O^i + r_{x_s}; N_s = N_s + 1;$
    for $n = 1 : 4^{n_O^i - 1}$ (SB2) do
        play arm $i$; observe global state $s$ and local state as $x$
        $t = t + 1; T_s^i = T_s^i + 1; r_s^i = r_s^i + r_{x_s}; N_s = N_s + 1;$
    $n_O^i := n_O^i + 1$; set $\hat{\mu}_s^i(t), \hat{p}_{\tilde{s}s}(t)$ according to (4), (5), (6), respectively; $\gamma^i(n_O^i) = x$
set $\hat{V}_s^i(t) \; \forall s$ according to (7)
set $i_s^* = \arg\max_i \hat{V}_s^i(t) \; \forall s$
for $n = 1 : 2 \cdot 4^{n_I - 1}$ do
    play arm $i_s^*$; observe new global state $s$ and local state as $x$
    $t := t + 1; N_s = N_s + 1$
  $n_I := n_I + 1;$
```

the order of magnitude of the sample size required to find the best arm in each global state with a required probability.

We point out that in order to derive $\overline{D}_s^i$, we should know the system parameters $\{p_{s\check{s}}\}, \{\mu_{\check{s}}^i\}$. Since the reward means and the transition probabilities are unknown, we estimate $\overline{D}_s^i$ by replacing $\mu_s^i, p_{s\check{s}}$ by their estimators:

$$\hat{\mu}_s^i(t) = \frac{1}{T_s^i(t)} \sum_{n=1}^{T_s^i(t)} x_s^i(t_s^i(n)), \tag{4}$$

$$\hat{p}_{s\check{s}}(t) = \frac{N_{s\check{s}}(t)}{N_s(t)}. \tag{5}$$

where $t_s^i(n)$ is the time index of the n^{th} play on arm i in global state s in sub-block SB2 only (SB2 is detailed in Section III-B), $T_s^i(t)$ is the number of samples from arm i in global state s in sub-block SB2 up to time t, $N_s(t)$ is the number of occurrences of the state s until time t, and $N_{s\check{s}}(t)$ is the number of transitions from s to $\check{s}$ up to time t. We also define: $\Delta_s^i \triangleq (V_s^* - V_s^i)^2$, $\Delta_s \triangleq \min_i \Delta_s^i$, and we define $0 < \Delta \leq \min_s \Delta_s$ to be a known lower bound on $\min_s \Delta_s$.

Denote the estimator of $\overline{D}_s^i$ by:

$$\hat{D}_s^i(t) \triangleq \frac{4L}{\max\left\{\Delta, (\hat{V}_s^*(t) - \hat{V}_s^i(t))^2 - \epsilon\right\}}, \tag{6}$$

where:

$$\hat{V}_s^i(t) \triangleq \sum_{\check{s} \in \mathcal{S}} \hat{p}_{s\check{s}}(t) \hat{\mu}_{\check{s}}^i(t), \tag{7}$$

$\hat{V}_s^*(t) \triangleq \max_i \hat{V}_s^i(t)$, $\epsilon > 0$ is a fixed tuning parameter, and L is defined in (12).

Using $\{\hat{D}_s^i(t)\}$, which are updated dynamically over time and controlled by the corresponding estimators, we can design an adaptive arm selection for sampling arm i at state s that will converge to its exploration rate required for efficient learning, as time increases. That is, the LEMP algorithm generates adaptive sizes of exploration phases, designed to explore each arm in each global state with the appropriate number of samples. Thus, LEMP avoids oversampling bad arms, and at the same time identifies the best arms with sufficiently high probability. Whether we succeed to obtain a logarithmic regret order depends on how fast $\hat{D}_s^i(t)$ converges to a value that is no smaller than $\overline{D}_s^i$ (so that we take at least $\overline{D}_s^i$ samples from bad arms in most of the times).

B. Description of the Exploration Phases

Due to the restless nature of both active and passive arms, learning the Markovian reward statistics requires that arms will be played in a consecutive manner for a period of time (i.e., phase). Therefore, the exploration phases are divided into sub-blocks SB1 and SB2. Consider time t (and we remove the time index t for convenience). We define $n_O^i(t)$ as the number of exploration phases in which arm i was played up to time t. Let $\gamma^i(n_O^i - 1)$ be the last reward state observed at the $(n_O^i - 1)^{th}$ exploration phase for arm i. As illustrated in Fig. 2, once the player starts the $(n_O^i)^{th}$ exploration phase, it first plays a random period of time, also known as a random hitting time, until observing state $\gamma^i(n_O^i - 1)$. This random period of time is referred to as SB1. Then, the player plays arm i until it observes $4^{n_O^i}$ samples. This period of time is referred to as SB2. The player stores the $(4^{n_O^i})^{th}$ state $\gamma^i(n_O^i)$ observed at the current $(n_O^i)^{th}$ exploration phase, and so on. We define the set of time indices during SB2 sub-blocks by $\mathcal{V}_i$. This procedure ensures that each interval in $\mathcal{V}_i$ starts from the last state that was observed in the previous interval. Therefore, cascading these intervals forms a sample path which is equivalent to a sample path generated by continuously sampling the Markov chain.

C. Description of the Exploitation Phases

Let $n_I(t)$ be the number of exploitation phases up to time t. The player plays the exploitation phase for a deterministic period of time with length $2 \cdot 4^{n_I(t)-1}$ according to the following rule: at each time slot the player computes the expected value, $\hat{V}_s^i(t)$, of each arm given the observed global state when entering the $(n_I)^{th}$ exploitation phase, and plays the arm that maximizes the expected value.

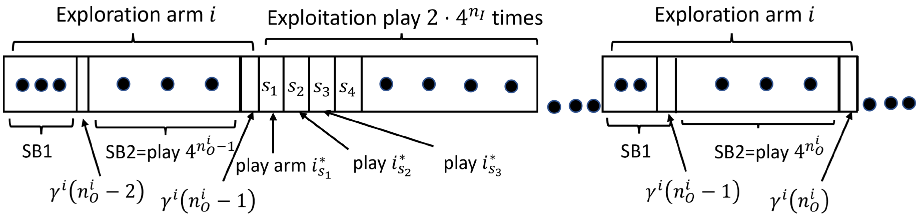

Fig. 2. An illustration of the exploration and exploitation phases of LEMP Algorithm.

D. Phase Selection Conditions

At the beginning of each phase, the player needs to decide whether to enter an exploration phase for one of the N arms, or whether to enter an exploitation phase. We recall that the purpose of the exploration phases is to estimate both the expected rewards of the arms, and the transition probabilities of the global process. We therefore define:

$$I_L \triangleq \frac{\bar{\lambda}_{\min}}{3072\left((x_{\max}+2)^2 \cdot |\mathcal{X}_{\max}| \cdot \hat{\pi}_{\max} \cdot |\mathcal{S}| \cdot (V^*_{\max}+2)\right)^2}, \tag{8}$$

$$I_G \triangleq \frac{1}{128\left((x_{\max}+2) \cdot |\mathcal{S}| \cdot (V^*_{\max}+2)\right)^2}, \tag{9}$$

which we denote as the local and global (respectively) minimal rate functions. The decision to explore or exploit will be made due to the next two conditions: first, if there exists an arm i and a global state s such that the following condition holds:

$$T^i_s(t) \leq \max\left\{\hat{D}^i_s(t), \frac{2}{\epsilon^2 \cdot I_L}\right\} \cdot \log t, \tag{10}$$

then the player enters an exploration phase for arm i. Second, if there exists a global state $s \in \mathcal{S}$ where

$$N_s(t) \leq \frac{2}{\epsilon^2 \cdot I_G} \cdot \log t, \tag{11}$$

then the player enters an exploration phase for arm i_M where $i_M \triangleq \arg\min_i\{\min_s \hat{D}^i_s(t)\}$. Otherwise, the player enters an exploitation phase. This selection rule of the LEMP algorithm is thus designed based on the following insights. First, the algorithm must take at least $\overline{D}^i_s(t) \cdot \log(t)$ samples from each sub-optimal arm for computing a sufficiently accurate estimate of the expected value $\hat{V}^i_s(t)$. Since $\overline{D}^i_s(t)$ depends on the expected values which are unknown, the algorithm replaces the unknown value $\overline{D}^i_s(t)$ by $\hat{D}^i_s(t)$, which overestimates $\overline{D}^i_s(t)$ to obtain the desired property. Second, since $\hat{D}^i_s(t)$ is a random variable, we need to make sure that the desired property holds with a sufficiently high probability. The parameters I_L, I_G in (10), (11) are used to guarantee the desired property.

IV. Regret Analysis

In the following theorem we establish a finite-sample bound on the expected regret as the function of time, resulting in a logarithmic regret order.

Theorem 1: Assume that LEMP algorithm is implemented and the assumptions on the system model described in Section II hold, and an upper bound on Δ in known. Let λ^i_s be the second largest eigenvalue of $P_{\mathcal{X}^i_s}$, and let $M^{s,i}_{x,y}$ be the mean hitting time of state y starting at initial state x for arm i in global state s. Define $x_{\max} \triangleq \max_{s\in\mathcal{S}, i\in\mathcal{N}} x^i_s$, $|\mathcal{X}_{\max}| \triangleq \max_{s\in\mathcal{S}, i\in\mathcal{N}} |\mathcal{X}^i_s|$, $\pi_{\min} \triangleq \min_{s\in\mathcal{S}, i\in\mathcal{N}, x\in\mathcal{X}^i_s} \pi^i_s(x)$, $\hat{\pi}_{\max} \triangleq \max_{s\in\mathcal{S}, i\in\mathcal{N}, x\in\mathcal{X}^i_s}\{\pi^i_s(x), 1-\pi^i_s(x)\}$, $V^*_{\max} =\triangleq \max_{s\in\mathcal{S}} V^*_s$, $\lambda_{\max} \triangleq \max_{s\in\mathcal{S}, i\in\mathcal{N}} \quad \lambda^i_s$, $\bar{\lambda}_{\min} \triangleq 1-\lambda_{\max}$, $\overline{\lambda^i_s} \triangleq 1-\lambda^i_s$, $M^i_{s,\max} \triangleq \max_{x,y\in\mathcal{X}^i_s, x\neq y} M^{s,i}_{x,y}$, $M^i_{\max} \triangleq \max_s M^i_{s,\max}$,

$$L \geq \frac{1}{16(V^*_{\max}+2)^2} \cdot \max\left\{\frac{1}{I_L}, \frac{1}{I_G}\right\}. \tag{12}$$

Then, the regret at time t is upper bounded by:

$$\begin{aligned} \mathbb{E}_\phi[r(t)] \leq x_{\max} \cdot \Bigg[& \sum_{i=1}^{N} \Bigg(\frac{1}{3}\left[4\left(3A_i \cdot \log(t)+1\right)-1\right] \\ & + M^i_{\max} \cdot \log_4\left(3A_i \log(t)+1\right)\Bigg) \\ & + 6N|\mathcal{S}|\left(\frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} + 2|\mathcal{S}|\right) \max_s \pi_s \\ & \cdot \left\lceil \log_4\left(\frac{3}{2}t+1\right)\right\rceil \Bigg] + \beta, \end{aligned} \tag{13}$$

where

$$A_i \triangleq \begin{cases} \max\{\frac{2}{\epsilon^2 I_L}, \frac{2}{\epsilon^2 I_G}, \max_s \overline{D}^i_{s,\max}\}, & \text{if } \forall s: i \in \mathcal{K}_s \\ \max\{\frac{2}{\epsilon^2 I_L}, \frac{2}{\epsilon^2 I_G}, 4L/\Delta\}, & \text{if } \exists s: i \notin \mathcal{K}_s \end{cases}, \tag{14}$$

$\overline{D}^i_{s,\max} \triangleq \frac{4L}{(V^*_s - V^i_s)^2 - 2\epsilon}$, $\mathcal{K}_s$ is defined as the set of all indices $i \in \{2,\ldots,N\}$ in global state s that satisfy:

$$\left(V^*_s - V^{\sigma(i)}_s\right)^2 - 2\epsilon > \Delta_s,$$

and

$$\beta \leq 8|\mathcal{S}|N\frac{|\mathcal{X}_{\max}|}{\pi_{\min}} \cdot \left[\zeta\left(2+\delta\right) + \frac{1}{1+\delta}\zeta\left(1+\delta\right)\right],$$

for some arbitrarily small $\delta > 0$, where $\zeta(\cdot)$ is the Reimann zeta function.

Theorem 1 shows that the regret under LEMP has a logarithmic order with time. The scaling of the regret with the mean hitting time $M^i_{\max}$ under LEMP (which arises due to the construction of SB1 in the exploration phases) is of order $O(\sum_i M^i_{\max} \log\log t)$. We point out that the scaling with $M^i_{\max}$ under LEMP is significantly better than the scaling under other algorithms that use regenerative cycles. For example, in [18], the RCA algorithm performs random regenerative cycles until catching predefined states in *each phase*, thus the scaling with the mean hitting time (which scales at least polynomially with the state space) is $O\big(\sum_i M^i_{\max}\log t\big)$. The scaling with N and Δ under LEMP is of order $O\big(\big(\frac{1}{\sqrt{\Delta}}+N-2\big)\log t\big)$ since every bad arm is sampled according to its unique exploration rate which is estimated by the adaptive sequencing rules. We point out that the scaling with N and Δ under LEMP is significantly better than the scaling under other algorithms that divides the time horizon into exploration and exploitation phases. For example, in [19], the scaling under the DSEE algorithm is of order $O\big(\big(\frac{1}{\sqrt{\Delta}}+\frac{N-2}{\Delta}\big)\log t\big)$ since all bad arms are explored according to the worst exploration rate. We also point out that the scaling of the regret under LEMP with the size of the global space $|\mathcal{S}|$ and local space $|\mathcal{X}_{\max}|$ is similar to the scaling of these parameters in classic works in the literature (e.g., [58] for $|\mathcal{S}|$ and [18], [31] for $|\mathcal{X}^i_s|$). Moreover, if the exogenous Markov process has just a single state, we obtain the typical regret bound for restless Markovian MAB (e.g., as in [20]), as expected. However, we note that if the local arms evolve as an i.i.d process, we do not obtain the bound of stochastic K-armed bandit. This is due to the fact that under the restless Markovian model we are required to design a much more complicated algorithm compared to algorithms in the stochastic K-armed bandit (e.g., UCB1) in order to obtain the logarithmic regret.

We finally note that the model considered in this paper can alternatively be adjusted such that we allow transitions of local arms at each time unit, and allow transitions of the global arm at each K time units, as done, for example, in queuing system applications, when some servers may be faster than others. We can apply the LEMP algorithm in this model as well by adjusting the estimation of the expected reward based on the same global state for K time units. Note that the assumptions in the analysis still hold, and we can still use Lezaud's result [59] in Lemma 5 (which bounds the probability of a large deviation from the stationary distribution) since it is applied to the local arms. Also, (26) in Lemma 1 (which bounds the estimation error of the global transition probabilities) holds as well. This adjustment would result in an additional K constant term in the first term of the regret, as a consequences of (11). This additional constant does not violate the logarithmic order of the regret.

Before proceeding to prove Theorem 1, we define the following auxiliary notation.

Definition 1: Let T_1 be the smallest integer, such that for all $t\geq T_1$ the following holds: $\overline{D}^i_s \leq \widehat{D}^i_s(t)$ for all $i\in\mathcal{N}, s\in\mathcal{S}$, and also $\widehat{D}^i_s(t)\leq\overline{D}^i_{s,\max}$ for all $i\in\mathcal{K}_s, s\in\mathcal{S}$.

The term T_1 captures the random time by which the exploration rates for all arms are sufficiently close to the desired exploration rates needed for achieving the desired logarithmic regret bound (as shown later).

Proof of Theorem 1: The layout of the proof is as follows, first we show in Lemma 1 that the expectation of the random time T_1 is bounded independently of t. Then, based on Lemma 2 we show in Lemmas 3 and 4, that a logarithmic regret is obtained for all $t>T_1$, which yields the desired expected regret.

In the next Lemma we show that the expected value of T_1 is bounded under the LEMP algorithm.

Lemma 1: Assume that the LEMP algorithm is implemented as described in Section III. Then, $\mathbb{E}[T_1]<\infty$ is bounded independent of t.

The proof of the Lemma 1 is given in Appendix A.

The second step of the proof is to show that a logarithmic regret is obtained for all $t>T_1$, which yields the desired expected regret.

Lemma 2: Let $\tilde{T}^i(t)\triangleq\sum_{n=1}^{t}\mathbb{1}_{\{\phi(n)=i\neq i^*_n\}}$ denote the number of times arm i was played when it was not the best arm during the t first rounds. Then the expected regret is upper bounded by:

$$\mathbb{E}_\phi[r(t)]\leq x_{\max}\mathbb{E}_\phi[T_1]+x_{\max}\sum_{i=1}^{N}\mathbb{E}_\phi[\tilde{T}^i(t)]. \qquad (15)$$

We present the proof of this lemma in Appendix B.

From (15) we observe that it is sufficient to upper-bound the expected number of times an arm i is played when this arm is sub-optimal. We will bound (15) for the exploration and exploitation phases separately. Specifically, let $T^i_O(t)$, and $T^i_I(t)$, denote the time spent on sub-optimal arm i in exploration and exploitation phases, respectively, by time t. Thus,

$$\tilde{T}^i(t)=T^i_O(t)+T^i_I(t).$$

The following two lemmas show that both $\mathbb{E}[T^i_O(t)]$ and $\mathbb{E}[T^i_I(t)]$ have a logarithmic order with time.

Lemma 3: The time spent by time t in exploration phases for sub-optimal arm i is bounded by:

$$\mathbb{E}[T^i_O(t)]\leq\sum_{i=1}^{N}\Big[\tfrac{1}{3}[4(3A_i\cdot\log(t)+1)-1] + M^i_{\max}\cdot\log_4(3A_i\log(t)+1)\Big].$$

Lemma 4: The time spent by time t in exploitation phases for sub-optimal arm i is bounded by:

$$\mathbb{E}\left[T^i_I(t)\right]\leq 6|\mathcal{S}|\cdot\left(\frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}}+2|\mathcal{S}|\right)\cdot\max_s\pi_s\cdot\left\lceil\log_4\left(\frac{3}{2}t+1\right)\right\rceil.$$

The proofs of Lemmas 3 and 4 are given in Appendix C and D, respectively.

Finally, we show at Appendix E, that combining the above four lemmas concludes the proof of Theorem 1. ■

V. Simulation Results

In this section we evaluate the regret of the LEMP algorithm numerically in four different scenarios.

We compare the LEMP algorithm to an extended version of the DSEE algorithm [19] which is an efficient and widely used algorithm in the RMAB settings, and to a strategy that chooses in the exploitation phases the best arm on average (i.e., competing against weak regret as in [18], [20]). The DSEE algorithm uses deterministic sequencing of exploration and exploitation phases, however, it does not estimate the hardness parameter, and explores each arm $\frac{4L}{\Delta}\cdot\log(t)$ times, which results in oversampling bad arms to achieve the desired logarithmic regret. We simulate and compare the regret of these three algorithms averaged over 1000 Monte-Carlo experiments, under four scenarios, denoted

TABLE II
A DESCRIPTION OF THE EXPERIMENT PARAMETERS

	N	$\lvert\mathcal{S}\rvert$	Δ
S1	3	2	0.4
S2	6	2	0.4
S3	3	3	0.4
S4	3	2	0.2

We set $|\mathcal{X}_s^i| = 2$ in all scenarios. The values of all transition probabilities and state rewards under each scenario are described in the text.

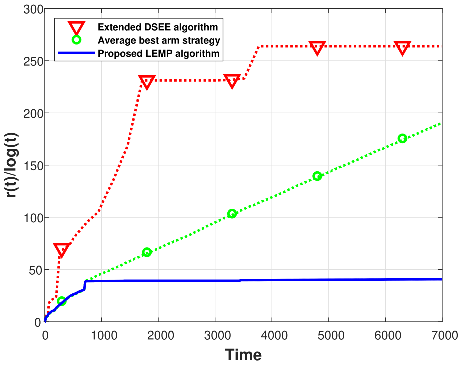

Fig. 3. S1: Performance comparison of the regret (normalized by log t).

by S1, S2, S3, S4. In Table II we summarize the number of arms, the number of global states, and the difference between the highest and the second highest values for each scenario. In all scenarios we set $L = 6, I_L = 0.1, I_G = 0.1$.

1) RMAB With Exogenous Process (S1): In Fig. 3 we simulated S1 scenario. Here, the global state models the presence of the primary user that uses the entire bandwidth by a Gilbert-Eliot model [60] that comprises a Markov chain with two binary states, where global state $s = 1$ denotes a transmitting primary user and $s = 0$ denotes a vacant channel, i.e., inactive primary user. To limit the interference to the primary user, a secondary user may choose to transmit over one of three possible channels (i.e., $N = 3$), where the channels are modeled by a Finite-State Markovian Channel (FSMC), which is a tractable model widely used to capture the time-varying behavior of a radio communication channel (e.g., a Rayleigh fading channels [12]). The transition probabilities of the global chain are $p_{00} = 0.4, p_{10} = 0.75$. In global state 1, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.5, 0.6, 0.7]$, $p_{21} = [0.5, 0.4, 0.3]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [4, 5.8, 1]$, $r_2 = [6, 8.2, 2]$. In global state 2, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.55, 0.65, 0.75]$, $p_{21} = [0.45, 0.35, 0.25]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [10, 9, 2.5]$, $r_2 = [14, 11, 3]$. It can be seen that LEMP significantly outperforms the extended DSEE algorithm. Fig. 3 also shows the superior of LEMP against a strategy that chooses the best arm on average, demonstrating the gain in tracking the best arm at each step according to the global process evolution.

2) Increasing the Number of Arms (S2): Next, we are interested in examine the regret in a larger system, i.e., we increased the number of arms to 6 (S2). The results are depicted in Fig. 4. The transition probabilities of the global process is as in the previous scenario. In global state 1, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.5, 0.6, 0.7, 0.7, 0.6, 0.5]$, $p_{21} = [0.5, 0.4, 0.3, 0.3, 0.4, 0.5]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [4, 5.8, 1, 1.1, 0.6, 1.2]$, $r_2 = [6, 8.2, 2, 1.9, 0.9, 2.2]$. In global state 2, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.55, 0.65, 0.75, 0.75, 0.65, 0.55]$, $p_{21} = [0.45, 0.35, 0.25, 0.25, 0.35, 0.45]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [10, 9, 2.5, 3, 2.56, 2.7]$, $r_2 = [14, 11, 3, 2.8, 3.1, 3.3]$. Increasing the number of arms is expected to decrease the performance under the extended DSEE algorithm, since more arms are sampled by the worst exploration rate. Indeed, it can be seen in Fig. 4, that the gap in the regret between LEMP and the extended DSEE algorithm is increased compared to the previous simulation. This is due to the fact that in the proposed LEMP algorithm, each arm is played according to its unique exploration rate (as a result of the online estimation of the hardness parameter), thus adding "bad" arms (i.e., arms with high exploration rate) does not significantly affect the LEMP performances. LEMP outperforms the strategy that chooses the best arm on average also in this scenario.

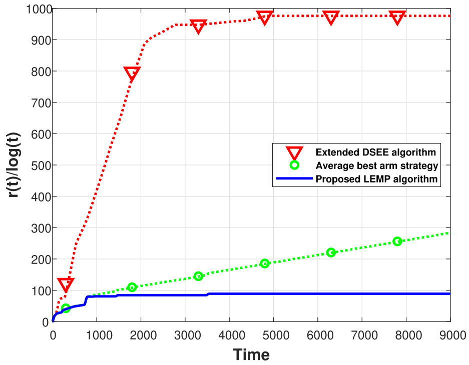

Fig. 4. S2: Performance comparison of the regret (normalized by log t).

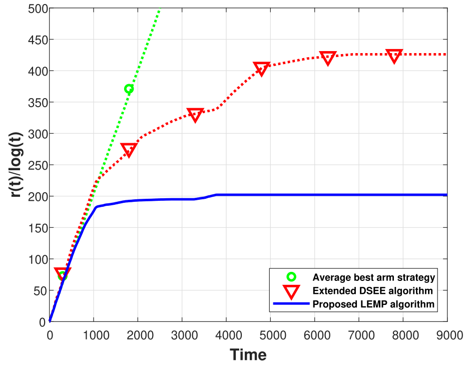

Fig. 5. S3: Performance comparison of the regret (normalized by log t).

3) Increasing the Number of Global States (S3): In Fig. 5 we increased the number of global states to 3 (S3). The transition probabilities of the global chain are $p_{00} = 0.85$, $p_{01} = 0.1$, $p_{02} = 0.05$, $p_{10} = 0.08$, $p_{11} = 0.85$, $p_{12} = 0.07$, $p_{20} = 0.06$, $p_{21} = 0.09$, $p_{22} = 0.85$. In global state 1, the local transition

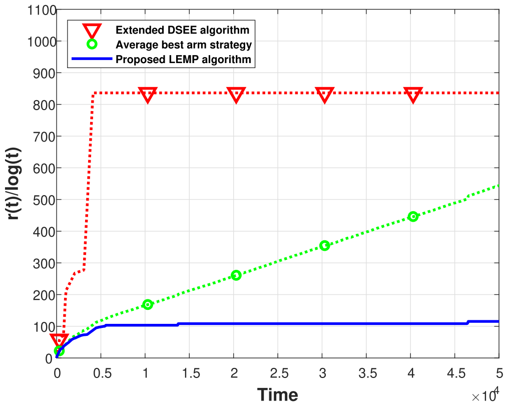

Fig. 6. S4: Performance comparison of the regret (normalized by log t).

probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.5, 0.6, 0.7]$, $p_{21} = [0.5, 0.4, 0.3]$, and the rewards for all arms at states 1,2 are $r_1 = [4, 1, 1.2]$, $r_2 = [6, 3, 1.8]$, respectively. In global state 2, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.55, 0.65, 0.75]$, $p_{21} = [0.45, 0.35, 0.25]$, and the rewards for all arms at states 1,2 are $r_1 = [5, 9, 4.5]$, $r_2 = [7, 11, 8.5]$, respectively. In global state 3, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.52, 0.62, 0.72]$, $p_{21} = [0.48, 0.38, 0.28]$, and the rewards for all arms at states 1,2 are $r_1 = [9.9, 9.5, 14]$, $r_2 = [10.3, 11.5, 16]$, respectively. In this scenario, for each global state s, there is a different best arm that is significantly better then the other two arms. Thus, playing the best arm on average in each time slot results in poor performances compared to LEMP, that tracks the best arm at each step, as can be seen in Fig. 5. LEMP outperforms the extended DSEE algorithm also in this scenario.

4) Decreasing the Difference Between the Highest and the Second Highest Values (S4): Finally, we simulated S4, where we decreased the difference between the highest and the second highest values in global state 1, compared to this difference in the first simulation (i.e., in Fig. 3). The transition probabilities of the global chain are $p_{00} = 0.4$, $p_{10} = 0.75$. In global state 1, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.5, 0.6, 0.7]$, $p_{21} = [0.5, 0.4, 0.3]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [4, 5.8, 1]$, $r_2 = [6, 9.2, 2]$. In global state 2, the local transition probabilities for all arms to transition from 1 to 1 and from 2 to 1, respectively, are: $p_{11} = [0.55, 0.65, 0.75]$, $p_{21} = [0.45, 0.35, 0.25]$, and the rewards for all arms at states 1,2, respectively, are $r_1 = [10, 9, 2.5]$, $r_2 = [14, 11, 3]$. Decreasing the difference between the highest and the second highest values in global state 1 results in a high exploration rate used to distinguish between the two best arms in this state. As discussed in Section III, LEMP explores only these two arms using the high exploration rate, where the extended version of the DSEE algorithm explores all the arms with the high exploration rate. Indeed, as can be seen in Fig. 6, this effect results in a high regret under the extended DSEE algorithm as compared to LEMP, which also outperforms the strategy that chooses the best arm on average.

We finally note that practically, it is well known that there is often a gap between the sufficient conditions required by theoretical analysis (often due to union-bounding events in the analysis) and practical conditions used for efficient online learning. While the sufficient conditions provided by the theoretical analysis in Section IV require to overestimate $\overline{D}_s^i(t)$ as in (6), simulation results provide much better performance when higher values of I_L and I_G and lower value of L are used. Thus, we tuned the parameters in all algorithms that we tested to achieve the best performance.

VI. Conclusion

We developed a novel Learning under Exogenous Markov Process (LEMP) algorithm for an extended version of the RMAB problem, where an exogenous Markov global process governs the distribution of the arms. Inspired by recent developments of sequencing methods of exploration and exploitation phases, LEMP estimates the hardness parameter of the problem which controls the size of exploration phases. During the exploitation phases, LEMP switches arms dynamically according to the global process evolution. Simulation results support the theoretical analysis, and show superior performances of the proposed LEMP algorithm against competitive strategies. The model and analytical results presented in this paper lead to interesting future research directions. One direction would consider the case where the global process is not directly observed (e.g., when the reward is a probabilistic function of the unobserved finite-state Markov process), by incorporating methods from hidden Markov model literature [48]. Another open problem is to derive a lower bound for the regret in our model, to understand the optimality properties of the LEMP algorithm. We leave this for future work.

Appendix A
Proof of Lemma 1

Proof of Lemma 1: First note that $\mathbb{E}[T_1]$ can be written as follows:

$$\mathbb{E}[T_1] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}(T_1 = n) = \sum_{n=1}^{\infty} \mathbb{P}(T_1 \geq n)$$

$$\leq \sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{K}_s} \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} \mathbb{P}\left(\widehat{D}_s^i(j) < \overline{D}_s^i \text{ or } \widehat{D}_s^i(j) > \overline{D}_{s,\max}^i\right)$$

$$+ \sum_{s\in\mathcal{S}} \sum_{i\notin\mathcal{K}_s} \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} \mathbb{P}\left(\widehat{D}_s^i(j) < \overline{D}_s^i\right).$$

Note that if we show that

$$\mathbb{P}\left(\widehat{D}_s^i(j) < \overline{D}_s^i \text{ or } \widehat{D}_s^i(j) > \overline{D}_{s,\max}^i\right) \leq C \cdot j^{-(2+\delta)}, \quad (16)$$

for some constants $C > 0, \delta > 0$ for all $i \in \mathcal{K}_s, s \in \mathcal{S}$ for all $j \geq n$, then we get:

$$\sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{K}_s} \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} \mathbb{P}\left(\widehat{D}_s^i(j) < \overline{D}_s^i \text{ or } \widehat{D}_s^i(j) > \overline{D}_{s,\max}^i\right)$$

$$
\begin{aligned}
&\le |\mathcal{S}|NC\left[\sum_{j=1}^{\infty} j^{-(2+\delta)} + \sum_{n=2}^{\infty}\sum_{j=n}^{\infty} j^{-(2+\delta)}\right]\\
&\le |\mathcal{S}|NC\left[\sum_{j=1}^{\infty} j^{-(2+\delta)} + \sum_{n=2}^{\infty}\int_{n-1}^{\infty} j^{-(2+\delta)}dj\right]\\
&= |\mathcal{S}|NC\left[\sum_{j=1}^{\infty} j^{-(2+\delta)} + \frac{1}{1+\delta}\sum_{n=2}^{\infty}(n-1)^{-(1+\delta)}\right]\\
&< \infty,
\end{aligned}
$$

which is bounded independent of t. Similarly, showing that $\mathbb{P}(\widehat{D}_s^i(j) < \overline{D}_s^i) \le C \cdot j^{-(2+\delta)}$ for some constants $C, \delta > 0$ for all $i \notin \mathcal{K}_s, s \in \mathcal{S}$ for all $j \ge n$ completes the statement.

Step 1. Simplifying (16): First,

$$
\begin{aligned}
&\mathbb{P}\left(\widehat{D}_s^i(t) < \overline{D}_s^i \quad \text{or} \quad \widehat{D}_s^i(t) > \overline{D}_{s,\max}^i\right)\\
&= \mathbb{P}\Bigg(\frac{4L}{\max\left\{\Delta, (\hat{V}_s^*(t) - \hat{V}_s^i(t))^2 - \epsilon\right\}} < \frac{4L}{(V_s^* - V_s^i)^2}\\
&\quad \bigcup \frac{4L}{\max\left\{\Delta, (\hat{V}_s^*(t) - \hat{V}_s^i(t))^2 - \epsilon\right\}} > \frac{4L}{\left(V_s^* - V_s^i\right)^2 - 2\epsilon}\Bigg)\\
&= \mathbb{P}\bigg(\bigg[\bigg(\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon > (V_s^* - V_s^i)^2\\
&\qquad \cap \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon \ge \Delta\bigg)\\
&\bigcup \left(\Delta > \left(V_s^* - V_s^i\right)^2 \cap \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon < \Delta\right)\bigg]\\
&\bigcup \bigg[\bigg(\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon < \left(V_s^* - V_s^i\right)^2 - 2\epsilon\\
&\qquad \cap \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon \ge \Delta\bigg)\\
&\cup \left(\Delta < \left(V_s^* - V_s^i\right)^2 - 2\epsilon \cap \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon < \Delta\right)\bigg]\bigg)\\
&\le \mathbb{P}\bigg(\bigg[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon > \left(V_s^* - V_s^i\right)^2\\
&\bigcup \Delta > \left(V_s^* - V_s^i\right)^2\bigg]\\
&\bigcup \bigg[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon < \left(V_s^* - V_s^i\right)^2 - 2\epsilon\\
&\cup \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \epsilon < \Delta\bigg]\bigg).
\end{aligned}
$$

The probability for the second event on the RHS is zero, and the forth event lies inside the measure of the third event due to the fact that $i \in \mathcal{K}_s$. Hence,

$$
\begin{aligned}
&\mathbb{P}\left(\widehat{D}_s^i(t) < \overline{D}_s^i \quad \mathit{or} \quad \widehat{D}_s^i(t) > \overline{D}_{s,\max}^i\right)\\
&\le \mathbb{P}\bigg(\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \left(V_s^* - V_s^i\right)^2 > \epsilon\\
&\bigcup \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \left(V_s^* - V_s^i\right)^2 < -\epsilon\bigg)\\
&= \mathbb{P}\bigg(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \left(V_s^* - V_s^i\right)^2\right| > \epsilon\bigg\}\\
&= \mathbb{P}\bigg(\bigg|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)^2 - \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)\left(V_s^* - V_s^i\right)\\
&\quad + \left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)\left(V_s^* - V_s^i\right) - \left(V_s^* - V_s^i\right)^2 | > \epsilon\bigg\}\\
&= \mathbb{P}\bigg(\bigg|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\\
&\quad + \left(V_s^* - V_s^i\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\bigg| > \epsilon\bigg\}\\
&\le \mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\right| > \frac{\epsilon}{2}\right)\\
&\quad + \mathbb{P}\left(\left|\left(V_s^* - V_s^i\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\right| > \frac{\epsilon}{2}\right).
\end{aligned} \tag{17}
$$

We continue by bounding the first term on the RHS of (17). For every $R > 0$, we have:

$$
\begin{aligned}
&\mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\right| > \frac{\epsilon}{2}\right)\\
&\le \mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > 1\right)\\
&\quad + \mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > \frac{\epsilon}{2(R+1)}\right)\\
&\quad + \mathbb{P}\left(\left|\left(V_s^* - V_s^i\right) + 1\right| > R\right)\\
&\le 2\mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > \frac{\epsilon}{2(R+1)}\right)\\
&\quad + \mathbb{P}\left(V_s^* + 1 > R\right).
\end{aligned}
$$

We choose $R = V_s^* + 1$. Then, the second term is equal to 0. We proceed with the first term:

$$
\begin{aligned}
&2 \cdot \mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > \frac{\epsilon}{2\left(V_s^* + 2\right)}\right)\\
&\qquad \le 2\mathbb{P}\left(\left|\hat{V}_s^*(t) - V_s^*\right| > \frac{\epsilon}{4(V_s^* + 2)}\right)\\
&\qquad\quad + 2\mathbb{P}\left(\left|\hat{V}_s^i(t) - V_s^i\right)\right| > \frac{\epsilon}{4(V_s^* + 2)}\right).
\end{aligned} \tag{18}
$$

We next bound the second term on the RHS of (17). For every $R' > 0$, we have:

$$
\begin{aligned}
&\mathbb{P}\left(\left|\left(V_s^* - V_s^i\right)\left[\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right]\right| > \frac{\epsilon}{2}\right)\\
&\le \mathbb{P}\left(V_s^* > R'\right)\\
&\quad + \mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > \frac{\epsilon}{2\left(R' + 1\right)}\right).
\end{aligned}
$$

We now choose $R' = R = V_s^* + 1$, so the first term is equal to 0. We continue with the second term:

$$
\mathbb{P}\left(\left|\left(\hat{V}_s^*(t) - \hat{V}_s^i(t)\right) - \left(V_s^* - V_s^i\right)\right| > \frac{\epsilon}{2\left(R' + 1\right)}\right)
$$

$$\le \mathbb{P}\left(\left|\hat{V}_s^*(t) - V_s^*\right| > \frac{\epsilon}{4(V_s^*+2)}\right)$$
$$+\,\mathbb{P}\left(\left|\hat{V}_s^i(t) - V_s^i)\right| > \frac{\epsilon}{4(V_s^*+2)}\right). \tag{19}$$

By combining (18) and (19) we get:

$$\mathbb{P}\left(\overline{D}_s^i(t) < \widehat{D}_s^i \quad \text{or} \quad \overline{D}_s^i(t) > \overline{D}_{s,\max}^i\right)$$
$$\le 6\cdot\max\left\{\mathbb{P}\left(\left|\hat{V}_s^*(t) - V_s^*\right| > \frac{\epsilon}{4(V_s^*+2)}\right),\right.$$
$$\left.\mathbb{P}\left(\left|\hat{V}_s^i(t) - V_s^i\right| > \frac{\epsilon}{4(V_s^*+2)}\right)\right\}. \tag{20}$$

Step 2. Bounding (20): We first bound the second term in (20) (the first term is bounded similarly).

$$\mathbb{P}\left(\left|\hat{V}_s^i(t) - V_s^i\right| > \frac{\epsilon}{4(V_s^*+2)}\right)$$
$$= \mathbb{P}\left(\left|\sum_{s'\in\mathcal{S}}\hat{p}_{ss'}(t)\,\hat{\mu}_{s'}^i(t) - \sum_{s'\in\mathcal{S}}p_{ss'}\mu_{s'}^i\right| > \frac{\epsilon}{4(V_s^*+2)}\right)$$
$$\le \mathbb{P}\left(\left|\sum_{s'\in\mathcal{S}}\left(\hat{p}_{ss'}(t)\,\hat{\mu}_{s'}^i(t) - p_{ss'}\mu_{s'}^i - \frac{\epsilon}{4(V_s^*+2)|\mathcal{S}|}\right)\right| > 0\right)$$
$$\le \sum_{s'\in\mathcal{S}}\mathbb{P}\left(\left|\hat{p}_{ss'}(t)\,\hat{\mu}_{s'}^i(t) - p_{ss'}\mu_{s'}^i\right| > \frac{\epsilon}{4(V_s^*+2)|\mathcal{S}|}\right).$$

Following similar steps as we did to obtain (20) from (17), we get

$$\mathbb{P}\left(\left|\hat{p}_{ss'}(t)\,\hat{\mu}_{s'}^i(t) - p_{ss'}\mu_{s'}^i\right| > \frac{\epsilon}{4(V_s^*+2)|\mathcal{S}|}\right)$$
$$\le 2\mathbb{P}\left(\left|\hat{p}_{ss'}(t) - p_{ss'}\right| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)\|\mathcal{S}|}\right) \tag{21}$$
$$+\,\mathbb{P}\left(\left|\hat{\mu}_{s'}^i(t) - \mu_{s'}^i\right| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}\right). \tag{22}$$

To complete the statement, we need to bound (21) and (22). To bound (22), we will use Lezaud's result [59].

Lemma 5 ([59]): Consider a finite-state, irreducible Markov chain $\{X_t\}_{t\ge1}$ with state space S, matrix of transition probabilities P, an initial distribution q, and stationary distribution π. Let $N_{\mathbf{q}} = \|(\frac{q_x}{\pi_x}, x\in S)\|_2$. Let $\widehat{P} = P'P$ be the multiplicative symmetrization of P where P' is the adjoint of P on $l_2(\pi)$. Let $\epsilon = 1-\lambda_2$, where λ_2 is the second largest eigenvalue of the matrix P'. ϵ will be referred to as the eigenvalue gap of P'. Let $f: S\to\mathcal{R}$ be such that $\sum_{y\in S}\pi_y f(y) = 0, \quad \|f\|_2 \le 1$ and $0\le\|f\|_2^2\le 1$ if P' is irreducible. Then, for any positive integer n and all $0<\lambda\le1$, we have:

$$Pr\left(\frac{1}{n}\sum_{t=1}^{n} f(X_t) \ge \lambda\right) \le N_{\mathbf{q}}\exp\left[-\frac{n\lambda^2\epsilon}{12}\right].$$

Consider an initial distribution $\mathbf{q}_s^i$ for the ith arm in global state s. We have:

$$\left\|\left(\frac{q_s^i(x)}{\pi_s^i(x)}, x\in\mathcal{X}_s^i\right)\right\|_2 \le \sum_{x\in\mathcal{X}_s^i}\left\|\frac{q_s^i(x)}{\pi_s^i(x)}\right\|_2 \le \frac{1}{\pi_{\min}}.$$

Before applying Lezaud's bound, we pay attention for the following: (i) The sample means $\{\hat{\mu}_s^i(t)\}$ are calculated only from measurements in the set $\mathcal{V}_i$. As discussed in Section III-B, these measurements are equivalent to a sample path generated by continuously sampling the Markov chain. Hence, we can apply Lezaud's bound to upper bound (22). (ii) By the construction of the algorithm, (10) ensures that once exploitation phases are executed (which are deterministic), the event $T_s^i(t) \ge \frac{(2+\delta)}{\epsilon^2 I_L}\log(t)$ for $\delta>0$ arbitrarily small surely occurs[2]. During exploration phases, the randomness of SB1 (say for arm $r\ne i$) affects $T_s^i(t)$ since SB1 can be very long (with small probability) and then $T_s^i(t) \ge \frac{(2+\delta)}{\epsilon^2 I_L}\log(t)$ might not hold until the end of the phase once the algorithm corrects the exploration gap by condition (10). Therefore, we define $E_i(t)$ as the event when all SB1 phases that have been executed by time t are smaller than $\delta\cdot t$. When event $E_i(t)$ occurs we have $T_s^i(t) \ge \frac{(2+\delta)}{\epsilon^2 I_L}\log(t)$ (for all $t>D$, for a sufficiently large finite deterministic value D). Then, for all i and s', we have:

$$\mathbb{P}\left(|\hat{\mu}_{s'}^i(t) - \mu_{s'}^i| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}\right)$$
$$= \mathbb{P}\left(|\hat{\mu}_{s'}^i(t) - \mu_{s'}^i| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}, E_i(t)\text{ occurs}\right)$$
$$+\,\mathbb{P}\left(|\hat{\mu}_{s'}^i(t) - \mu_{s'}^i| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|},\right.$$
$$\left.(E_i(t)\text{ does not occur})\vphantom{\frac{1}{1}}\right)$$
$$\le \mathbb{P}\left(|\hat{\mu}_{s'}^i(t) - \mu_{s'}^i| > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}, E_i(t)\text{ occurs}\right) \tag{23}$$
$$+\,\mathbb{P}\left(E_i(t)\text{ does not occur}\right). \tag{24}$$

We next bound (22) by bounding (23) and (24):
We define $O^{i,x}{}_s(t)$ as the number of occurrences of local state x on arm i in global state s up to time t, and we first look at:

$$\mathbb{P}\left(\hat{\mu}_{s'}^i(t) - \mu_{s'}^i > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}, E_i(t)\right)$$
$$= \mathbb{P}\left(\sum_{x\in\mathcal{X}_s^i} x\cdot O_s^{i,x}(t) - T_s^i(t)\sum_{x\in\mathcal{X}_s^i} x\cdot\pi_s^i(x)\right.$$
$$\left.> \frac{T_s^i(t)\cdot\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}, E_i(t)\right)$$

[2] We point out that a precise statement requires to set $(2+2\delta)$ in (10) and the statement holds for all $t>D$, where D is a finite deterministic value. However, since $\delta>0$ is arbitrarily small and is not a design parameter, we do not present it explicitly when describing the algorithm to simplify the presentation.

$$\leq \sum_{x\in\mathcal{X}_s^i} \mathbb{P}\Bigg(\frac{\sum_{n=1}^{t} \mathbf{1}\left(x_s^i\left(n\right)=x\right) - T_s^i\left(t\right)\pi_s^i\left(x\right)}{\hat{\pi}_s^i\left(x\right)\cdot T_s^i\left(t\right)}$$

$$> \frac{T_s^i\left(t\right)\cdot\epsilon}{16\left(V_s^*+2\right)\left(x_{\max}+2\right)\left|\mathcal{S}\right|\left|\mathcal{X}_s^i\right|\cdot x\hat{\pi}_s^i\left(x\right)}, E_i\left(t\right)\Bigg)$$

$$\leq |\mathcal{X}_s^i| N_{s,\mathbf{q}}^{(i)}$$

$$\cdot \exp\left(-T_s^i\left(t\right)\frac{\epsilon^2}{(16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|)^2\cdot x_{\max}^2|\mathcal{X}_s^i|^2\hat{\pi}_{\max}^2}\frac{\bar{\lambda}_{\min}}{12}\right)$$

and due to $E_i(t)$: $T_s^i(t) > \frac{2+\delta}{\epsilon^2 I_L}\cdot\log(t)$, so we have:

$$\mathbb{P}\left(\hat{\mu}_{s'}^i\left(t\right)-\mu_{s'}^i > \frac{\epsilon}{16\left(V_s^*+2\right)\left(x_{\max}+2\right)\left|\mathcal{S}\right|}, E_i\left(t\right)\right)$$

$$\leq \frac{|\mathcal{X}_{\max}|}{\pi_{\min}}$$

$$\exp\left(-\frac{(2+\delta)}{\epsilon^2 I_L}\frac{\epsilon^2\cdot\bar{\lambda}_{\min}}{12\cdot16^2(V_s^*+2)^2(x_{\max}+2)^2x_{\max}^2|\mathcal{X}_s^i|^2|\mathcal{S}|^2\hat{\pi}_{\max}^2}\log\left(t\right)\right)$$

$$\leq \frac{|\mathcal{X}_{\max}|}{\pi_{\min}}\cdot e^{-(2+\delta)\cdot\log(t)} = \frac{|\mathcal{X}_{\max}|}{\pi_{\min}}\cdot t^{-(2+\delta)}, \qquad (25)$$

for some $\delta > 0$ arbitrarily small.

Together with applying Lemma 3 to $-f$, we get the bound for (23). We next upper bound (24). When event $E_i(t)$ does not occur, there exists an SB1 phase (i.e., hitting time) which is greater than $\delta\cdot t$. Since all arms are finite state and irreducible Markov chains (i.e., the hitting time of some state x starting from state y is almost surely finite), we have that $\mathbb{P}(E_i(t)$ does not occur$)\leq e^1\cdot e^{-\frac{\delta}{|\mathbf{X}_{\max}|}t}$. Therefore for all $t > D$, for a sufficiently large finite deterministic value D, we have $\mathbb{P}(E_i(t)$ does not occur$)\leq \frac{|\mathcal{X}_{\max}|}{\pi_{\min}}t^{-(2+\delta)}$, which completes (22). (21) is bounded by:

$$\mathbb{P}\left(\hat{p}_{ss'}(t)-p_{ss'} > \frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}\right)$$

$$\leq \exp\left(-2N_s(t)\cdot\left(\frac{\epsilon}{16(V_s^*+2)(x_{\max}+2)|\mathcal{S}|}\right)^2\right) \qquad (26)$$

The bound in (26) follows similar steps as in Part A of Appendix A in [56]. Using condition (11), $N_s(t) > \frac{2+\delta}{\epsilon^2\cdot I_G}\cdot\log(t)$ for some arbitrarily small $\delta > 0$, we conclude that (26) is bounded by $t^{(2+\delta)}$, which proves that (16) is bounded by $4\frac{|\mathcal{X}_{\max}|}{\pi_{\min}}\cdot t^{-(2+\delta)}$.

Finally, showing that $\mathbb{P}(\widehat{D}_s^i(j) < \overline{D}_s^i)\leq j^{-(2+\delta)}$ for some $\delta > 0$ for all $i\notin\mathcal{K}_s$ for all $j\geq n$ follows similar steps as showed by handling $\widehat{D}_s^i(j) < \overline{D}_s^i$ when proving (16). Thus, Lemma 1 follows.

Appendix B
Proof of Lemma 2

Next, we prove the upper bound (15). Note that the regret can be written as follows:

$$\mathbb{E}_\phi[r(t)] = \mathbb{E}_\phi\left[\sum_{n=1}^{t}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]$$

$$= \mathbb{E}_\phi\left[\sum_{n=1}^{T_1}\sum_{i\in V_c(n)}\left(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n)\right)\mathbb{1}_{\{\phi(n)=i\}}\right] \qquad (27)$$

$$+\mathbb{E}_\phi\left[\sum_{n=T_1+1}^{t}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]. \qquad (28)$$

By applying Lemma 1, we obtain that (27) is bounded independent of t:

$$\mathbb{E}_\phi\left[\sum_{n=1}^{T_1}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]\leq x_{\max}\mathbb{E}_\phi[T_1],$$

which results in the additional constant term $O(1)$ in the regret bound in (13) which is independent of t.

Next, we upper bound (28). Note that for all $t > T_1$, we have:

$$\overline{D}_s^i\leq\widehat{D}_s^i(t)\leq\overline{D}_{s,\max}^i, \qquad (29)$$

for all $s\in\mathcal{S}, i\in\mathcal{K}_s$, and we have the LHS of the inequality for $i\notin\mathcal{K}_s$. For convenience, we will develop (28) between $n=1$ and t with (29) (and the LHS for $i\notin\mathcal{K}_s$) holds for all $1\leq n\leq t$, which upper bounds (28) between $n=T_1+1$ and t:

$$\mathbb{E}_\phi\left[\sum_{n=T_1+1}^{t}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]$$

$$\leq\mathbb{E}_\phi\left[\sum_{n=1}^{t}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]. \qquad (30)$$

Finally, note that:

$$\mathbb{E}_\phi\left[\sum_{n=1}^{t}\sum_{i\in V_c(n)}(x_{s_n}^{i_n^*}(n)-x_{s_n}^i(n))\mathbb{1}_{\{\phi(n)=i\}}\right]$$

$$\leq x_{\max}\sum_{i=1}^{N}\mathbb{E}_\phi[\tilde{T}^i(t)]. \qquad (31)$$

Appendix C
Proof of Lemma 3

Proof of Lemma 3: We first upper bound the number of exploration phases $n_O^i(t)$ for each arm (say i) by time t. If the player has started the n^{th} exploration phase, we have by (10) and the fact that $t\geq T_1$:

$$\sum_{n=1}^{n_O^i(t)}4^{n-1}=\frac{1}{3}\left(4^{n_O^i(t)}-1\right)\leq A_i\cdot\log(t).$$

Hence, $n_O^i(t)\leq\lfloor\log_4(3A_i\log(t)+1)\rfloor+1$.

Next, note that exploration phase $n_O^i(t)$ for arm i consists of the time until the last state observed at the $(n_O^i(t)-1)^{th}$ exploration phase $\gamma^i(n_O^i-1)$ is observed again (i.e., SB1 sub-block), and another $4^{n_O^i(t)}$ time slots. Thus, the time spent by time t in exploration phases for arm i is bounded by:

$$\begin{aligned}\mathbb{E}[T_O^i(t)] &\leq \sum_{n=0}^{n_O^i(t)-1} \left(4^n + M_{\max}^i\right) \\ &= \frac{1}{3}\left(4^{n_O^i(t)} - 1\right) + M_{\max}^i \cdot n_O^i(t) \\ &\leq \frac{1}{3}\left[4(3A_i \cdot \log(t) + 1) - 1\right] + M_{\max}^i \cdot \log_4(3A_i \log(t) + 1).\end{aligned}$$

Appendix D
Proof of Lemma 4

Proof of Lemma 4: We first upper bound the number of exploitation phases by time t, $n_I(t)$. By time t, at most t time slots have been spent on exploitation phases. Thus, we have:

$$\sum_{n=1}^{n_I(t)} 2 \cdot 4^{n-1} \leq t,$$

which implies that $\frac{2}{3}(4^{n_I(t)} - 1) \leq t$. Hence,

$$n_I(t) \leq \left\lceil \log_4\left(\frac{3}{2}t + 1\right)\right\rceil. \tag{32}$$

Next, we use (32) to bound the regret caused by choosing sub-optimal arms in exploitation phases. Let $T_{s,I}^i(t)$ denotes the time spent on sub-optimal arm i in global state s in exploitation phases, by time t (note that $T_I^i(t) = \sum_{s\in\mathcal{S}} T_{s,I}^i(t) \leq |\mathcal{S}| \max_s\{T_{s,I}^i(t)\}$). We define $Pr[i,s,n]$ as the probability that a sub-optimal arm i is played when the global state is s in the n^{th} exploitation phase. From (32) we have:

$$\begin{aligned}\mathbb{E}[T_{s,I}^i(t)] &\leq \sum_{n=1}^{n_I} 2 \cdot 4^{n-1} \cdot \pi_s \cdot Pr[i,s,n] \\ &\leq \sum_{n=1}^{\lceil \log_4(\frac{3}{2}t+1)\rceil} 2 \cdot 4^{n-1} \cdot \pi_s \cdot Pr[i,s,n] \\ &\leq \sum_{n=1}^{\lceil \log_4(\frac{3}{2}t+1)\rceil} 3t_n \cdot \pi_s \cdot Pr[i,s,n],\end{aligned} \tag{33}$$

where t_n denotes the starting time of the n^{th} exploitation phase and (33) follows from the fact that $t_n \geq \frac{2}{3}4^{n-1}$. Note that it suffices to show that $Pr[i,s,n]$ has an order of t_n^{-1} so as to obtain a logarithmic order with time for the summation in (33).

Next, we bound $Pr[i,s,n]$. We define $C_{s,t}^i = \sqrt{L\log(t)/T_s^i(t)}, C_{s,t}^* = \sqrt{L\log(t)/T_s^*(t)}$, where $T_s^*(t)$ denotes the number of plays on the best arm of global state s, i_s^*, by time t.

$$\begin{aligned}Pr[i,s,n] &= \mathbb{P}\left(\hat{V}_s^i(t_n) \geq \hat{V}_s^*(t_n)\right) \\ &\leq \mathbb{P}\left(\hat{V}_s^*(t_n) \leq V_s^* - C_{s,t_n}^*\right) \\ &\quad + \mathbb{P}\left(\hat{V}_s^i(t_n) \geq V_s^i + C_{s,t_n}^i\right) \\ &\quad + \mathbb{P}\left(V_s^* < V_s^i + C_{s,t_n}^i + C_{s,t_n}^*\right).\end{aligned} \tag{34}$$

We first show that the third term in (34) is zero. Note that from (10) we have:

$T_s^i(t) > \max\{\widehat{D}_s^i(t), \frac{2}{\epsilon^2 I_L}\} \cdot \log t_n$,

and from (29) and the fact that $\overline{D}_s^i \leq \max_i \overline{D}_s^i$, we have:

$\min\{T_s^*, T_s^i\} \geq \overline{D}_s^i \cdot \log t_n$.

As a result,

$$\begin{aligned}&\mathbb{P}\left(V_s^* < V_s^i + C_{s,t_n}^i + C_{s,t_n}^*\right) \\ &= \mathbb{P}\left(V_s^* - V_s^i < \sqrt{\frac{L\log t_n}{T_s^i(t_n)}} + \sqrt{\frac{L\log t_n}{T_s^*(t_n)}}\right) \\ &\leq \mathbb{P}\left(V_s^* - V_s^i < 2\sqrt{\frac{L\log t_n}{\min\left\{T_s^*(t_n), T_s^i(t_n)\right\}}}\right) \\ &= \mathbb{P}\left(\left(V_s^* - V_s^i\right)^2 < \frac{4\,L\log t_n}{\min\left\{T_s^*(t_n), T_s^i(t_n)\right\}}\right) \\ &= \mathbb{P}\left(\min\left\{T_s^*(t_n), T_s^i(t_n)\right\} < \overline{D}_s^i \cdot \log t_n\right) = 0.\end{aligned}$$

Therefore, we can rewrite (34) as follows:

$$\begin{aligned}Pr[i,s,n] &\leq \mathbb{P}\left(\hat{V}_s^i(t_n) \geq \hat{V}_s^*(t_n)\right) \\ &\leq \mathbb{P}\left(\hat{V}_s^*(t_n) \leq V_s^* - C_{s,t_n}^*\right) \\ &\quad + \mathbb{P}\left(\hat{V}_s^i(t_n) \geq V_s^i + C_{s,t_n}^i\right).\end{aligned} \tag{35}$$

Next, we bound both terms on the RHS of (35). Using similar steps as we used for bounding the second term in (20), we get:

$$\begin{aligned}&\mathbb{P}\left(\hat{V}_s^i(t_n) - V_s^i \geq C_{s,t_n}^i\right) \\ &\leq 2|\mathcal{S}|\mathbb{P}\left(|\hat{p}_{ss'}(t_n) - p_{ss'}| \geq \frac{1}{4|\mathcal{S}|(x_{\max}+2)} \cdot C_{s,t_n}^i\right) \\ &\quad + |\mathcal{S}|\mathbb{P}\left(|\hat{\mu}_{s'}^i(t_n) - \mu_{s'}^i| \geq \frac{1}{4|\mathcal{S}|(x_{\max}+2)} \cdot C_{s,t_n}^i\right).\end{aligned} \tag{36}$$

The second term in (36) is bounded similarly as in (22):

$$\begin{aligned}&|\mathcal{S}|\mathbb{P}\left(|\hat{\mu}_{s'}^i(t_n) - \mu_{s'}^i| \geq \frac{1}{4|\mathcal{S}|(x_{\max}+2)} \cdot C_{s,t_n}^i\right) \\ &\quad\leq \frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} \\ &\cdot \exp\left(-T_s^i(t_n)\,\tfrac{L\log(t_n)}{T_s^i(t_n)}\,\tfrac{1}{16(x_{\max}+2)^2|\mathcal{S}|^2}\,\tfrac{\bar{\lambda}_{\min}}{12(x_{\max}|\mathcal{X}_{\max}|\pi_{\max})^2}\right) \\ &\quad\leq \frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} \cdot t_n^{-1},\end{aligned}$$

where the last inequality is due to (12). The first term in (36) is bounded similarly as in (21):

$$\begin{aligned}&2|\mathcal{S}|\mathbb{P}\left(|\hat{p}_{ss'}(t_n) - p_{ss'}| \geq \frac{1}{4|\mathcal{S}|(x_{\max}+2)} \cdot C_{s,t_n}^i\right) \\ &\leq 2|\mathcal{S}|\exp\left(-2N_s(t_n) \cdot \frac{1}{16|\mathcal{S}|^2(x_{\max}+2)^2} \cdot \frac{L \cdot \log(t_n)}{N_s(t_n)}\right)\Bigg) \\ &\leq 2|\mathcal{S}|t_n^{-1},\end{aligned}$$

where the last inequality is due to (12), and the fact that $\forall i: N_s(t) > T_s^i(t)$. The first term in (35) is bounded similarly, and therefore:

$$Pr[i,s,n] \le 2\left(\frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} + 2|\mathcal{S}|\right)\cdot t_n^{-1}. \qquad (37)$$

Using (37), we can bound (33), and therefore:

$$\begin{aligned}\mathbb{E}\left[T_I^i(t)\right] \le\ & 6|\mathcal{S}|\cdot\left(\frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} + 2|\mathcal{S}|\right)\cdot\max_s \pi_s \\ & \cdot\left\lceil \log_4\left(\frac{3}{2}t+1\right)\right\rceil. \end{aligned} \qquad (38)$$

Appendix E
Incorporating the Regret Events to Prove Theorem 1

We conclude the proof of Theorem 1 by incorporating the regret events discussed above., i.e., regret that is caused by imprecise estimation of the exploration rate, regret that is caused by playing bad arms in exploration phases, and regret that is caused by playing bad arms in exploitation phases.

Lemmas 1 and 2 result in the additional constant term $O(1)$ in the regret bound (13) which is independent of t.

From Lemmas 3 and 2, the regret caused by playing bad arms in exploration phases by time t is bounded by:

$$\begin{aligned} x_{\max}\cdot\sum_{i=1}^{N}\Bigg[&\frac{1}{3}[4(3A_i\cdot\log(t)+1)-1] \\ &+ M_{\max}^i\cdot\log_4(3A_i\log(t)+1)\Bigg], \end{aligned}$$

which coincides with the first and second terms on the RHS of (13).

From Lemmas 4 and 2, the regret caused by playing sub-optimal arms in exploitation phases by time t is bounded by:

$$\begin{aligned} x_{\max}\cdot N\cdot 6|\mathcal{S}|\cdot&\left(\frac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} + 2|\mathcal{S}|\right) \\ &\cdot\max_s \pi_s\cdot\left\lceil \log_4\left(\frac{3}{2}t+1\right)\right\rceil, \end{aligned}$$

which coincides with the third term on the RHS of (13), and thus Theorem 1 follows.

Acknowledgment

In this journal version we include: (i) A detailed discussion on the implementation of the algorithm; (ii) a deep theoretical analysis of the algorithm with detailed proofs; (iii) more extensive simulation results; and (iv) a detailed discussion of the results, and comprehensive discussion and comparison with the existing literature.

References

[1] T. Gafni, M. Yemini, and K. Cohen, "Restless multi-armed bandits under exogenous global Markov process," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 5218–5222.

[2] J. Sun, Y. Zhao, N. Zhang, X. Chen, Q. Hu, and J. Song, "A dynamic distributed energy storage control strategy for providing primary frequency regulation using multi-armed bandits method," *IET Gener., Transmiss. Distrib.*, vol. 16, no. 4, pp. 669–679, Feb. 2022.

[3] S. L. Scott, "Multi-armed bandit experiments in the online service economy," *Appl. Stochastic Models Bus. Ind.*, vol. 31, no. 1, pp. 37–45, 2015.

[4] C. Bulucu, "Personalizing treatments via contextual multi-armed bandits by identifying relevance," Ph.D. dissertation, Elect. Elect. Eng., Bilkent Univ., Çankaya/Ankara, Turkey, 2019.

[5] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2903–2915, Dec. 2020.

[6] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 14–41, May 2022.

[7] Y. Xu, P. Cheng, Z. Chen, M. Ding, Y. Li, and B. Vucetic, "Task offloading for large-scale asynchronous mobile edge computing: An index policy approach," *IEEE Trans. Signal Process.*, vol. 69, pp. 401–416, 2021.

[8] O. Amar and K. Cohen, "Online learning for shortest path and backpressure routing in wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 2702–2707.

[9] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[10] S. Geirhofer, L. Tong, and B. M. Sadler, "A measurement-based model for dynamic spectrum access in wlan channels," in *Proc. IEEE Mil. Commun. Conf.*, 2006, pp. 1–7.

[11] K. Wang and L. Chen, "On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 300–309, Jan. 2012.

[12] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[13] G. Elena, K. Milos, and I. Eugene, "Survey of multiarmed bandit algorithms applied to recommendation systems," *Int. J. Open Inf. Technol.*, vol. 9, no. 4, pp. 12–27, 2021.

[14] S. Kulkarni and S. F. Rodd, "Context aware recommendation systems: A review of the state of the art techniques," *Comput. Sci. Rev.*, vol. 37, 2020, Art. no. 100255.

[15] K. Misra, E. M. Schwartz, and J. Abernethy, "Dynamic online pricing with incomplete information using multiarmed bandit experiments," *Marketing Sci.*, vol. 38, no. 2, pp. 226–252, 2019.

[16] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," in *Proc. IEEE 9th Annu. Conf. Struct. Complexity Theory*, 1994, pp. 318–322.

[17] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 977–982, Nov. 1987.

[18] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.

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

[20] T. Gafni and K. Cohen, "Learning in restless multi-armed bandits via adaptive arm sequencing rules," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 5029–5036, Oct. 2021.

[21] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 2940–2943.

[22] S. Bagheri and A. Scaglione, "The restless multi-armed bandit formulation of the cognitive compressive sensing problem," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1183–1198, Mar. 2015.

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

[24] 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.

[25] J.-Y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits.," in *Proc. COLT*, 2010, pp. 41–53.

[26] S. Shahrampour, M. Noshad, and V. Tarokh, "On sequential elimination algorithms for best-arm identification in multi-armed bandits," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4281–4292, Aug. 2017.

[27] C. Shen, "Universal best arm identification," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4464–4478, Sep. 2019.

[28] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.

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

[30] 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.

[31] C. Tekin and M. Liu, “Online algorithms for the multi-armed bandit problem with Markovian rewards,” in *Proc. IEEE 48th Annu. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 1675–1682.

[32] C. Tekin and M. Liu, “Approximately optimal adaptive learning in opportunistic spectrum access,” in *Proc. IEEE INFOCOM*, 2012, pp. 1548–1556.

[33] J. Xu, L. Chen, and O. Tang, “An online algorithm for the risk-aware restless bandit,” *Eur. J. Oper. Res.*, vol. 290, no. 2, pp. 622–639, 2021.

[34] P. Karthik and R. Sundaresan, “Learning to detect an odd restless Markov arm,” in *Proc. IEEE Int. Symp. Inf. Theory*, 2021, pp. 1457–1462.

[35] T. Gafni and K. Cohen, “A distributed stable strategy learning algorithm for multi-user dynamic spectrum access,” in *Proc. IEEE 57th Annu. Allerton Conf. Commun., Control, Comput.*, 2019, pp. 347–351.

[36] T. Gafni and K. Cohen, “Distributed learning over Markovian fading channels for stable spectrum access,” *IEEE Access*, vol. 10, pp. 46652–46669 2022.

[37] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *J. Appl. Probability*, vol. 25, no. A, pp. 287–298, 1988.

[38] R. R. Weber and G. Weiss, “On an index policy for restless bandits,” *J. Appl. Probability*, vol. 27, no. 3, pp. 637–648, 1990.

[39] N. Ehsan and M. Liu, “On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services,” in *Proc. IEEE INFOCOM*, 2004, vol. 3, pp. 1974–1983.

[40] K. Cohen, Q. Zhao, and A. Scaglione, “Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access,” in *Proc. 48th Asilomar Conf. Signals, Syst., Comput.*, 2014, pp. 1575–1578.

[41] Q. Zhao, B. Krishnamachari, and K. Liu, “On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

[42] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, “Optimality of myopic sensing in multichannel opportunistic access,” *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.

[43] S. H. A. Ahmad and M. Liu, “Multi-channel opportunistic access: A case of restless bandits with multiple plays,” in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, 2009, pp. 1361–1368.

[44] K. Liu and Q. Zhao, “Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access,” *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, Nov. 2010.

[45] K. Wang, L. Chen, and Q. Liu, “On optimality of myopic policy for opportunistic access with nonidentical channels and imperfect sensing,” *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2478–2483, Jun. 2014.

[46] Q. Zhao and B. Krishnamachari, “Structure and optimality of myopic sensing for opportunistic spectrum access,” in *Proc. IEEE Int. Conf. Commun.*, pp. 6476–6481, 2007.

[47] K. Liu, R. Weber, and Q. Zhao, “Indexability and whittle index for restless bandit problems involving reset processes,” in *Proc. IEEE 50th Conf. Decis. Control Eur. Control Conf.*, 2011, pp. 7690–7696.

[48] V. Krishnamurthy and B. Wahlberg, “Partially observed Markov decision process multiarmed bandits–structural results,” *Math. Operations Res.*, vol. 34, no. 2, pp. 287–302, 2009.

[49] V. Krishnamurthy and R. J. Evans, “Hidden Markov model multiarm bandits: A methodology for beam scheduling in multitarget tracking,” *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 2893–2908, Dec. 2001.

[50] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag, “Multi-armed bandit, dynamic environments and meta-bandits,” in *Proc. Workshop, Online Trading Between Exploration Exploitation*, Whistler, Canada, 2006.

[51] J. Y. Yu and S. Mannor, “Piecewise-stationary bandit problems with side observations,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1177–1184.

[52] A. Slivkins and E. Upfal, “Adapting to a changing environment: The Brownian restless bandits,” in *Proc. COLT*, 2008, pp. 343–354.

[53] Z. Wang, R. Zhou, and C. Shen, “Regional multi-armed bandits with partial informativeness,” *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5705–5717, Nov. 2018.

[54] S. Baltaoglu, L. Tong, and Q. Zhao, “Online learning and optimization of Markov jump linear models,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 2289–2293.

[55] S. Baltaoglu, L. Tong, and Q. Zhao, “Online learning and optimization of Markov jump affine models,” 2016, *arXiv:1605.02213*.

[56] M. Yemini, A. Leshem, and A. Somekh-Baruch, “The restless hidden Markov bandit with linear rewards and side information,” *IEEE Trans. Signal Process.*, vol. 69, pp. 1108–1123, 2021.

[57] A. Al-Tahmeesschi, M. López-Benítez, J. Lehtomäki, and K. Umebayashi, “Accurate estimation of primary user traffic based on periodic spectrum sensing,” in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2018, pp. 1–6.

[58] P. Auer, T. Jaksch, and R. Ortner, “Near-optimal regret bounds for reinforcement learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 89–96.

[59] P. Lezaud, “Chernoff-type bound for finite Markov chains,” *Ann. Appl. Probability*, vol. 8, no. 3, pp. 849–867, 1998.

[60] K. Liu and Q. Zhao, “Link throughput of multi-channel opportunistic access with limited sensing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2997–3000.

Tomer Gafni received the B.Sc. and M.Sc. degrees in electrical and computer engineering from Ben-Gurion University, Beersheba, Israel, in 2019 and 2020, respectively. He is currently working toward the Ph.D. degree with the school of electrical and computer engineering, Ben-Gurion University, Israel. His main research interests include sequential learning, federated learning, decision theory, and statistical inference and learning, with applications in large-scale systems and wireless networks. He was the recipient of the Kaufmann Award for the Highest Student Achievement in Electrical Engineering, Ben-Gurion university.

Michal Yemini (Member, IEEE) received the B.Sc. degree in computer engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2011, and the Ph.D. degree from the joint M.Sc.-Ph.D. Program from the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel. He is currently an Assistant Professor with Bar-Ilan University. Prior to that, she was an Associate Research Scholar with Princeton University, Princeton, NJ, USA, a Postdoctoral Researcher with Stanford University, Stanford, CA, USA, and a Visiting Postdoctoral Researcher with Princeton University. Her main research interests include distributed optimization, sequential decision-making, learning theory, information theory, and percolation theory. She was the recipient of the Eric and Wendy Schmidt Postdoctoral Award for Women in Mathematical and Computing Sciences, the Council of Higher Education’s Postdoctoral Fellowships Program for Outstanding Women in Science, and the Bar-Ilan University’s Postdoctoral Fellowship for Women.

Kobi Cohen (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in electrical engineering from Bar-Ilan University, Ramat Gan, Israel, in 2007 and 2013, respectively. He was with the Coordinated Science Laboratory with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, and the Department of Electrical and Computer Engineering with the University of California, Davis, Davis, CA, USA, as a Postdoctoral Research Associate. In October 2015, he joined the School of Electrical and Computer Engineering with Ben-Gurion University of the Negev (BGU), Beersheba, Israel, where he is currently an Associate Professor. He is also a Member of the Cyber Security Research Center, and the Data Science Research Center with BGU. His main research interests include statistical inference and learning, signal processing, communication networks, decision theory and stochastic optimization with applications to large-scale systems, cyber systems, and wireless and wireline networks. Since 2021, he has sbeen an Associate Editor for the IEEE Transactions on Signal Processing, and the IEEE Transactions on Wireless Communications. Other selected awards and Honors include highlighting in top 50 popular paper list, IEEE Transactions on Wireless Communications in 2019 and 2020 for paper: Deep multi-user reinforcement learning for distributed dynamic spectrum access, highlighting in popular paper list, *IEEE Signal Processing Magazine* in 2022 for paper: Federated Learning: A signal processing perspective, receiving the Best Paper Award in the International Symposium on Modeling and Optimization in Mobile, Ad hoc and Wireless Networks (WiOpt) 2015, the Feder Family Award (second prize), awarded by the Advanced Communication Center at Tel Aviv University, Tel Aviv, Israel, (2011), and President Fellowship (2008–2012) and top Honor List’s prizes in 2006, 2010, and 2011, respectively, from Bar-Ilan University.