Learning in Restless Multi-Armed Bandits using Adaptive Arm Sequencing Rules

Tomer Gafni, Kobi Cohen
Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501 Israel
Email: : {gafnito, yakovsec}@bgu.ac.il

***Abstract*—We consider a class of restless multi-armed bandit (RMAB) problems with unknown arm dynamics. At each time, a player chooses an arm out of N arms to play, referred to as an active arm, and receives a random reward from a finite set of reward states. The reward state of the active arm transits according to an unknown Markovian dynamic. The reward state of passive arms (which are not chosen to play at time t) evolves according to an arbitrary unknown random process. The objective is an arm-selection policy that minimizes the regret, defined as the reward loss with respect to a player that always plays the most rewarding arm. This class of RMAB problems has been studied recently in the context of communication networks and financial investment applications. We develop a strategy that selects arms to be played in a consecutive manner in which the selection sequencing rules are adaptively updated controlled by the current sample reward means, referred to as Adaptive Sequencing Rules (ASR) algorithm. By designing judiciously the adaptive sequencing rules of the chosen arms, we show that ASR algorithm achieves a logarithmic regret order with time and a finite-sample bound on the regret is established. Although existing methods have shown a logarithmic regret order with time in this RMAB setting, the theoretical analysis presents significant improvement in the regret scaling with respect to the system parameters under ASR. Extensive simulation results support the theoretical study and demonstrate strong performance of the algorithm as compared to existing methods.**

I. Introduction

Restless Multi-Armed Bandit (RMAB) problems are generalizations of the classic Multi-Armed Bandit (MAB) problem [1]–[3]. 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. The RMAB problem under the Bayesian formulation with known Markovian dynamics has been shown to be P-SPACE hard in general [4].

In this paper, we consider the following RMAB problem with unknown arm dynamics. At each time, a player chooses an arm out of N arms to play, referred to as an active arm. Once playing an arm, a random reward is received from a finite set of rewards. The reward state of the active arm transits according to an unknown Markovian dynamics. The reward state of passive arms (which are not chosen to play at time t) might change as well and evolve according to an arbitrary unknown random process.

This class of RMAB problems has been studied recently in the context of communication networks, and financial investment applications [5], [6]. For example, in the hierarchical opportunistic spectrum access model in cognitive radio networks, a secondary user (unlicensed) is allowed to transmit data over a channel among a set of available channels (i.e., arms) when primary (licensed) users do not transmit. The temporal spectrum usage of the primary user is modeled by a Markovian dynamics, which leads to a Markovian reward model. Thus, the secondary user aims at designing a good channel selection policy without knowing the dynamics of the primary users, with the goal of maximizing its long-term rate (i.e., accumulated reward). Other related models have studied channel selection under unknown fading dynamics and financial investments (see [5], [6] and references therein).

A. Performance Measure of Learning in RMAB

Although optimal solutions have been obtained for some special cases of RMAB models (see references in Section I-D), solving RMAB problems directly is intractable in general [4]. Thus, a widely used performance measure of an algorithm is the *regret*, defined as the reward loss with respect to a player with a side information on the model. An algorithm that achieves a sublinear scaling rate of the regret with time approaches the performance of the player with the side information as time increases. The essence of the problem is thus to design an algorithm that learns the side information effectively so as to achieve the best sublinear scaling of the regret with time.

In this paper we use the definition of regret that was introduced in [7] and used later in [5], [6] for a similar RMAB model as considered here. Specifically, the regret is defined as the reward loss of an algorithm with respect to a player that knows the expected reward of all arms and always plays the arm with the highest expected reward. It should be noted that computing the optimal policy for RMABs is P-SPACE hard even when the Markovian model is known [4]. Nevertheless, always playing the arm with the highest expected reward is known to be optimal in the classic MAB under i.i.d. or rested Markovian rewards (up to an additional constant term [3]). Thus, it is commonly used in RMAB with unknown dynamics settings for measuring the algorithm performance in a tractable manner.

B. Existing Random and Deterministic Approaches

We are facing an online learning problem with the well known exploration versus exploitation dilemma. On the one hand, a player should explore all arms in order to infer their states. On the other hand, it should exploit the information gathered so far to play the best arm. Due to the restless nature of both active and passive arms and potential reward loss due to transient effect as compared to steady state when

978-1-5386-4780-6/18/$31.00©2018 IEEE

switching arms, learning the Markovian reward statistics requires that arms will be played in a consecutive manner for a period of time (i.e., epoch). In [5], [6], regenerative cycle algorithm (RCA), and deterministic sequencing of exploration and exploitation (DSEE) algorithm, respectively, have been proposed based on these insights. The RCA algorithm chooses the active arms based on the upper confidence bound (UCB) index [8] when entering each epoch, and a logarithmic regret with time was shown. However, since RCA performs random regenerative cycles until catching predefined states at each epoch (i.e., hitting times) the scaling with the mean hitting time M (which scales at least polynomially with the state space) is of order $O(M \log t)$. The DSEE algorithm overcomes this issue by using deterministic sequencing of exploration and exploitation epochs. A logarithmic regret with time was shown under DSEE. However, applying the deterministic sequencing method by DSEE results in oversampling bad arms to achieve the desired logarithmic regret, which scales as $O\left((\frac{1}{\sqrt{\Delta}} + \frac{N-2}{\Delta}) \log t\right)$, where N is the number of arms and $0 < \Delta < (\mu_{\sigma(1)} - \mu_{\sigma(2)})^2$ is a known lower bound on the square difference between the highest reward mean $\mu_{\sigma(1)}$ and the second highest reward mean $\mu_{\sigma(2)}$. Increasing the mean hitting times (e.g., by increasing the state space, or decreasing the probability of switching between states) decreases performance under RCA. Increasing N when $(\mu_{\sigma(1)} - \mu_{\sigma(2)})$ is small as compared to the differences between $\mu_{\sigma(1)}$ and the reward means of other arms decreases performance under DSEE.

C. Main Results

1) *Algorithm development:* We propose a novel Adaptive Sequencing Rules (ASR) algorithm for solving the RMAB problem. The basic idea of ASR is to estimate online the desired (unknown) exploration rate of each arm required for efficient learning. Thus, by sampling each arm according to the desired exploration rate, ASR avoids oversampling bad arms as in DSEE, and at the same time it avoids using too frequent regenerative cycles as in RCA. Interestingly, the size of the exploitation epochs is deterministic and the size of the exploration epochs is random under ASR. The sequencing 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 in a closed-loop manner.
2) *Theoretical performance analysis:* We establish a finite-sample upper bound on the regret under the proposed ASR algorithm. Our analysis is valid for both model settings in [5], and [6]. Thus, performance comparison between the algorithms can be conducted analytically. Specifically, similar to RCA [5] and DSEE [6], we show that the proposed ASR algorithm achieves a logarithmic regret order with time as well. The scaling with the mean hitting time under ASR, however, is significantly better than the scaling under RCA ($O(M \log \log t)$ under ASR as compared to $O(M \log t)$ under RCA). The scaling with the number of arms and Δ under ASR is significantly better than the scaling under DSEE ($O\left((\frac{1}{\sqrt{\Delta}} + N - 2) \log t\right)$ under ASR as compared to $O\left((\frac{1}{\sqrt{\Delta}} + \frac{N-2}{\Delta}) \log t\right)$ under DSEE).
3) *Simulation results:* We performed extensive simulation experiments that support our theoretical results under various parameter settings. Significant performance gain of ASR over RCA and DSEE has been observed.

D. Related Work

RMAB problems have been studied in the literature under both the non-Bayesian [5], [6], [9]–[11] and Bayesian [12]–[20] settings. Under the non-Bayesian setting, special cases of Markovian dynamics have been studied in [5], [9], [11]. Under the Bayesian setting with known dynamics, the objective is exact optimality in terms of the total expected reward over time. The structure of the optimal policy for a general RMAB remains open. There are a number of studies on special classes of RMABs. In particular, the optimality of the myopic policy was shown under positively correlated two-state Markovian arms [15]–[18] under the model where a player receives a unit reward for each arm observed in a good state. In [19], [21], the indexability of a special classes of RMAB have been established. In [20], optimality conditions of a myopic policy have been established for a family of regular reward functions. In our previous work, we have derived optimality conditions of a myopic policy under arm activation constraints in the context of dynamic spectrum access [22]. Other related approaches include game theoretic, and reinforcement learning algorithms (see [23]–[27] and references therein).

II. System Model and Problem Formulation

We consider N arms indexed by $i = 1, 2, \cdots, N$. The i^{th} arm is modeled as a discrete-time, irreducible and aperiodic Markov chain with finite state space S^i. The maximal cardinality among the set spacecs is defined by: $|S_{max}| \triangleq \max_{1 \leq i \leq N}\{|S^i|\}$. 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. Let $s_i(t)$ denote the state of arm i at time t. The highest sum of rewards among arms is defined by $r_{max} \triangleq \max_{1 \leq i \leq N}\{\sum_{s \in S^i} s_i\}$. Let P^i denote the transition probability matrix and $\vec{\pi_i} = \{\pi_i(s)\}_{s \in S^i}$ the stationary distribution of arm i. The minimal stationary distribution among all arms and states is defined by $\pi_{min} \triangleq \min_{1 \leq i \leq N, s \in S^i} \ \pi_i(s)$. Let λ_i be the second largest eigenvalue of P^i, and let $\lambda_{max} \triangleq \max_{1 \leq i \leq N} \ \lambda_i$ be the maximal one among all arms. Also let $\overline{\lambda}_{min} \triangleq 1 - \lambda_{max}$, and let $\overline{\lambda_i} \triangleq 1 - \lambda_i$ be the eigenvalue gap. Let $M^i_{x,y}$ be the mean hitting time of state y starting at initial state x for arm i, and let $M^i_{max} = \max_{x,y \in S_i, x \neq y} M^i_{x,y}$. We also define:

$$\begin{aligned} A_{max} &\triangleq \max_i \, (\min_{s \in S^i} \ \pi_i(s))^{-1} \sum_{s \in S^i} s, \\ L &\triangleq \frac{30 r^2_{max}}{(3 - 2\sqrt{2})\overline{\lambda}_{min}}. \end{aligned} \tag{1}$$

We assume that the arms are restless. Specifically, the reward state of the active arm (say i) transits according to the unknown Markovian rule P^i, while the reward state of passive arms (which are not chosen to play at time t) evolves according to an arbitrary unknown random process. The stationary reward mean μ_i is given by $\mu_i = \sum\limits_{s \in S^i} s\pi_i(s)$. Let σ be a permutation of $\{1, ..., N\}$ such that

$$\mu^* \triangleq \mu_{\sigma(1)} \geq \mu_{\sigma(2)} \geq \cdots \geq \mu_{\sigma(N)}.$$

Let $t^i(n)$ denote the time index of the n^{th} play on arm i, and $T^i(t)$ denote the total number of plays on arm i by time t. Thus, the total reward by time t is given by:

$$R(t) = \sum_{i=1}^{N} \sum_{n=1}^{T^i(t)} s_i(t^i(n)). \tag{2}$$

For a policy ϕ, we define the regret $r^\phi(t)$ as the difference between the expected total reward that can be obtained by playing the arm with the highest mean, and the expected total reward obtained from using policy ϕ up to time t:

$$r_\phi(t) = t\mu_{\sigma(1)} - \mathbb{E}_\phi[R(t)]. \tag{3}$$

The objective is to minimize the growth rate of the regret with time.

III. The Adaptive Sequencing Rules (ASR) Algorithm

The basic idea of ASR is to sample each arm according to its learning rate needed for a sufficiently accurate inference. We show in the analysis that we must explore a bad arm $\sigma(i)$, $i = 2, 3, ..., N$, at least $\overline{D}_i \log t$ times for being able to distinguishing it from μ^* with a sufficiently high accuracy, where

$$\overline{D}_i \triangleq \frac{4L}{(\mu^* - \mu_{\sigma(i)})^2}. \tag{4}$$

The smaller the mean difference, the more samples we must take for exploring bad arms. *Since the reward means* $\left\{\mu_{\sigma(i)}\right\}_{i=1}^{N}$*, are unknown, however, we can estimate* $\overline{D}_i$ *by replacing* $\mu_{\sigma(i)}$ *by its sample reward mean. Using the estimate of* $\overline{D}_i$ *(which is updated dynamically during time and controlled by the sample means), we can design an adaptive sequencing rule for sampling arm i that will converge to its learning rate, required for obtaining a sufficiently accurate inference as time increases.* Whether we succeed to obtain a logarithmic regret order depends on how fast the estimate of $\overline{D}_i$ converges to a value which is no smaller than $\overline{D}_i$ (so we take at least $\overline{D}_i$ samples from bad arms in most of the times). To guarantee the desired convergence speed, we judiciously overestimate $\overline{D}_i$ as detailed in Section III-D.

A. Playing arms consecutively during exploration and exploitation epochs:

As discussed in I-B, learning the Markovian reward statistics requires that arms will be played in a consecutive manner for a period of time. For instance, RCA selects arms based on UCB and plays the arm a random period of time which depends on hitting time events. On the other hand, DSEE samples arms a deterministic periods of time that grow geometrically with time. Interestingly, we show that by judiciously combining these two sampling methods, while determining the exploration frequency for each arm according to its adaptive sequencing rule (described in Sec. III-D), we can achieve tremendous improvement in both theoretical and simulation performance as shown in Section IV.

Specifically, we divide the time horizon into exploration and exploitation epochs, as illustrated in Fig. 1. An exploration epoch is dedicated to play a certain arm determined by its adaptive sequencing rule (described in Sections III-D, III-E). Let $n_O^i(t)$ be the number of exploration epochs in which arm i was played up to time t. An exploitation epoch is dedicated to play the arm with the highest sample mean, whenever exploration is not being performed. Let $n_I(t)$ be the number of exploitation epochs up to time t. In Fig. 1, we illustrate the exploration epochs for arm i only, for the ease of illustration. In general, an interleaving of exploration epochs for all arms with exploitation epochs (for the best arm) is performed.

B. The structure of exploration epoch:

The exploration epochs for each arm are divided into two sub-bloks: a random-size sub-block SB1, and a deterministic-size sub-block SB2. Consider time t (and we remove the time index t for convenience). Let $\gamma^i(n_O^i - 1)$ be the last reward state observed at the $(n_O^i - 1)^{th}$ exploration epoch. 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 a deterministic period of time with length of $4^{n_O^i}$. This deterministic period of time is referred to as SB2. The player stores the last reward 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 epoch:

Let $\overline{s}_i$ be the sample reward mean of arm i when entering the $(n_I)^{th}$ exploitation epoch. Then, the player plays the arm with the highest sample mean $\max_i \overline{s}_i$ for a deterministic period of time with length $2 \cdot 4^{n_I - 1}$ (there are no arm switchings inside epochs). We define the set of time indices in exploitation epochs by $\mathcal{W}_i$. Computing sample mean $\overline{s}_i$ for each arm is based on observations taken from $\mathcal{V}_i$ and $\mathcal{W}_i$. Observations from SB1 sub-blocks are removed to ensure consistency of the estimators.

D. The Selection rule (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 for the arm with the highest sample mean. Let $\widetilde{s}_i(t)$ be the sample reward mean of arm i, computed based on observations taken from $\mathcal{V}_i$ only at time t (see discussion and detailed analysis in [28]). Let

$$\widehat{D}_i(t) \triangleq \frac{4L}{\max\left\{\Delta, (\max_j \widetilde{s}_j(t) - \widetilde{s}_i(t))^2 - \epsilon\right\}}, \tag{5}$$

where $0 < \Delta < (\mu_{\sigma(1)} - \mu_{\sigma(2)})^2$ is a known lower bound on the square difference $(\mu_{\sigma(1)} - \mu_{\sigma(2)})^2$, and $\epsilon > 0$ is a fixed tuning parameter (a discussion on the implementation is given in Sec. III-E). We also define:

$$I \triangleq \frac{\epsilon^2 \cdot \overline{\lambda}_{min}}{192(r_{max} + 1)^2}. \tag{6}$$

The design of the selection rule is based on the following insights. First, we need to make sure that the algorithm takes at least $\overline{D}_i \log t$ samples from each bad arm ($\overline{D}_i$ is given in (4)) for computing a sufficiently accurate sample means $\overline{s}_i$.

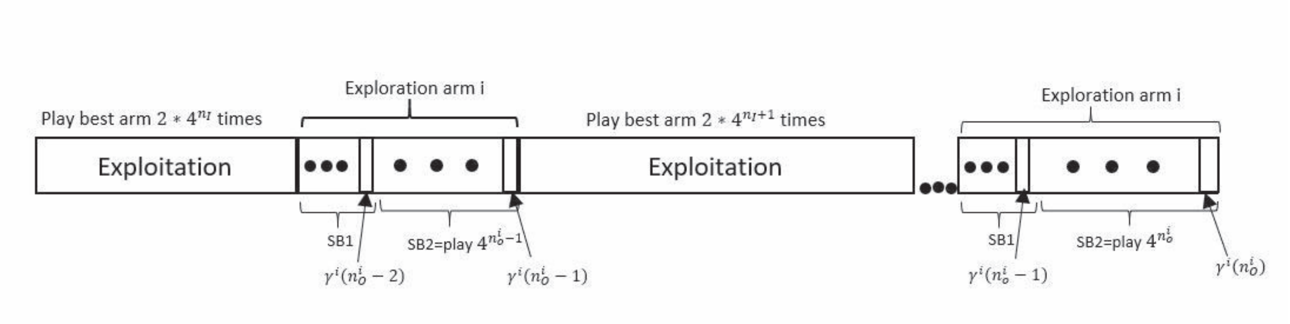

Fig. 1. An illustration of the exploration and exploitation epochs under ASR. As explained in Sec. III-C, during exploitation epoch the player plays the same arm that had the highest sample mean in the beginning of the epoch. As explained in Sec. III-B, an exploration epoch is divided into a random-size sub-block SB1 and a deterministic (geometrically growing) size sub-block SB2. SB1 of an arm (say i as in the figure) is a random hitting time until catching the last state γ^i that arm i observed in the previous exploration epoch. Selecting which epoch to play is determined by the selection rule described in Sec. III-D.

Therefore, the algorithm replaces the unknown value $\overline{D}_i$ by $\widehat{D}_i(t)$. $\widehat{D}_i(t)$ overestimates $\overline{D}_i$ to obtain the desired property. Second, since $\widehat{D}_i(t)$ is a random variable, we need to make sure that the desired property holds with a sufficiently high probability. I can be viewed as the minimal rate function of the estimates among all arms and used to guarantee the desired property. Consider a beginning of each epoch at time t, and let $\mathcal{V}_i(t)$ be the set of all time indices during SB2 sub-blocks up to time t. Then, if there exists an arm (say i) such that the following condition holds:

$$|\mathcal{V}_i(t)| \leq \max\left\{\widehat{D}_i(t), \frac{2}{I}\right\} \cdot \log t, \qquad (7)$$

then the player enters an exploration epoch for arm i (ties between arms are broken arbitrarily). Otherwise, it enters an exploitation epoch. As a result, the selection rule for each arm that governs the arm sequencing policy is adaptive in the sense that it is updated dynamically with time and controlled by the random sample mean in a closed loop manner.

E. High-level pseudocode and implementation of ASR:

In summary, the player performs the following algorithm:
1) (Initialization:) For all N arms, execute an exploration epoch where a single observation is taken from each arm.
2) If condition (7) holds for some arm (say i), then execute an exploration epoch for arm i (as described in Sec. III-B) and when finishing go to Step 2 again. Otherwise, go to Step 3.
3) Execute an exploitation epoch (as described in Sec. III-C). When finishing, go to Step 2.

We next discuss technical implementation details when executing ASR algorithm. (i) From a theoretical perspective, ASR and DSEE requires the same knowledge on the system parameters to guarantee the theoretical performance. RCA requires the same parameters, excepts that Δ is not needed; (ii) 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 analysis) and practical conditions for obtaining good performance. For example, in [6] the authors simulated DSEE with exploration rate $10 \cdot \log t$ while the theoretical sufficient conditions were $\approx 1,000 \cdot \log t$. A similar gap was observed in RCA. Indeed, this is the case in ASR as well. While analysis requires to overestimate $\overline{D}_i$ as in (5), simulation results provide much better performance when estimating $\overline{D}_i$ directly by setting $\widehat{D}_i(t) \leftarrow \frac{4L}{(\max_j \widetilde{s}_j(t) - \widetilde{s}_i(t))^2}$. Thus, in practice Δ is not needed and the parameters can be estimated on the fly. In Fig. 2, we simulated exactly the same parameters that the authors pick and tuned in [6, Figure 4]. We indeed obtained the same curves for DSEE and RCA. *We then executed ASR without the knowledge of Δ and without tuning ϵ (set to zero). $\widehat{D}_i(t)$ was estimated on the fly.* It can be seen that ASR significantly outperforms both DSEE and RCA. A more extensive empirical study that demonstrates the efficiency of ASR can be found in [28].

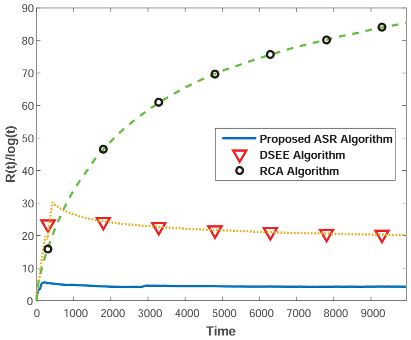

Fig. 2. The regret (normalized by $\log t$) for RMAB with 5 arms under ASR, DSEE, and RCA as a function of time.

IV. Theoretical Regret Analysis

In the following theorem we establish a finite-sample bound on the regret with time. The proof can be found in the extended version of this paper [28].

Theorem 1: Assume that the proposed ASR algorithm is implemented and the assumptions on the system model described in Section II hold. Then, the regret at time t is upper bounded by:

$$\begin{aligned} r(t) &\leq C_1 \cdot \log_4(t) + C_2 \cdot \log(t) \\ &+ \sum_{i=2}^{N} \left(\mu_{\sigma(1)} - \mu_{\sigma(i)}\right) M_{max}^i \cdot \log_4(\log(t)) + O(1), \end{aligned} \qquad (8)$$

where

$$\begin{aligned} C_1 &= A_{max} + 3\sum_{i=2}^{N} \frac{\mu_{\sigma(1)}-\mu_{\sigma(i)}}{\pi_{min}} \times \\ &\qquad \sum_{k=1,i} \left(\frac{1}{\log(2)} + \frac{\sqrt{2\lambda_k}\sqrt{L}}{10 \sum\limits_{s\in S_k} s} |S_k| \right), \\ C_2 &= 4\sum_{i=2}^{N} \Big[\mathbf{1}_{\mathcal{K}}(i) \max\Big\{ (\mu_{\sigma(1)} - \mu_{\sigma(i)})\frac{2}{I}\ , \\ &\qquad \frac{4L}{(\mu_{\sigma(1)}-\mu_{\sigma(i)})+\sqrt{2\epsilon}} + \frac{4L\cdot\sqrt{2\epsilon}}{(\mu_{\sigma(1)}-\mu_{\sigma(i)})^2-2\epsilon} \Big\} \\ &\qquad + \mathbf{1}_{\mathcal{K}^C}(i) \left(\mu_{\sigma(1)} - \mu_{\sigma(i)}\right) \max\left\{ \frac{2}{I}, \frac{4L}{\Delta} \right\} \Big]\ , \end{aligned} \tag{9}$$

where $\mathcal{K}$ is defined as the set of all indices $i \in \{2, ..., N\}$ that satisfy: $(\mu_{\sigma(1)} - \mu_{\sigma(i)})^2 - 2\epsilon > (\mu_{\sigma(1)} - \mu_{\sigma(2)})^2$, and $\mathbf{1}_{\mathcal{K}}(i)$ is the indicator function on the set $\mathcal{K}$, i.e., $\mathbf{1}_{\mathcal{K}}(i) = 1$ if $i \in \mathcal{K}$ and $\mathbf{1}_{\mathcal{K}}(i) = 0$ otherwise. $\mathcal{K}^C$ is the complementary set of $\mathcal{K}$.

Theoretical comparison with RCA and DSEE: The theorem shows that similar to RCA [10] and DSEE [6], the regret under ASR has logarithmic order with time. The scaling with the state space under ASR, however, is significantly better than the scaling under RCA. Since RCA performs a random regenerative cycles until catching predefined states in each epoch, the scaling with the mean hitting time (which scales polynomially with the state space) is $O(\sum_i M_{max}^i \log t)$. On the other hand, ASR scales only with $O(\sum_i M_{max}^i \log\log t)$. The scaling with N and Δ under ASR is significantly better than the scaling under DSEE. Specifically, DSEE scales with $O\left((\frac{1}{\sqrt{\Delta}} + \frac{N-2}{\Delta})\log t\right)$, whereas ASR scales only with $O\left((\frac{1}{\sqrt{\Delta}} + N - 2)\log t\right)$ since the adaptive sequencing rules estimate the desired learning rate for each arm.

V. Conclusion

Inspired by recent developments of sequencing methods of exploration and exploitation epochs, we develop a novel algorithm that introduces the concept of adaptive sequencing rules for arm selection in RMAB problems. The arm selection rule is adaptive in the sense that it estimates the required learning rate of each arm and updated dynamically with time, controlled by the random sample mean in a closed loop manner. Significant performance gain over RCA and DSEE has been analyzed theoretically and numerically.

VI. Acknowledgement

This research was supported by the Cyber Security Research Center at Ben-Gurion University of the Negev.

References

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 148–177, 1979.

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

[3] 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.

[4] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[5] 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.

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

[7] 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.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

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

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

[11] 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.

[12] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, pp. 287–298, 1988.

[13] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, pp. 637–648, 1990.

[14] N. Ehsan and M. Liu, "On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services," in *In Proc. IEEE Conf. Computer and Communications*, vol. 3, pp. 1974–1983, 2004.

[15] Q. Zhao and B. Krishnamachari, "Structure and optimality of myopic sensing for opportunistic spectrum access," in *IEEE International Conference on Communications (ICC)*, pp. 6476–6481, 2007.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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, 2012.

[21] K. Liu, R. Weber, and Q. Zhao, "Indexability and whittle index for restless bandit problems involving reset processes," in *IEEE Conference on Decision and Control (CDC)*, pp. 7690–7696, Dec. 2011.

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

[23] K. Cohen, A. Leshem, and E. Zehavi, "Game theoretic aspects of the multi-channel aloha protocol in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2276–2288, 2013.

[24] K. Cohen and A. Leshem, "Distributed game-theoretic optimization and management of multichannel aloha networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1718–1731, 2016.

[25] K. Cohen, A. Nedić, and R. Srikant, "Distributed learning algorithms for spectrum sharing in spatial random access wireless networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2854–2869, 2017.

[26] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks," *in Proc. of the IEEE Global Communications Conference (GLOBECOM), arXiv preprint arXiv:1704.02613*, 2017.

[27] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *submitted to IEEE Transactions on Wireless Communications, 2018, arXiv preprint arXiv:1704.02613.*

[28] T. Gafni and K. Cohen, "Learning in restless multi-armed bandits via adaptive arm sequencing rules," *submitted to IEEE Transactions on Automatic Control, 2018, available at arXiv.*