Restless Multi-Armed Bandits under Exogenous Global Markov Process

Tomer Gafni[1], *Michal Yemini*[2], *and Kobi Cohen*[1]
[1]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]Department of Electrical and Computer Engineering, Princeton University, Princeton NJ 08540 USA
gafnito@post.bgu.ac.il, yemini.michal@gmail.com, yakovsec@bgu.ac.il

***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. The objective is an arm-selection policy that minimizes the regret, 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. We develop the Learning under Exogenous Markov Process (LEMP) algorithm, that achieves a logarithmic regret order with time, and a finite-sample bound on the regret is established. Simulation results support the theoretical study and demonstrate strong performances of LEMP.**

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

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.

This model captures a common application of dynamic spectrum access, where a network consists of a wide-band primary user and a narrow-band secondary user. The exogenous Markov process models the presence/absence of the primary user in the wide-band frequency (e.g., the Gilbert–Elliott model [1]), and the dynamics of the arms captures the quality of the different narrow-band frequencies.

As commonly adopted in RMAB problems, the objective is to select the arm that has the highest immediate expected value at each time slot under unknown arm dynamics [2]–[4]. In this paper, the value function (and thus the arm selection) depends on both the mean reward of the arms and the transition probabilities of the global Markov process. 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. Due to the exogenous process, each global state is associated with different "best" arm (i.e., the arm that maximizes the expected value given the current global state). Thus, we note that the regret is not defined with respect to the best arm on average, but with respect to a strategy tracking the best arm at each step, which is stronger.

A. 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., epoch) [5]–[7]. We thus divide the time horizon into separated exploration and exploitation phases. The goal of the exploration phase is to identify the best arm for each global state before entering the exploitation epoch. It is well known that UCB policies used to identifying the best arm require parameter tuning depending on the unobserved hardness of the task [8]. 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. As shown in [8], the hardness parameter is indeed 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. However, since the hardness parameter is unknown, existing algorithms use an upper bound on $\max_i H_i$ (e.g., [5]), which increases the order of magnitude of exploration epochs, and consequently the regret.

Our main results are summarized next. First, we develop a novel algorithm, dubbed Learning under Exogenous Markov Process (LEMP), that estimates online the appropriate hardness parameter from past observations (Sec. III-A), resulting in adaptive sizes of exploration epochs, 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 (Sec. III-B). In the exploitation epochs, LEMP dynamically chooses the best estimated arm, based on the evaluation of the global state (Sec.III-C). Interestingly, the size of the exploitation epochs is deterministic and the size of the exploration epochs is random. The rules that decide when to enter each epoch 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 (Sec. III-D). Second, we provide a rigorous theoretical analysis of LEMP. Specifically, we establish a finite sample upper bound on the 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. Finally, we provide numerical simulations that support the theoretical results.

B. Related Work

The extended RMAB model considered here is a generalization of the classic MAB problem [9]–[13]. RMAB problems have been studied under both the non- Bayesian [5]–[7], [14]–[18], and Bayesian [19]–[21], [22]–[28] settings. Under the non-Bayesian setting, special cases of Markovian dynamics have been studied in [6], [14], [16]. 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 [22]–[24], [29] under the model where a player receives a unit reward for each arm that was observed in a good state. In [25], [30], the indexability of a special classes of RMAB has been established. It is also related to models of partially observed Markov decision process (POMDP) [31], [32], with the goal of balancing between increasing the immediate reward and the benefits of improving the learning accuracy of the unknown states. Other related studies can be found in [33], [34].

The setting in this paper is also related to the non-stationary bandit problems, where distributions of rewards may change in time [3], [35]–[37]. However, the distribution that governs the non-stationary models in these studies differs from our settings, and leads to a different problem structure. Finally, [38], [39] and recently [40] considered the setting of global Markov process that governs the reward distribution. However, they addressed the linear/affine model, which is different from the RMAB formulation in this paper.

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 a finite space, irreducible, and aperiodic discrete time $\mathcal{S}$ Markov chain 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 $\mathcal{X}_s^i$ Markov chain with unknown transition matrix $P_{\mathcal{X}_s^i}$. We assume that $\mathcal{X}_{\tilde{s}}^i \bigcap \mathcal{X}_{\check{s}}^i = \emptyset$ for all $i, \tilde{s}, \check{s}$ (i.e., we can recover the global state in each time slot). We also define the stationary distribution of state x in arm i at global state s to be $\pi_s^i(x)$.

At each time, the player chooses one arm to play. Each arm, when played, offers a certain positive reward that defines the current state of the arm, $x_{s_t}^i$. 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 $P_{\mathcal{X}_s^i}$. The unknown stationary reward mean of arm i at global state s, μ_s^i, is given by:

$$\mu_s^i = \sum_{x \in \mathcal{X}_s^i} x \pi_s^i(x).$$

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

$$V_s^i \triangleq \sum_{\tilde{s} \in \mathcal{S}} p_{s\tilde{s}} \mu_{\tilde{s}}^i. \tag{1}$$

Let i_t^* be the arm with the highest expected value at time t, i.e., $i_t^* \triangleq \arg\max_i V_{s_t}^i$, and let $\phi(t) \in \{1, 2, ..., N\}$ 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 $\mathcal{N}$. The expected regret of policy ϕ is defined as:

$$\mathbb{E}_\phi[r(t)] = \mathbb{E}_\phi\left[\sum_{n=1}^{t} \sum_{i: V_{s_n}^i < V_{s_n}^*} (x_{s_n}^{i_n^*}(n) - x_{s_n}^i(n)) \mathbb{1}_{\{I_n = i\}}\right]. \tag{2}$$

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 [2] and in [3] 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, but with respect to the best arm at each step according to the global state, which is a stronger regret.

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

The LEMP algorithm divides the time horizon into exploration and exploitation phases. The 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 probabilities estimation in a closed loop manner.

A. Design Principles of LEMP

In order to ensure sufficient small regret in exploitation epochs (i.e., to reduce the probability for choosing suboptimal arms in exploitation), we should take a sufficiently large number of samples in the exploration epochs. From (1) we observe that we should estimate accurately two terms: the mean reward of each arm i in each global state s, μ_s^i, 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 $\overline{D}_s^i \log(t)$ times for being able to distinguishing it from i_s^* (i.e., the arm that maximizes the expeted value in state s) with a sufficiently high accuracy, where

$$\overline{D}_s^i \triangleq \frac{4L}{(V_s^* - V_s^i)^2}, \tag{3}$$

where $V_s^* \triangleq \max_i V_s^i$, and L is constants that depends on the system parameters, defined in (11). The $\overline{D}_s^i$ parameter is a type of hardness parameter [8], appropriate for the setting considered in this paper, in the sense that it determines 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 $\{p_{s\check{s}}\}, \{\mu_{\check{s}}^i\}$. 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:

$$\hat{\mu}_s^i(t) = \frac{1}{T_s^i(t)} \sum_{n=1}^{T_s^i(t)} x_s^i(t_s^i(n)), \quad \hat{p}_{\tilde{s}\check{s}}(t) = \frac{N_{\tilde{s}\check{s}}(t)}{N_{\tilde{s}}(t)}, \tag{4}$$

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 Sec. 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 $N_{\tilde{s}\check{s}}(t)$ is the number of transitions from $\tilde{s}$ 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 $\Delta \triangleq \min_s \Delta_s$.

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

$$\hat{D}_s^i(t) \triangleq \frac{4L}{\max\{\Delta, (\hat{V}_s^*(t) - \hat{V}_s^i(t))^2 - \epsilon\}}, \tag{5}$$

where:

$$\hat{V}_s^i(t) \triangleq \sum_{\check{s} \in \mathcal{S}} \hat{p}_{s\check{s}}(t) \hat{\mu}_{\check{s}}^i(t), \tag{6}$$

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

Using $\{\hat{D}_s^i(t)\}$, which are updated dynamically during 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. Whether we succeed to obtain a logarithmic regret order depends on how fast $\hat{D}_s^i(t)$ converges to a value which 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).

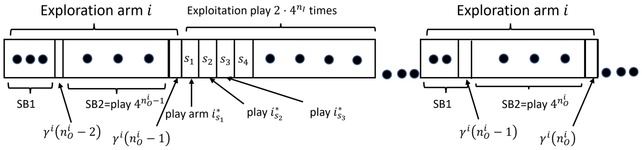

Fig. 1. An illustration of the exploration and exploitation epochs of LEMP Algorithm.

B. The structure of exploration epochs:

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., epoch). Therefore, the exploration epochs 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 epochs 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 epoch for arm i. As illustrated in Fig. 1, once the player starts the $(n_O^i)^{th}$ exploration epoch, it first plays a random period of time until observing $\gamma^i(n_O^i - 1)$ (i.e., a random hitting time). 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 epoch, and so on. We define the set of time indices during SB2 sub-blocks by $\mathcal{V}_i$.

C. The structure of exploitation epochs:

Let $n_I(t)$ be the number of exploitation epochs up to time t. The player plays the exploitation epoch 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 of each arm given the observed global state $\hat{V}_s^i(t)$ when entering the $(n_I)^{th}$ exploitation epoch, and plays the arm that maximizes the expected value.

D. Choosing between epoch types:

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

$$I_L \triangleq \frac{\bar{\lambda}_{\min}}{3072((x_{\max}+2)^2 \cdot |\mathcal{X}_{\max}| \cdot \hat{\pi}_{\max} \cdot |\mathcal{S}| \cdot (V_{\max}^* + 2))^2}, \tag{7}$$

$$I_G \triangleq \frac{1}{128((x_{\max}+2) \cdot |S| \cdot (V_{\max}^* + 2))^2}. \tag{8}$$

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:

$$T_s^i(t) \leq \max\left\{\hat{D}_s^i(t), \frac{2}{\epsilon^2 \cdot I_L}\right\} \cdot \log t, \tag{9}$$

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

$$N_s(t) \leq \frac{2}{\epsilon^2 \cdot I_G} \cdot \log t, \tag{10}$$

then the player enters an exploration epoch for arm i_M which satisfies $i_M \triangleq \arg\min_i\{\min_s \hat{D}_s^i(t)\}$. Otherwise, the player enters an exploitation epoch.

IV. Regret Analysis

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

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

$$L \triangleq \frac{1}{16(V_{\max}^* + 2)^2} \cdot \max\left\{\frac{1}{I_L}, \frac{1}{I_G}\right\}. \tag{11}$$

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

$$\begin{aligned}\mathbb{E}_\phi[r(t)] \leq x_{\max} \cdot \Bigg[\sum_{i=1}^{N} \Big(\frac{1}{3}[4(3A_i \cdot \log(t) + 1) - 1] \\ + M_{max}^i \cdot \log_4(3A_i \log(t) + 1)\Big) \\ +6N|\mathcal{S}|(\tfrac{|\mathcal{S}||\mathcal{X}_{\max}|}{\pi_{\min}} + 2|\mathcal{S}|) \max_s \pi_s \cdot \lceil \log_4(\tfrac{3}{2}t + 1) \rceil\Bigg] \\ +O(1),\end{aligned} \tag{12}$$

where

$$A_i \triangleq \begin{cases} \max\{2/I_L\ , 2/I_G\ ,\ \max\limits_s \overline{D}_{s,max}^i\}\ , & \textit{if } \forall s : i \in \mathcal{K}_s \\ \max\{2/I_L\ , 2/I_G\ ,\ 4L/\Delta\}\ , & \textit{if } \exists s : i \notin \mathcal{K}_s \end{cases}, \tag{13}$$

$\overline{D}_{s,max}^i \triangleq \dfrac{4L}{(V_s^* - V_s^i)^2 - 2\epsilon}$, *and $\mathcal{K}_s$ is defined as the set of all indices $i \in \{2, ..., N\}$ in global state s that satisfy:*

$$(V_s^* - V_s^i)^2 - 2\epsilon > \Delta_s.$$

The proof separates the regret of the exploration rounds from that of the exploitation rounds. It can be found in the extended version of this paper [41].

We next analyze the regret numerically. In Fig. 2 we simulated LEMP for $|\mathcal{S}| = 2, N = 3$. The global state models the presence of the primary user that uses the entire bandwidth by a Gilbert-Eliot model [1] 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) [42]. We compared the LEMP algorithm to an extended version of the DSEE algorithm [5], and to an algorithm that chooses in the exploitation epochs the best arm on average. The DSEE algorithm uses deterministic sequencing of exploration and exploitation epochs, however, it does not estimate the hardness parameter, and explores each arm $\Delta \cdot \log(t)$ times, which results in oversampling bad arms, to achieve the desired logarithmic regret. It can be seen that LEMP significantly outperforms the extended DSEE algorithm. Fig. 2 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. The full parameter setting and more extensive simulation results can be found in the extended version of this paper [41].

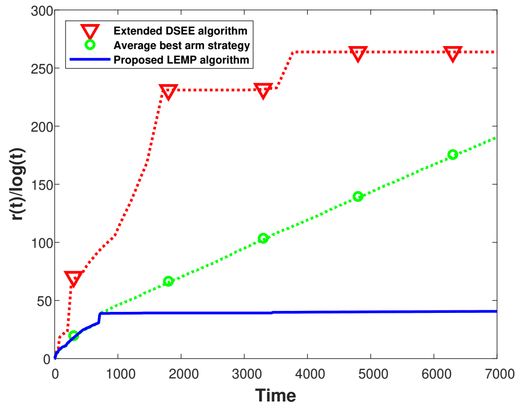

Fig. 2. Performance comparison of the regret (normalized by log t).

V. 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 epochs, LEMP estimates the hardness parameter of the problem which controls the size of exploration epochs. During the exploitation epochs, LEMP switches arms dynamically according to the global process evolution. Simulation results support the theoretical analysis, and shows superior performances of LEMP against competitive strategies.

References

[1] K. Liu and Q. Zhao, “Link throughput of multi-channel opportunistic access with limited sensing,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2997–3000, IEEE, 2008.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The nonstochastic multiarmed bandit problem,” *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[3] A. Garivier and E. Moulines, “On upper-confidence bound policies for switching bandit problems,” in *International Conference on Algorithmic Learning Theory*, pp. 174–188, Springer, 2011.

[4] S. Bagheri and A. Scaglione, “The restless multi-armed bandit formulation of the cognitive compressive sensing problem,” *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1183–1198, 2015.

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

[6] C. Tekin and M. Liu, “Online learning of rested and restless bandits,” *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[7] T. Gafni and K. Cohen, “Learning in restless multi-armed bandits via adaptive arm sequencing rules,” *IEEE Transactions on Automatic Control*, 2020.

[8] J.-Y. Audibert, S. Bubeck, and R. Munos, “Best arm identification in multi-armed bandits.,” in *COLT*, pp. 41–53, Citeseer, 2010.

[9] J. C. Gittins, “Bandit processes and dynamic allocation indices,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.

[10] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[11] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards,” *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, 1987.

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

[13] C. Tekin and M. Liu, “Online algorithms for the multi-armed bandit problem with markovian rewards,” in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1675–1682, IEEE, 2010.

[14] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, “The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2940–2943, IEEE, 2011.

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

[16] J. Oksanen and V. Koivunen, “An order optimal policy for exploiting idle spectrum in cognitive radio networks,” *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1214–1227, 2015.

[17] T. Gafni and K. Cohen, “A distributed stable strategy learning algorithm for multi-user dynamic spectrum access,” in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 347–351, IEEE, 2019.

[18] T. Gafni and K. Cohen, “Distributed learning over markovian fading channels for stable spectrum access,” *arXiv preprint arXiv:2101.11292*, 2021.

[19] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

[20] R. R. Weber and G. Weiss, “On an index policy for restless bandits,” *Journal of applied probability*, vol. 27, no. 3, pp. 637–648, 1990.

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

[22] Q. Zhao, B. Krishnamachari, and K. Liu, “On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

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

[24] S. H. A. Ahmad and M. Liu, “Multi-channel opportunistic access: A case of restless bandits with multiple plays,” in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1361–1368, IEEE, 2009.

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

[26] K. Wang and L. Chen, “On optimality of myopic policy for restless multi-armed bandit problem: An axiomatic approach,” *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 300–309, 2011.

[27] K. Wang, L. Chen, and Q. Liu, “On optimality of myopic policy for opportunistic access with nonidentical channels and imperfect sensing,” *IEEE transactions on vehicular technology*, vol. 63, no. 5, pp. 2478–2483, 2013.

[28] K. Cohen, Q. Zhao, and A. Scaglione, “Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access,” in *2014 48th Asilomar Conference on Signals, Systems and Computers*, pp. 1575–1578, IEEE, 2014.

[29] Q. Zhao and B. Krishnamachari, “Structure and optimality of myopic sensing for opportunistic spectrum access,” in *2007 IEEE International Conference on Communications*, pp. 6476–6481, IEEE, 2007.

[30] K. Liu, R. Weber, and Q. Zhao, “Indexability and whittle index for restless bandit problems involving reset processes,” in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 7690–7696, IEEE, 2011.

[31] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[32] V. Krishnamurthy and B. Wahlberg, “Partially observed markov decision process multiarmed bandits—structural results,” *Mathematics of Operations Research*, vol. 34, no. 2, pp. 287–302, 2009.

[33] V. Krishnamurthy and R. J. Evans, “Hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking,” *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 2893–2908, 2001.

[34] T. Javidi, “Information acquisition and sequential belief refinement,” in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7635–7654, IEEE, 2016.

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

[36] J. Y. Yu and S. Mannor, “Piecewise-stationary bandit problems with side observations,” in *Proceedings of the 26th annual international conference on machine learning*, pp. 1177–1184, 2009.

[37] A. Slivkins and E. Upfal, “Adapting to a changing environment: the brownian restless bandits.,” in *COLT*, pp. 343–354, 2008.

[38] S. Baltaoglu, L. Tong, and Q. Zhao, “Online learning and optimization of markov jump linear models,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2289–2293, IEEE, 2016.

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

[40] M. Yemini, A. Leshem, and A. Somekh-Baruch, “The restless hidden markov bandit with linear rewards and side information,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 1108–1123, 2021.

[41] T. Gafni, M. Yemini, and K. Cohen, “Learning in restless bandits under exogenous global markov process,” *arXiv preprint arXiv:2112.09484*, 2021.

[42] H. S. Wang and N. Moayeri, “Finite-state markov channel-a useful model for radio communication channels,” *IEEE transactions on vehicular technology*, vol. 44, no. 1, pp. 163–171, 1995.