Stochastic Multi-Player Multi-Armed Bandits with Multiple Plays for Uncoordinated Spectrum Access

Marie-Joséphа Youssef, Venugopal V. Veeravalli, Joumana Farah, Charbel Abdel Nour

Stochastic Multi-Player Multi-Armed Bandits with Multiple Plays for Uncoordinated Spectrum Access

Marie-Josepha Youssef[(1)], Venugopal V. Veeravalli[(2)], Joumana Farah[(3)] and Charbel Abdel Nour[(1)]
[(1)]IMT Atlantique, Department of Electronics, Lab-STICC - UMR 6285, Technopôle Brest Iroise, CS 83 818 - 29238 Brest Cedex, France
[(2)]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
[(3)]Department of Electricity and Electronics, Faculty of Engineering, Lebanese University, Roumieh, Lebanon

***Abstract*—In this paper, an algorithm based on the multi-player multi-armed bandit (MAB) framework is proposed to solve an uncoordinated spectrum access problem. The proposed technique does not require any communication or coordination between users. The case of varying channel rewards across users is considered. In contrast to previous work, the users are permitted to choose multiple channels for transmission, resulting in a MAB model with multiple plays. The proposed algorithm has an expected regret of the order $\mathcal{O}(\log^2 T)$, which is validated by simulation results.**

***Index Terms*—uncoordinated spectrum access, multi-armed bandits with multiple plays, varying reward distribution.**

I. Introduction

With the introduction of the Internet-of-Things (IoT), wireless communication networks are faced with an increasingly growing number of machine-type-devices (MTD) [1]. MTD applications (e.g., smart meters, e-health, etc.) generally result in mobile traffic that mostly relies on the uplink transmission of short packet messages. Compared to the small packet sizes of useful information, the signaling overhead resulting from acquiring the channel state information (CSI) at MTDs and sending scheduling requests to a central unit is large. Therefore, optimizing the uplink scheduling of MTDs for efficient spectrum use is of utmost importance.

Uncoordinated spectrum access has received significant interest in recent literature [2]–[8]. In a system relying on uncoordinated spectrum access, a user chooses a channel and transmits whenever it has data to send, without formulating a scheduling request to the receiver or control unit. By doing so, the handshake required to receive a scheduling slot is avoided, reducing the transmission latency and signaling overhead. Because of the distributed nature and the lack of coordination in the resulting spectrum access, the problem of collisions arises. A collision occurs when multiple users choose the same channel for transmission. In this case, the receiver cannot distinguish between the collided messages, possibly leading to a failed reception of all colliding transmissions. To avoid collisions, two techniques can be leveraged. One is to apply novel multiple access techniques, such as non-orthogonal multiple access (NOMA) [5]. The other is to exploit learning algorithms so users can distributively adjust their transmissions, minimizing collisions. This work focuses on the latter technique, leaving the study of the combination of both solutions for future work.

The use of reinforcement learning (RL) for uncoordinated spectrum access has recently garnered some attention. To maximize the accumulated data rate and the number of successful transmissions, [3] and [4] adopt Q-learning, while [5] considers a NOMA system and applies deep RL. The related framework of multi-player multi-armed bandits (MAB) [9] has also been widely used to study uncoordinated spectrum access. In this framework, users learn how to dynamically adjust their transmissions on the available channels to optimize system performance. In [10] and [11], the MAB model is used to study the opportunistic spectrum access problem in cognitive radio networks where secondary users compete to access the part of the spectrum not occupied by primary users. In contrast, in [6] and [7], the MAB is employed to study the uncoordinated spectrum access problem without user distinction.

Most of the previous MAB related studies assumed that the reward distribution for each channel is the same across users [6], [10], [12]. However, in wireless networks, this assumption may not hold. Recently, [7], [8], [13] introduced new algorithms for the case of distinct reward distributions across users. In [13], forced collisions are used, allowing users to implicitly communicate the estimated means with each other to reach the optimal matching. In [7], [8], a game-theoretic approach is proposed, based on the dynamics introduced in [14]. However, all previous work employing MAB for uncoordinated spectrum access has restricted each user to transmit on one channel. A few studies [15], [16] consider the MAB problem with multiple arm plays, but are restricted to the single-player case. In wireless networks, allowing users to choose multiple channels for transmission enhances the achieved performance.

In this work, we consider the uplink of an uncoordinated spectrum access wireless system, where users aim to dynamically adjust their transmissions, without communicating with each other, to avoid collisions and optimize system performance. In addition to considering varying channel rewards between users, we study the case where each user chooses multiple channels at each timeslot. An algorithm, extending that of [7], [8] and achieving an expected regret of $\mathcal{O}(\log^2 T)$, is proposed. To the best of our knowledge, this is the first work that studies the problem of multi-player MAB with multiple plays and varying channel rewards across users.

The rest of this paper is organized as follows. Section II presents the system model. In sections III, IV and V, the proposed algorithm is presented along with an analysis of the system-wide regret. Simulation results are presented in section VI and conclusions in section VII.

978-1-7281-4490-0/20/$31.00

II. System Model

Consider the uplink of a cellular system where K users aim to communicate either with a base station (BS) or with each other in an uncoordinated manner. The communication occurs over a finite time-horizon T, which is an integer number of timeslots that may not be known in advance to the users. At each timeslot t, every user k chooses one or multiple channels to transmit its data bits. If two or more users choose the same channel for communication, the messages of all involved users collide and none of the transmissions succeed, i.e., all users achieve zero rate. We assume that each user k chooses N channels at each timeslot. Furthermore, we assume that system bandwidth is partitioned into M channels, such that $M \geq KN$. Hence, a stable assignment of users and channels, in which no two users collide, exists. It is assumed that users do not *a priori* know the rewards they get over each channel. Moreover, these rewards are distinct for each user. We model this problem of uncoordinated spectrum access as a stochastic multi-player multi-armed bandit problem with multiple plays. In this setting, the set of players is the set of users $\mathcal{K} = \{1, \ldots, K\}$, and the set of arms is the set of channels $\mathcal{M} = \{1, \ldots, M\}$. At each timeslot, every user is allowed to pull N channels. To this end, the action of user k at timeslot t is $\boldsymbol{a}_k^{(t)} \in \{0,1\}^{M\times 1}$ such that $a_k^{(t)}(m) = 1$ if user k pulls channel m at timeslot t. Moreover, $\sum_{m=1}^{M} a_k^{(t)}(m) = N,\ \forall k \in \mathcal{K}$. The action space of each user k is $\mathcal{A}_k$ and consists of all possible combinations of N channels, hence $|\mathcal{A}_k| = \binom{M}{N}$. Let $\boldsymbol{a}^{(t)}$ denote the strategy profile of all users at timeslot t, i.e., $\boldsymbol{a}^{(t)} = \{\boldsymbol{a}_1^{(t)}, \ldots, \boldsymbol{a}_K^{(t)}\}$. Also, let $\mathcal{A} = \prod_{k=1}^{K} \mathcal{A}_k$ be the action space of all users. Upon choosing an action $\boldsymbol{a}_k^{(t)}$, user k receives the following reward:

$$g_k^t(\boldsymbol{a}^{(t)}) = \sum_{m=1}^{M} a_k^{(t)}(m)\mu(k,m)\eta_m(\boldsymbol{a}^{(t)}), \tag{1}$$

where $\mu(k,m)$ is the mean reward of user k over channel m and $\eta_m(\boldsymbol{a}^{(t)}) = 1$ if no collision occurred on m when $\boldsymbol{a}^{(t)}$ is played. When a collision takes place, $\eta_m(\boldsymbol{a}^{(t)}) = 0$.

The mean reward of user k over channel m reflects the channel gain of k over m and is an indicator of the rate user k achieves over channel m. We assume that $\mu(k,m) \in [0,1]$ leading to $g_k^t(\boldsymbol{a}^{(t)}) \in [0,N]$.

Users are assumed to make their decisions in a totally distributed manner. In other words, each user k does not observe neither the channels chosen by other users nor their received rewards. Each user k can only observe the reward it gets on its chosen channels. We assume that, in addition to observing the total achieved reward on its chosen channels, user k can observe the reward achieved over each chosen channel. Moreover, users are assumed to know whether or not a collision has occurred with probability one.

Our aim is to propose a distributed algorithm that allows users to settle on channels, without coordination, in such a way as to maximize the system sum reward. Since $M \geq KN$, the optimal allocation must not involve any collisions. In fact, in the case of a collision, each colliding user can choose some other channel, without collision, and receive a non-zero reward. Let $\boldsymbol{a}^*$ be the action profile that yields the highest sum reward:

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}\in\mathcal{A}}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\mu(k,m)\eta_m(\boldsymbol{a}). \tag{2}$$

The expected regret incurred during the time horizon T is:

$$\bar{R} = T\sum_{k,m} a_k^*(m)\mu(k,m) - \mathbb{E}\left(\sum_{t,k,m} a_k^{(t)}(m)\mu(k,m)\eta_m(\boldsymbol{a}^{(t)})\right). \tag{3}$$

III. Proposed Algorithm

In all prior work on uncoordinated spectrum access based on MABs, each user is assumed to choose one channel at each timeslot. However, relaxing this assumption would result in a better performance for the users if an adequate algorithm is formulated. Indeed, when a user can access multiple channels simultaneously, both the probability of a successful transmission and the achieved reward or rate increase. The considered setting is closest to the ones considered in [7] and [8], where a game-theoretic approach to find the assignment of users to channels maximizing the user sum rewards is used. In both of these works, each user chooses one channel at each timeslot. In this work, we adapt the algorithms of both of these papers to the case of a user pulling N channels at each timeslot. The proposed technique is shown in Algorithm 1.

Since the time horizon T is not known in advance, Algorithm 1 proceeds in epochs, with each epoch consisting of an integer number of timeslots. In each epoch, three phases, namely, exploration, matching and exploitation, take place. The exploration phase aims at estimating the previously unknown means of each channel. During this phase, each user uniformly accesses one channel at each timeslot to get channel estimates. This phase runs for a constant number of timeslots given by T_0, the value of which is determined in section IV. Upon termination, all users have an estimation $\hat{\boldsymbol{\mu}}$ of the channel means. These estimated means are used in the second phase of the algorithm where users play a non-cooperative game with the aim of maximizing the achieved sum rewards. In other words, when user k chooses a channel m, if the received reward is non-zero, user k assumes that this reward is equal to the estimated one $\hat{\mu}(k,m)$. The dynamics of this matching phase, adopted from [14], are described in section III-A. The matching phase runs for $c_1 l^{1+\delta}$ timeslots, l being the epoch number. The third and final phase is an exploitation phase in which the users settle on the channels that resulted in the best performance in the previous matching phase. The exploitation phase runs for $c_2 2^l$ timeslots. Note that the duration of the matching and exploitation phases grows with the epoch number, whereas the exploration phase is performed for a constant number of timeslots in each epoch.

A. Matching Dynamics

Each user k is associated with a state $[\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, S_k]$. Following the terminology of [14] regarding the state of each user,

Algorithm 1

Initialization: Set $\hat{\mu}(k,m) = 0, V_k^t(m) = 0, y_k^t(m) = 0,\ \forall k \in \mathcal{K}, m \in \mathcal{M}$. Let $\epsilon > 0$ and $c > KN$.

for $l = 1, \ldots, L_T$ **do**

1- Exploration Phase:

  for $t = 1 : T_0$ **do**

    Choose one channel $m \in \mathcal{M}$ uniformly.

    Receive reward $x_k^t(m)$.

    if $x_k^t(m) > 0$ (i.e., no collision on channel m) **then**

      $V_k^t(m) = V_k^{t-1}(m) + 1,\ y_k^t(m) = y_k^{t-1}(m) + x_k^t(m)$. // $V_k^t(m)$ is a counter of the number of times m was accessed by k without collision until timeslot t. $y_k^t(m)$ is the sum reward achieved by k over m until timeslot t.

    end if

  end for

  Estimate means: $\hat{\mu}(k,m) = y_k^t(m)/V_k^t(m),\ \forall m \in \mathcal{M}$.

end for

2- Matching Phase: for the next $c_1 l^{1+\delta}$ timeslots, play according to the dynamics described in section III-A.

If $S_k^t = C$, choose the action to play according to (5). If $S_k^t = D$, choose the action according to (6).

If the received reward for some chosen channel is 0, the user becomes discontent as per (8).

If $\boldsymbol{a}_k \neq \bar{\boldsymbol{a}}_k$ or $\boldsymbol{u}_k \neq \bar{\boldsymbol{u}}_k$ or player k is discontent, the state transition happens according to (9).

Each user k keeps a counter of the number of times each action $\boldsymbol{a}_k$ was played and resulted in it being content:

$$F_k^l(\boldsymbol{a}_k) = \sum_{t=T_0+1}^{T_0+c_2 l^{1+\delta}} \mathbb{I}\left(\boldsymbol{a}_k^{(t)} = \boldsymbol{a}_k, S_k^t = C\right), \tag{4}$$

$\mathbb{I}$ being the indicator function.

3- Exploitation phase: for $c_2 2^l$ timeslots:

Play the action $\boldsymbol{a}_k^{(l)*}$ satisfying: $\boldsymbol{a}_k^{(l)*} = \underset{\boldsymbol{a}_k \in \mathcal{A}_k}{\operatorname{argmax}} F_k^l(\boldsymbol{a}_k)$.

$\bar{\boldsymbol{a}}_k \in \{0,1\}^{M\times 1}$ is the baseline action of user k, satisfying $\sum_{m=1}^{M} \bar{a}_k(m) = N$, $\bar{\boldsymbol{u}}_k$ is the baseline utility of user k satisfying $|\bar{\boldsymbol{u}}_k| = N$. $S_k \in \{C, D\}$ is the mood of user k reflecting whether k is content or discontent with the current action and utility. Denote $u_{k,\max} = \underset{\boldsymbol{a}_k}{\operatorname{argmax}} \sum_{m=1}^{M} a_k(m)\hat{\mu}(k,m)$. That is, $u_{k,\max}$ is the highest reward achievable by user k, resulting from playing the N channels having the highest means, without experiencing collision on any of these channels. At each timeslot t during the matching phase, user k adheres to the following dynamics to decide on the chosen action:

- A content user plays its baseline action with high probability:

$$p_k^{\boldsymbol{a}_k} = \begin{cases} \frac{\epsilon^c}{\binom{M}{N}-1}, & \text{if } \boldsymbol{a}_k \neq \bar{\boldsymbol{a}}_k, \\ 1-\epsilon^c, & \text{if } \boldsymbol{a}_k = \bar{\boldsymbol{a}}_k, \end{cases} \tag{5}$$

  where $\epsilon > 0$ is a small perturbation and c is a constant satisfying $c \geq KN$ [14].
- A discontent user chooses its action uniformly at random:

$$p_k^{\boldsymbol{a}_k} = \frac{1}{\binom{M}{N}}, \quad \forall\ \boldsymbol{a}_k \in \mathcal{A}_k. \tag{6}$$

After deciding on the action and observing the reward, the state transition of each user k occurs according to:

- If $\boldsymbol{a}_k = \bar{\boldsymbol{a}}_k$ and $\boldsymbol{u}_k = \bar{\boldsymbol{u}}_k$, then a content player remains content with probability one:

$$[\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, C] \to [\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, C]. \tag{7}$$

- If $u_{k,n} = 0$ for some $n \in [N]$, then player k becomes discontent with probability one:

$$[\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, C/D] \to [\boldsymbol{a}_k, \boldsymbol{u}_k, D]. \tag{8}$$

- If $\boldsymbol{a}_k \neq \bar{\boldsymbol{a}}_k$ or $\boldsymbol{u}_k \neq \bar{\boldsymbol{u}}_k$ or player k is discontent, then the state transitions occur according to:

$$[\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{u}}_k, C/D] \to \begin{cases} [\boldsymbol{a}_k, \boldsymbol{u}_k, C] & \text{w.p. } \epsilon^{u_{k,\max} - \sum_{n=1}^{N} u_{k,n}}, \\ [\boldsymbol{a}_k, \boldsymbol{u}_k, D] & \text{w.p. } 1 - \epsilon^{u_{k,\max} - \sum_{n=1}^{N} u_{k,n}}. \end{cases} \tag{9}$$

B. Regret Analysis

The time horizon T can be lower bounded by [8]:

$$T \geq \sum_{l=1}^{L_T - 1} (T_0 + c_1 l^{1+\delta} + c_2 2^l) \geq c_2(2^{L_T} - 2), \tag{10}$$

where L_T is the total number of epochs occurring within the time horizon T. Hence, L_T is upper bounded by:

$$L_T \leq \log\left(T/c_2 + 2\right). \tag{11}$$

1) Regret in the Exploration Phase: In the exploration phase, each user samples channels uniformly to get estimates of their means. Even though the purpose of this work is to assign to each user N channels at each timeslot, the number of channels sampled by each user at a timeslot is one in the exploration phase. It is chosen this way to decrease the number of collisions, which increases the number of samples with a non-zero reward of each channel, leading to better estimates.

The expected regret in the exploration phase R_1 in all epochs can be upper bounded by:

$$R_1 \leq \sum_{l=1}^{L_T} KNT_0 \leq KNT_0 \log\left(T/c_2 + 2\right). \tag{12}$$

2) Regret in the Matching Phase: The expected regret in the matching phase R_2 can be upper bounded by:

$$R_2 \leq \sum_{l=1}^{L_T} KNc_1 l^{1+\delta} \leq KNc_1 L_T^{2+\delta} \leq KNc_1 \log^{2+\delta}\left(T/c_2 + 2\right). \tag{13}$$

3) Regret in the Exploitation Phase: In the exploitation phase of epoch l, each user k plays the most played action resulting in content behavior in the matching phase of epoch l. The exploitation phase fails in two cases:

1) The exploration phase of epoch l fails, meaning that the mean estimates $\hat{\boldsymbol{\mu}}$ differ significantly from the actual means $\boldsymbol{\mu}$. This happens with a probability $\leq 4M^2 e^{-l}$ as shown in Lemma 2.

2) The most played action of the matching epoch differs from the optimal action. This happens with a probability $\leq A_1 e^{-l^{1+\delta}}$ as shown in Lemma 4.

The expected regret in the exploitation phase can be upper bounded by:

$$\begin{aligned} R_3 &\leq \sum_{l=1}^{L_T} KNc_2 2^l (4M^2 e^{-l} + A_1 e^{-l^{1+\delta}}) \\ &\leq \frac{2KNc_2(4M^2+A_1)}{e-2} = A_3. \end{aligned} \tag{14}$$

4) Regret of the Proposed Technique:

Theorem 1. The expected regret of the proposed technique can be upper bounded as:

$$R \leq R_1 + R_2 + R_3 = \mathcal{O}\left(\log^{2+\delta}(T)\right). \tag{15}$$

IV. Exploration Phase

The exploration phase is performed so users can learn estimates of the mean channel rewards. Since the estimation may not be always perfect, the optimal assignment with the estimated means $\hat{\boldsymbol{\mu}}$ might differ from the optimal assignment calculated from the correct means $\boldsymbol{\mu}$. However, if the estimation inaccuracy is kept small as in [7] and [8], the optimal assignment does not change due to the estimation inaccuracy.

Lemma 1. Let J_1 and J_2 be the sum reward achieved by the optimal and the second best assignments, and let $\Delta = \frac{J_1 - J_2}{2KN}$. If the difference between the estimated and the correct channel reward means satisfies:

$$|\mu(k,m) - \hat{\mu}(k,m)| < \Delta, \ \forall k \in \mathcal{K}, m \in \mathcal{M}, \tag{16}$$

then the optimal assignment does not change due to the estimation inaccuracy. In other words:

$$\begin{aligned} &\underset{\boldsymbol{a}\in\mathcal{A}}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\mu(k,m)\eta_m(\boldsymbol{a}) = \\ &\underset{\boldsymbol{a}\in\mathcal{A}}{\operatorname{argmax}} \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\hat{\mu}(k,m)\eta_m(\boldsymbol{a}). \end{aligned} \tag{17}$$

Proof. Denote the optimal assignment by $\boldsymbol{a}^{(1)}$ and the sum rewards achieved when $\boldsymbol{a}^{(1)}$ is played by J_1. Furthermore, denote the second best assignment and the sum reward achieved under it by $\boldsymbol{a}^{(2)}$ and J_2 respectively. Note that an optimal assignment does not have any collisions since $M \geq KN$. Hence, $J_1 = \sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\mu(k,m)$. Let the estimated mean of user k over channel m be written as:

$$\hat{\mu}(k,m) = \mu(k,m) + z(k,m), \tag{18}$$

where $z(k,m)$ is the estimation inaccuracy satisfying $|z(k,m)| \leq \Delta$. The sum reward achieved when $\boldsymbol{a}^{(1)}$ is played with the estimated channel means satisfies:

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\hat{\mu}(k,m) = \sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)(\mu(k,m)+z(k,m)) \\ &> \sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\mu(k,m) - \Delta KN. \end{aligned} \tag{19}$$

Any other assignment $\boldsymbol{a} \neq \boldsymbol{a}^{(1)} \neq \boldsymbol{a}^{(2)}$ must perform at most as well as $\boldsymbol{a}^{(2)}$:

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\hat{\mu}(k,m)\eta_m(\boldsymbol{a}) = \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)(\mu(k,m)+ \\ &z(k,m))\eta_m(\boldsymbol{a}) < \sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(2)}(m)\mu(k,m)\eta_m(\boldsymbol{a}^{(2)}) + \Delta KN. \end{aligned} \tag{20}$$

To avoid changing the optimal assignment because of the estimation inaccuracy, the following must hold $\forall \boldsymbol{a} \neq \boldsymbol{a}^{(1)}$:

$$\sum_{k=1}^{K}\sum_{m=1}^{M} a_k^{(1)}(m)\hat{\mu}(k,m) > \sum_{k=1}^{K}\sum_{m=1}^{M} a_k(m)\hat{\mu}(k,m)\eta_m(\boldsymbol{a}). \tag{21}$$

This happens when:

$$J_1 - \Delta KN > J_2 + \Delta KN. \tag{22}$$

For (22) to hold, Δ must satisfy:

$$\Delta < \frac{J_1 - J_2}{2KN}. \tag{23}$$

■

Next, we upper bound the probability of error, i.e., the probability of having reward estimates that do not satisfy (16) in the exploration epoch l. We also provide a lower bound on the length of the exploration epoch T_0.

Lemma 2. If $T_0 = \lceil \frac{2Me}{\Delta^2} \rceil$, all players have an estimation of the channel means satisfying (16), with probability $\geq 1 - \gamma_{e,l}$. Moreover, the probability of error in the l^{th} exploration epoch, $\gamma_{e,l}$, is upper bounded by:

$$\gamma_{e,l} \leq 4M^2 e^{-l}. \tag{24}$$

Proof. As in [6] and [12], we first find Q, the required number of observations of each channel by each user to guarantee (16). In order to do so, we first need to bound the probability of each user not having a correct estimation of the channel means. Let $\gamma = \gamma_{e,l}/2$. Define the following events:

- A: all players have an estimate satisfying (16),
- B: all players have $\geq Q$ observations of each channel,
- A_k: player k has an estimate satisfying (16),
- B_k: player k observed each channel $\geq Q$ times.

We need the following to hold:

$$\Pr(\bar{A}_k | B_k) \leq \gamma/K. \tag{25}$$

In fact, we have:

$$\begin{aligned} &\Pr(\bar{A}_k|B_k) \leq \Pr\left(\exists\, m, \text{s.t. } |\mu(k,m) - \hat{\mu}(k,m)| > \Delta \mid B_k\right) \overset{(a)}{\leq} \\ &\sum_{m=1}^{M} \Pr\left(|\mu(k,m) - \hat{\mu}(k,m)| > \Delta \mid B_k\right) = \\ &\sum_{m=1}^{M}\sum_{q=Q}^{\infty} \Pr\left(|\mu(k,m) - \hat{\mu}(k,m)| > \Delta \mid k \text{ has } q \text{ observations}\right. \\ &\left.\text{of each arm}\right) \Pr\left(q \text{ views} | q \geq Q\right) \overset{(b)}{\leq} \sum_{m=1}^{M}\sum_{q=Q}^{\infty} 2e^{(-2q\Delta^2)} \times \end{aligned}$$

$$\Pr\ (q \text{ views}|\ q \geq Q) = \sum_{m=1}^{M} 2e^{(-2Q\Delta^2)} = 2Me^{(-2Q\Delta^2)}, \quad (26)$$

where (a) results from applying the union bound and (b) from using Hoeffding's inequality [17].

To ensure this probability is lower than $\frac{\gamma}{K}$, Q must satisfy:

$$Q \geq \frac{1}{2\Delta^2}\log(\frac{2KM}{\gamma}). \quad (27)$$

Then, the following holds:

$$\Pr\ (A|B) = 1-\Pr(\bar{A}|B) \geq 1-\sum_{k=1}^{K}\Pr\ (\bar{A}_k|B_k) \geq 1-K\frac{\gamma}{K} = 1-\gamma. \quad (28)$$

Hence, if Q satisfies (27), all users have an estimate for every channel satisfying (16) with a probability higher than $1-\gamma$.

Next, we need to find a time horizon T_h for the exploration phases large enough such that all players have $\geq Q$ observations of each arm with probability higher than $1-\gamma$. Note that the length of each exploration phase T_0 does not necessarily satisfy $T_0 \geq T_h$. In other words, all players can get $\geq Q$ observations of each arm with probability higher than $1-\gamma$ after multiple exploration phases.

Let $A_{k,m}(t) = 1$ if player k observed channel m at timeslot t, and 0 otherwise. We have:

$$\begin{aligned}&\Pr\ (k \text{ has} \leq \frac{1}{2}T_h\mathbb{E}[A_{k,m}(t)] \text{ observations}) = \\ &\Pr\ (\sum_{t=1}^{T_h} A_{k,m}(t) \leq \frac{1}{2}T_h\mathbb{E}[A_{k,m}(t)] \text{ observations}) \overset{(a)}{\leq} \\ &e^{\left(\frac{-\frac{1}{4}T_h\mathbb{E}[A_{k,m}(t)]}{2}\right)},\end{aligned} \quad (29)$$

where (a) results from applying the Chernoff bound.

By using a union bound on (29), we get:

$$\begin{aligned}\Pr\ (\exists k,m \text{ s.t. } k \text{ has} \leq \frac{1}{2}T_h\mathbb{E}[A_{k,m}(t)] \text{ observations}) \\ \leq KMe^{\left(\frac{-\frac{1}{4}T_h\mathbb{E}[A_{k,m}(t)]}{2}\right)}.\end{aligned} \quad (30)$$

To bound the above probability by γ, T_h must satisfy:

$$T_h \geq \frac{8}{\mathbb{E}[A_{k,m}(t)]}\log\left(\frac{KM}{\gamma}\right). \quad (31)$$

Moreover, the number of observations of each arm $\sum_{t=1}^{T_h} A_{k,m}(t)$ must be at least Q. Hence, we need:

$$\sum_{t=1}^{T_h} A_{k,m}(t) > \frac{1}{2}T_h\mathbb{E}[A_{k,m}(t)] \geq Q > \frac{1}{2\Delta^2}\log\left(\frac{2KM}{\gamma}\right), \quad (32)$$

which is warranted if:

$$T_h \geq \left\lceil \max\left\{\frac{8\log\left(\frac{KM}{\gamma}\right)}{\mathbb{E}[A_{k,m}(t)]}, \frac{\log\left(\frac{2KM}{\gamma}\right)}{\Delta^2\mathbb{E}[A_{k,m}(t)]}\right\}\right\rceil. \quad (33)$$

Note that $\mathbb{E}[A_{k,m}(t)] = \frac{1}{M}\left(1-\frac{1}{M}\right)^{K-1} \geq \left(Me^{\left(\frac{K-1}{M-1}\right)}\right)^{-1}$ which follows since $(1-\frac{1}{x})^{x-1} \geq \frac{1}{e}$. If K is unknown, since $M > K$, the following holds: $\mathbb{E}[A_{k,m}(t)] \geq \frac{1}{Me}$. Hence, (33) can be formulated as:

$$T_h \geq \left\lceil \max\left\{8Me\log\left(\frac{M^2}{\gamma}\right), \frac{Me\log\left(\frac{2M^2}{\gamma}\right)}{\Delta^2}\right\}\right\rceil. \quad (34)$$

Having T_h, the probability of all users having an estimate of every channel mean satisfying (16) is given by:

$$\begin{aligned}&\Pr(A) = 1-\Pr(\bar{A}) = 1-\big(\Pr(\bar{A}|B)\Pr(B)+\Pr(\bar{A}|\bar{B})\Pr(\bar{B})\big) \\ &\geq 1-\big(\Pr(\bar{A}|B))+\Pr(\bar{B})\big) \geq 1-(\gamma+\gamma) = 1-\gamma_{e,l}.\end{aligned} \quad (35)$$

Moreover, $\Delta \leq \frac{J_1-J_2}{2KN} \leq \frac{KN-0}{2KN} \leq \frac{1}{2}$. Hence, (34) is satisfied if:

$$T_h = \frac{2Me}{\Delta^2}\log\left(\frac{4M^2}{\gamma_{e,l}}\right). \quad (36)$$

To upper bound the error probability in the l^{th} exploration epoch, we first note that:

$$T_0 \times l = T_h = \frac{2Me}{\Delta^2}\log\left(\frac{4M^2}{\gamma_{e,l}}\right). \quad (37)$$

To have $\gamma_{e,l} \leq 4M^2e^{-l}$, the length of each exploration epoch must satisfy:

$$T_0 \geq \frac{2Me}{\Delta^2}. \quad (38)$$

■

V. Matching Phase

The purpose of the matching phase is to reach a final assignment in which every user accesses N channels without collision, such that the sum reward achieved by all users is maximized. The dynamics introduced in section III-A induce a Markov chain over the state space $\mathcal{Z} = \prod_{k=1}^{K}\{\mathcal{A}_k \times [0,1]^{N\times 1} \times \{C,D\}\}$. Let P^ϵ denote the transition matrix of the regular perturbed Markov chain $\mathcal{Z}$. [14] guarantees that, when playing according to these dynamics, the optimal state, which maximizes the sum rewards, is most frequently played. To prove this, the authors of [14] rely on the theory of resistance trees for regular perturbed Markov chains [18]. The dynamics used in this paper differ from those in [14] in two aspects:

1) If user k experiences a collision on some channel m, i.e., its received reward on m is 0, user k is discontent with probability one. In [14], the game is assumed to be inter-dependent. Interdependency means that it is not possible to partition users into two groups that do not interact with each other. However, this property does not hold in the considered setting as shown in [8]. Therefore, as in [8], to characterize the stable states of the unperturbed chain when $\epsilon = 0$, a player with 0 reward on some channels is discontent with probability one.
2) For the transition probabilities between content and discontent in (9), instead of using $\epsilon^{N-\sum_{n=1}^{N} u_{k,n}}$, we use $\epsilon^{u_{k,\max}-\sum_{n=1}^{N} u_{k,n}}$, since the maximum utility achievable by each user k is $u_{k,\max}$.

Lemma 3. Let D^0 denote the set of states where all users are discontent. Moreover, let C^0 denote all singleton states where

all users are content and their baseline actions and utilities are aligned. As proved in [14], the only recurrence states of $\mathcal{Z}$ are D^0 and all singletons in C^0.

Similarly to [14], the resistance of moving from D^0 to any state $z \in C^0$ is $r(D^0 \to z) = \sum_{k=1}^{K} \left(u_{k,\max} - \sum_{n=1}^{N} u_{k,n} \right)$. The transition $z \in C^0 \to D$ has a resistance of $r(z \to D^0) = c$ and the resistance of moving from any state $z \in C^0$ to $z' \in C^0$ is $c \leq r(z \to z') \leq 2c$. The stochastic potential of any state $z \in C^0$ is of the form: $\zeta(z) = c[|C^0| - 1] + \sum_{k=1}^{K} \left(u_{k,\max} - \sum_{n=1}^{N} u_{k,n} \right)$. From Theorem 1 of [14], the stable state is the one minimizing the stochastic potential, hence the one maximizing the achieved sum reward. This stable state is guaranteed to be played the majority of times for a small enough perturbation ϵ [8], [14]. In the exploitation phase, since the state that was most played and that resulted most in the players being content is played, the stable state is hence expected to be played with high probability.

Let π denote the stationary distribution of the Markov chain $\mathcal{Z}$ and let $\boldsymbol{z}^* = [\bar{\boldsymbol{a}}^*, \bar{\boldsymbol{u}}^*, C^K]$ denote the optimal state. According to [8], $\pi(\boldsymbol{z}^*) > 1/2$ for a small enough perturbation ϵ. The following lemma characterizes the probability of error in the matching phase of the l^{th} epoch, $\delta_{m,l}$.

Lemma 4. Let $\boldsymbol{a}^{(l)}$ denote the action that was most played in some epoch l. The probability of error in the matching phase in epoch l, $\delta_{m,l}$, is upper bounded by:

$$\delta_{m,l} = \Pr(\boldsymbol{a}^* \neq \boldsymbol{a}^{(l)}) \leq A_0 \left\|\phi\right\|_\pi e^{\left(\frac{-\theta^2 \pi(\boldsymbol{z}^*) c_2 l^{1+\delta}}{72 T_m(1/8)}\right)}. \tag{39}$$

In (39), $\boldsymbol{a}^*$ is the optimal action defined in (2), A_0 is a constant, ϕ_π is the probability distribution of the initial state played in epoch l and $T_m(1/8)$ is the mixing time of the Markov chain $\mathcal{Z}$ with an accuracy of 1/8 [19].

Proof. In the matching phase, each user k keeps a counter of the number of times each action was played and resulted in k being content. At the end of the matching phase, user k chooses the action that was most played and resulted in k being content and continuously plays it in the exploitation phase. If the optimal strategy profile $\boldsymbol{a}^*$ was played more than $c_1 l^{1+\delta}/2$ times during matching phase l, then each user has played the optimal action more than half of the timeslots during the matching phase. Hence, the optimal action is played during the exploitation phase. Therefore, we can upper bound the probability of error in the matching phase by finding the probability of the optimal action being played less than half of the timeslots in the matching phase of epoch l. Let $f(\boldsymbol{z}) = \mathbb{I}(\boldsymbol{z} = \boldsymbol{z}^*)$, where $\mathbb{I}$ is the indicator function. Then,

$$\delta_{m,l} = \Pr(\boldsymbol{a}^* \neq \boldsymbol{a}^{(l)}) \leq \Pr(\sum_{\tau=1}^{c_1 l^{1+\delta}} f(\boldsymbol{z}(\boldsymbol{\tau})) \leq \frac{c_1 l^{1+\delta}}{2}). \tag{40}$$

Let $\theta = 1 - \frac{1}{2\pi(\boldsymbol{z}^*)}$, where $0 \leq \theta \leq 1$ since $\pi(\boldsymbol{z}^*) \geq 1/2$. Then,

$$\delta_{m,l} \leq \Pr\left(\sum_{\tau=1}^{c_1 l^{1+\delta}} f(\boldsymbol{z}(\boldsymbol{\tau})) \leq (1-\theta)\pi(\boldsymbol{z}^*) c_1 l^{1+\delta} \right). \tag{41}$$

Using the concentration bound of Theorem 3 in [19], the above probability can be bounded by:

$$\delta_{m,l} \leq A_0 \left\|\phi\right\|_\pi e^{\left(\frac{-\theta^2 \pi(\boldsymbol{z}^*) c_1 l^{1+\delta}}{72 T_m(1/8)}\right)}, \tag{42}$$

where A_0 is a constant, ϕ_π is the distribution of the initial state of the matching phase at the l^{th} epoch, and $T_m(1/8)$ is the mixing time of the Markov chain with an accuracy of 1/8.

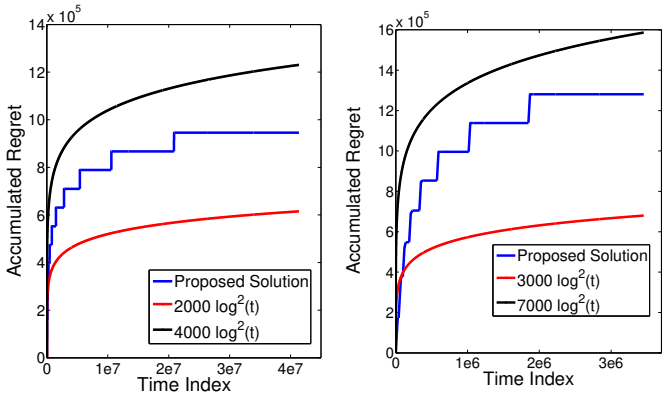
$\delta_{m,l}$ can be further upper bounded as:

$$\delta_{m,l} \leq A_0 \left\|\phi\right\|_\pi e^{\left(\frac{-\theta^2 \pi(\boldsymbol{z}^*) c_1 l^{1+\delta}}{72 T_m(1/8)}\right)} \leq A_1 e^{(-l^{1+\delta})}. \tag{43}$$

■

VI. Simulation Results

Extensive simulations of the proposed algorithm were conducted to validate its performance. The following simulation parameters were chosen: $c_1 = 3000, c_2 = 5000, \epsilon = 10^{-4}, c = KN, \delta = 0$.



(a) (b)

Fig. 1: Accumulated regret as time progresses for (a) $K = 2, M = 6, N = 3$, (b) $K = 4, M = 8, N = 2$.

Fig. 1 shows the average accumulated regret as a function of time, averaged over 50 realizations of the algorithm for different system settings. The results show that the average accumulated regret increases with time as $\mathcal{O}(\log(t)^2)$. More specifically, the regret for a system consisting of $K = 2$ users, $M = 6$ channels, and where each user pulls $N = 3$ channels in each timeslot, is bounded between $2000\log(t)^2$ and $4000\log(t)^2$, as shown in Fig. 1a. The regret of a system consisting of $K = 4$ users, $M = 8$ channels and with $N = 2$ channels pulled at each timeslot, is bounded between $3000\log(t)^2$ and $7000\log(t)^2$ as shown in Fig. 1b. The higher regret observed for the case of $K = 4$ users is due to the system taking a longer time to converge as shown next.

In Fig. 2, the mean achieved reward normalized by the reward of the optimal allocation is shown for different system settings. The performance of the proposed method is compared against that of the upper confidence bound (UCB) algorithm [9]. The

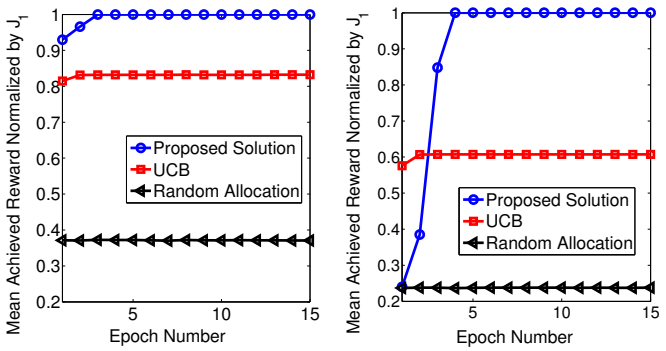


Fig. 2: Mean of the achieved reward normalized by the reward of the optimal allocation J_1 for (a) $K = 2, M = 6, N = 3$, (b) $K = 4, M = 8, N = 2$.

simulated UCB algorithm consists of each player choosing N arms according to the UCB technique without consideration for the other players in the system. The performance of the random selection technique, where each user chooses an action uniformly at random (i.e., with a probability given by (6)), is also shown for comparison. Fig. 2 shows the superior performance of the proposed method. Moreover, Fig. 2 shows that the algorithm convergence is much faster for a system consisting of a smaller number of users.

TABLE I

Method	**Average Hitting Time of Optimal Allocation** $(K = 2, M = 6, N = 3)$, $(K = 4, M = 8, N = 2)$	**Percentage of Optimal Allocation Hits** $(K = 2, M = 6, N = 3)$, $(K = 4, M = 8, N = 2)$
Proposed Solution	$(1.8 \times 10^4), (2.3 \times 10^4)$	$(97), (96)$
UCB	$(9 \times 10^5), (10^7)$	$(17.9), (0.1)$
Random Selection	$(1.8 \times 10^4), (2.4 \times 10^6)$	$(0.25), (2 \times 10^{-4})$

Finally, Table I compares the average hitting time of the optimal allocation and the percentage of hitting the optimal allocation for different system settings. The hitting time of the optimal allocation is defined as the first time this optimal allocation is played. Table I also shows the superiority of the proposed method over the UCB algorithm and the random selection technique. In fact, the proposed algorithm has a smaller average hitting time for both simulated system settings. Moreover, the percentage of optimal allocation hits of the proposed method greatly exceeds the percentage of the UCB and the random technique.

VII. Conclusion

In this paper, the uncoordinated spectrum access problem, where each user is allowed to choose N channels at each timeslot, was modeled as a multi-user MAB with multiple plays and varying user rewards. A game-theoretic approach was used to develop an algorithm with a sub-linear regret of $\mathcal{O}(\log^2 T)$. Simulation results validated the sub-linear regret of the proposed method and showed its superior performance, when compared with two other algorithms. The case of non-zero rewards on collision is an interesting extension of this work, and will be studied in a future work.

Acknowledgment

This work has been funded with support from the UBL, the GdR ISIS, the Lebanese University, and the US National Science Foundation SpecEES program under grant number 1730882, throughout the University of Illinois at Urbana-Champaign (UIUC). The first author would also like to thank Ms. Akshayaa Magesh (UIUC) for her help with the subject of MABs and for useful discussions.

References

[1] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: potential, challenges, and solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

[2] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-free radio access for short-packet communications over 5G networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–7.

[3] S. K. Sharma and X. Wang, "Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, Apr. 2019.

[4] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[5] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of uplink grant-free NOMA system," *IEEE Internet Things J., Early Access*, pp. 1–11, Feb. 2020.

[6] M. Bande and V. V. Veeravalli, "Multi-user multi-armed bandits for uncoordinated spectrum access," 2018. [Online]. Available: arxiv:1807.00867

[7] A. Magesh and V. V. Veeravalli, "Multi-player multi-armed bandits with non-zero rewards on collisions for uncoordinated spectrum access," 2019. [Online]. Available: arXiv:1910.09089

[8] I. Bistritz and A. Leshem, "Distributed multi-player bandits - a game of thrones approach," in *32nd Proc. Int. Conf. on Neural Inf. Process. Syst.*, ser. NIPS'18, Montreal, Canada, 2018, pp. 7222–7232.

[9] T. Lattimore and C. Szepesvri, *Bandit Algorithms*. Cambridge Univ. Press, 2020.

[10] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *IEEE Int. Conf. on Commun. Workshops*, May 2008, pp. 487–492.

[11] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[12] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits-a musical chairs approach," in *Int. Conf. on Mach. Learn.*, 2016, pp. 155–163.

[13] E. Boursier, V. Perchet, E. Kaufmann, and A. Mehrabian, "A practical algorithm for multiplayer bandits when arm means vary among players," 2019. [Online]. Available: arXiv:1902.01239

[14] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," in *SIAM J. Control Optim.*, no. 5, 2014, pp. 2753–2770.

[15] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.

[16] Y. Xia, T. Qin, W. Ma, N. Yu, and T. Liu, "Budgeted multi-armed bandits with multiple plays," *25th Proc. Int. Joint Conf. on Artif. Intell.*, pp. 2210–2216, July 2016.

[17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[18] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, 1993. [Online]. Available: http://www.jstor.org/stable/2951778

[19] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified," *29th Symp. Theor. Aspects of Comput. Sci.*, pp. 124–135, Feb. 2012.