Multi-User MABs with User Dependent Rewards for Uncoordinated Spectrum Access

Akshayaa Magesh*, Venugopal V. Veeravalli †
Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Emails: *amagesh2@illinois.edu, †vvv@illinois.edu

***Abstract*—The uncoordinated spectrum access problem is studied using a multi-player multi-armed bandits framework. In the considered system model there is no central control and the users cannot communicate with each other. Furthermore, the environment may appear differently to different users, *i.e.*, the mean rewards as seen by different users for a particular channel may be different. With this setup, we present a policy that achieves expected regret of $O(\log T)$ over a time horizon of duration T.**

I. Introduction

Current spectrum management protocols treat frequency spectrum as a fixed commodity, which leads to spectrum under utilization. Dynamic spectrum access techniques have emerged as good strategies to improve spectrum utilisation. Existing techniques for dynamic spectrum access have focused primarily on the primary/secondary user paradigm, where secondary users detect vacant bandwidths when available and vacate the occupied channel when a primary user wants to transmit.

In this work, we consider the uncoordinated spectrum access model, where there is no such hierarchy among users and users are not allowed to communicate with each other. The users follow a common protocol designed to maximize the system performance rather than an individual's reward. This problem is studied using the stochastic multi-user multi-armed bandit framework. The channels are treated as the arms and the channel gains or the rates received from the channels can be interpreted as the rewards.

In most of the previous work in this area (eg. [1], [2], [3] and [4]), it is assumed that the reward distributions for each channel is the same across all users. In [1], [2] and [4], the assumption is that, in the case of a collision (when more than one user access a channel) all the colliding users receive zero reward, which is the assumption we work with in this paper. The algorithm in [1] combines an ϵ-greedy approach with a collision avoiding mechanism and achieves expected regret of $O(T^{\frac{2}{3}})$. The musical chairs algorithm in [2] has an exploration phase where the users estimate the mean rewards for the channels and the number of users, and the users occupy one of the best arms for the remainder of the time horizon. In [3], a model is studied in which more than one user can access a channel simultaneously and receive non-zero rewards and an approach similar to the musical chairs algorithm [2] is presented. The policies provided in [2] and [3] provide guarantees of constant regret with high probability. The algorithm presented in [4] achieves a regret of $O(\log T)$. The lower bound for the single user case is $O(\log T)$, and since we are considering a scenario without communication between the users, we expect the regret lower bound to be $O(\log T)$ at best, which we show is achieved by the policy presented in this paper.

Given that in a practical scenario, users are not colocated in a wireless network, assuming that different users have different reward distributions for the channels results in a more realistic model. There has been some work covering varying reward distributions across users (eg. [5], [6], [7] , [8] and [9]). The algorithms presented in [5] and [6] consider such a model, with an assumption that the players can sense what happens on a channel, *i.e.*, if someone is using the channel or not, or if there is a collision on it. However, such an assumption might be unrealistic for the uncoordinated spectrum access problem. In our work, we consider a fully distributed scenario where players only have access to their previous actions and rewards. The work in [7] considers such a fully distributed setting, and the proposed algorithm achieves a regret of $O(\log^2 T)$. They consider a game-theoretic approach to solve for the pareto-optimal matching that maximizes the system welfare. The work in [9] extends this game-theoretic approach to the case of non-zero rewards on collisions.

The work in [4] introduces the idea of using forced collisions as a way to communicate among the users in the homogeneous setting where the reward distributions for the channels are the same across users. In this work, we use the idea of forced collisions in the heterogeneous case. The algorithm presented here was developed independently of the work in [8], where an approach similar to ours is explored, *i.e.*, a fully distributed setting with user dependent rewards and with zero reward on collision. However, while the algorithm presented in [8] achieves logarithmic regret only in the case of a unique optimal matching, the policy presented in this paper results in logarithmic regret even in the case of multiple optimal matchings.

This research was supported by the US NSF SpecEES under grant number 1730882, through the University of Illinois at Urbana-Champaign.

Authorized licensed use limited to: University of Illinois. Downloaded on January 22,2021 at 20:12:30 UTC from IEEE Xplore. Restrictions apply.

II. System Model

We consider the scenario of a multi-user game involving K users and M channels as the arms in a stochastic multi-armed bandit setup. We assume that $K < M$ and that the rewards for the channels are bounded in $[0, 1]$. Let the mean reward for user j on channel m be denoted by $\mu_j(m)$. Consider a time horizon T, and let the action taken by user (arm chosen by the user) j at time $t \leq T$ be $a_{t,j}$. In the case of a collision, *i.e.*, when multiple users access the same channel, all the colliding users receive zero reward.

Let $\mathcal{A}(K, M)$ denote all the possible user channel matchings, *i.e.*, $\mathbf{a} = [a_1, a_2, ..., a_K] \in \mathcal{A}(K, M)$, with a_j denoting the action taken by user j. Since we assume that colliding users get zero rewards, we only consider matchings that are unique, *i.e.*, once for which all the users are assigned to distinct arms. Let $\mathbf{a}^* \in \mathcal{A}(K, M)$ be such that

$$\mathbf{a}^* \in \underset{\mathbf{a}\in\mathcal{A}(K,M)}{\arg\max} \sum_{j=1}^{K} \mu_j(a_j).$$

Define J_1 to be the system reward for the optimal matching, *i.e.*, $J_1 = \sum_{j=1}^{K} \mu_j(a_j^*)$, and J_2 to be the system reward for the second optimal matching. In our algorithm, we assume that we have access to a lower bound on the parameter Δ defined as follows (see also [5], [7]):

$$\Delta = \frac{J_1 - J_2}{2M}$$

Note that this quantity is strictly positive even in the case of multiple optimal matchings.
The expected regret of the system is defined as

$$R(T) = T \sum_{j=1}^{K} \mu_j(a_j^*) - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} \mu_j(a_{t,j})\right]$$

where the expectation is over the actions of the players.

III. Algorithm

We assume that the players are time synchronised and that they enter the system at $t = 0$. The algorithm proceeds in epochs for each user. This allows us to proceed without knowing the time horizon T. Each epoch has three phases.

The first phase is the exploration phase, which has two parts. The first part is the fixing phase and is done for each user to obtain a unique ID. Each user accesses the arms uniformly and once a free arm is found, *i.e.*, non-zero reward is received, that arm is played for the rest of the fixing phase. The channel numbers they settle on serve as their IDs. At the end of the fixing phase, the users that are not fixed occupy channel 1, and the fixed users access channel 1 in order of their IDs sequentially. If the fixed users do not face a collision during this step, all users have obtained unique IDs. Once all the users obtain unique IDs, say at epoch ℓ_f, this part of the exploration phase is no longer done from epoch $\ell_f + 1$. The second part of the exploration phase is for the users to get estimates of the mean rewards of the arms. The users start from the channels corresponding to their IDs, and sample each arm for γ time units in a round-robin fashion.

The second phase is the matching phase and its purpose is for the users to arrive at the optimal matching. The first part of the matching phase is for each user to communicate their estimates of the mean rewards of all the channels to the other users. The users transmit the estimated mean rewards $\hat{\mu}_j(m)$ for all the channels in the order of their IDs. Since the players are not allowed to directly communicate with each other, collisions are used as a way to exchange information among players. The use of forced collisions as a form of communication was introduced in [4]. The main idea is that there are M channels available to the users and the transmitting user j occupies a certain channel. When the receiving users access the channels one at a time, the channel number on which they face a collision gives them information about what was being transmitted. The value of the estimate $\hat{\mu}_j(m)$ that each user has received at the end of the matching phase is denoted by $\hat{\hat{\mu}}_j(m)$. Note that this is similar to truncating the value of $\hat{\mu}_j(m)$ to a finite number of bits as $\hat{\hat{\mu}}_j(m)$. At the end of the first part of the matching phase, $|\hat{\mu}_j(m) - \hat{\hat{\mu}}_j(m)| \leq \Delta/2$ for $j \in [K]$ and $m \in [M]$ given that all the users have a unique ID.

Once this communication part of the matching phase is completed, each user has the same approximated values of estimated mean rewards. Each user then independently solves for the set of optimal matchings from the matrix of $\hat{\hat{\mu}}_j(m)$ values. If there is a unique optimal matching, the users play the arm according to that matching for the exploitation phase. If there are multiple optimal matchings, it is necessary for each user to choose the same optimal matching from this set. This is achieved in the following manner. The user with ID number one chooses one of the optimal matchings and occupies that channel. The remaining users access the M channels in order of their IDs to get the channel number chosen by the user with ID one and update the set of optimal matchings that correspond with the channel chosen by user one. This process is repeated until all the users settle on the same optimal matching. This matching is played by all the users for the exploitation phase.

In our algorithm, the estimated mean rewards of all the channels are communicated to each user through the idea of forced collisions, and hence all the users can independently solve the assignment problem to arrive at the same set of optimal matchings. However, in [8], the communication mechanism is adapted from [4], where a leader-follower protocol is employed. The followers send the values of estimated mean rewards to the leader, and the leader computes the matching that has to be played by the users. While the algorithm presented here gives guarantees of logarithmic regret for the cases of unique optimal matching and multiple optimal matchings, the work in [8] provides guarantees of logarithmic regret only for unique optimal matchings and quasi-logarithmic regret in the case of multiple optimal matchings. Note that the constants in the upper bound for average regret of our algorithm are comparable to the ones obtained in [8].

Authorized licensed use limited to: University of Illinois. Downloaded on January 22,2021 at 20:12:30 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Decentralized MUMAB Algorithm

Initialization: Set $\hat{\mu}_j(m) = 0$ for all values of $j \in [K]$ and all $m \in [M]$ and L_T as the last epoch with time horizon T.

for $\ell = 1, ..., L_T$ **do**

- **Exploration phase**:
- Fixing phase : Access channels uniformly for T_f time units till find a channel with no collision; fix on that for remainder of sub-phase. The channel number fixed on serves as unique ID.
- If not fixed during fixing phase, send a flag by occupying channel 1.
- If fixed during part fixing phase, access channel 1 in order of ID. Once all users are fixed, skip fixing phase.
- Access each channel in a round robin fashion for γ time units to get estimates $\hat{\mu}_j(m)$.
- **Matching phase**: Enter the matching algorithm to convey the estimates to all users, receive their estimates and calculate the optimal matching.
- **Exploitation phase**: Occupy the channel resulting from the matching algorithm for 2^ℓ time units.

end

Algorithm 2: Matching Algorithm for user i

Initialization: Transmit in order of ID.

for *Transmitting user* $j' = 1, ..., K$ **do**

- If turn to transmit, transmit $\hat{\mu}_{j'}(m)$ for $m = 1, ..., M$ as:
- Occupy channel $h_1 = \lceil M\hat{\mu}_{j'}(m) \rceil$ for M time units.
- Occupy $h_r = \lceil M^r(\hat{\mu}_{j'}(m) - \sum_{n=1}^{r-1} \frac{h_n - 1}{M^n}) \rceil$ for M time units for each subsequent round r for $\frac{1}{\log M} \log(\frac{1}{\Delta})$ rounds.
- If not turn to transmit, in order of IDs:
- Access channels in round robin fashion and initialize the estimates of the transmitted value as $\hat{\hat{\mu}}_{j'}(m) = \frac{2h-1}{2M}$ in the first round.
- Update in the subsequent round r as $\hat{\hat{\mu}}_{j'}(m) = \sum_{n=1}^{r-1} \frac{h_n - 1}{M^n} + \frac{2h_r - 1}{2M^r}$.

end

Calculate the set S_M of optimal matchings with values of estimates.

if S_M *is a singleton set* **then**

- Assign channel according to the optimal matching.

else

- User 1 chooses one of the optimal matchings.
- Remaining users update their set of optimal matchings accordingly.
- Similar mechanism for subsequent users allows users to settle on one of the optimal matchings.

end

IV. Regret Analysis

Theorem 1. *Assuming the rewards of each channel for all users are bounded in* $[0, 1]$ *and i.i.d. for all* $t \leq T$ *for a time horizon* T *and* $\Delta = \frac{J_1 - J_2}{2M}$ *is known, the regret of the decentralized MUMAB algorithm is* $O(\log T)$

Proof. The regret incurred during the L_T epochs can be analyzed as the sum of the regrets incurred in the three stages of the algorithm. From the structure of the epochs, we have that $L_T < \log T$.

1) Exploration phase: Let the regret incurred during the exploration phase for all epochs be R_1. The exploration goes on for $T_f + K + \gamma M$ time units till epoch ℓ_f (when all the users get fixed) and for γM time units after that. Choosing $T_f = M \log(20K)$ and $\gamma = \frac{1}{2\Delta^2}$, we have that

$$\begin{aligned} R_1 &\leq \sum_{\ell=1}^{L_T} \left(M \left(\frac{1}{2\Delta^2} + 1 + \log 20K \right) \right) \\ &\leq \left(M \left(\frac{1}{2\Delta^2} + 1 + \log 20K \right) \right) \log T. \end{aligned} \tag{1}$$

2) Matching phase: Let the regret incurred during the matching phase be R_2. The matching phase runs for $\frac{KM^3}{\log M} \log \frac{1}{\Delta} + (M-k)M^2 + M^3$ when there are multiple optimal matchings and for $\frac{KM^3}{\log M} \log \frac{1}{\Delta} + (M-k)M^2$ time units when there is an unique optimal matching. Thus

$$\begin{aligned} R_2 &\leq \sum_{\ell=1}^{L_T} \frac{KM^3}{\log M} \log \frac{1}{\Delta} + (M-k)M^2 + M^3 \\ &\leq \left(\frac{K}{\log M} \log \frac{1}{\Delta} + 2 \right) M^3 \log T. \end{aligned} \tag{2}$$

3) Exploitation phase: Let R_3 denote the regret incurred during the exploitation phase. Regret is incurred in the exploitation phase in epoch ℓ only in case of the following two events:

   a) The users do not get a unique ID in the first part of the exploration phase. Let $P_\ell(A)$ denote the probability that after the exploration phase of epoch ℓ the users do not have a unique ID

   b) Given that users have unique IDs, for some $j \in [k]$ and some $m \in [M]$, $|\hat{\mu}_j(m) - \mu_j(m)| > \Delta$. Let $P_\ell(B)$ denote the probability of this event.

   Thus we have that

$$R_3 = \sum_{\ell=1}^{L_T} 2^\ell (P_\ell(A) + P_\ell(B)). \tag{3}$$

   Let p_f denote the probability that with the fixing phase running for T_f time units, all users are fixed. From Lemma 1 of [4], we have that $p_f \geq (1 - Ke^{\frac{T_f}{M}})$. From the definition of the event A, we have that

$$P_\ell(A) = (1 - p_f)^\ell \tag{4}$$

Authorized licensed use limited to: University of Illinois. Downloaded on January 22,2021 at 20:12:30 UTC from IEEE Xplore. Restrictions apply.

Our choice of T_f results in $p_f > \frac{3(e-1)}{2e}$ and thus,

$$\sum_{\ell=1}^{L_T} 2^\ell P_\ell(A) = \sum_{\ell=1}^{L_T} 2^\ell (1-p_f)^\ell \leq \frac{e}{2e-3}. \quad (5)$$

We therefore get the first term of R_3 to be bounded by a finite number.

Let F denote the event that in epoch ℓ all the users have unique IDs. This means that in some epoch $\ell_f \in 0, 1, ..., \ell-1$, the system was fixed. Thus

$$\begin{aligned} &P_\ell(B) \\ &= P(|\hat{\mu}_j(m) - \mu_j(m)| > \Delta | F) \\ &= \sum_{i=1}^{\ell-1} P(|\hat{\mu}_j(m) - \mu_j(m)| > \Delta | F, \ell_f = i) P(\ell_f = i) \\ &\leq \sum_{i=1}^{\ell-1} 2e^{-2\Delta^2 \gamma(\ell-i)} p_f (1-p_f)^{i-1} \\ &= 2e^{-2\Delta^2\gamma\ell} \sum_{i=1}^{\ell-1} e^{2\Delta^2\gamma i} p_f (1-p_f)^{i-1}. \end{aligned} \quad (6)$$

Choosing $T_f = M \log(20K)$ gives $p_f > \frac{3(e-1)}{2e}$, and using $\gamma = \frac{1}{2\Delta^2}$ yields

$$P_\ell(B) \leq \frac{4e}{e-1} e^{-\ell} \quad (7)$$

and

$$\begin{aligned} \sum_{\ell=1}^{L_T} 2^\ell P_\ell(B) &\leq \frac{4e}{e-1} \sum_{\ell=1}^{L_T} \left(\frac{2}{e}\right)^{-\ell} \\ &\leq \frac{8e}{(e-1)(e-2)}. \end{aligned} \quad (8)$$

Thus

$$R_3 \leq C = \frac{e}{2e-3} + \frac{8e}{(e-1)(e-2)}. \quad (9)$$

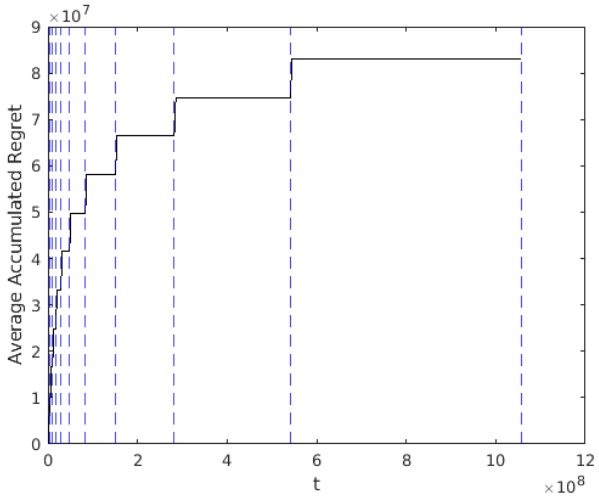
Therefore

$$\begin{aligned} R(T) &= R_1 + R_2 + R_3 \\ &\leq \left(\frac{M}{2\Delta^2} + \frac{KM^3}{\log M} \log \frac{1}{\Delta} + 4M^3 \right) \log T + C \\ &= O(\log T). \end{aligned} \quad (10)$$

□

V. Experimental Results

In this section, we present experimental results to validate the performance of our algorithm. We applied the proposed algorithm to a system with $K = 10$ users and $M = 10$ channels. Figure 1 shows the plot of average accumulated regret across the time horizon. The algorithm was run for 10 epochs. We see from the figure that the average accumulated regret grows sub-linearly with time and the regret is bounded by $\log T$.



VI. Conclusion

In this work, we have studied the uncoordinated spectrum access problem using a multi-player multi-armed bandit framework where there is no central control and users cannot communicate with each other. We have considered the case where the mean rewards as seen by different users for a particular channel may be different. Under this setup, we have presented an algorithm that provides a regret of $O(\log T)$. It is of interest to extend the results presented here to the case where the users receive non-zero rewards on collisions. See [9] for an initial study along these lines.

References

[1] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, 2014, pp. 66–81.

[2] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits: A musical chairs approach," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016, pp. 155–163.

[3] M. Bande and V. V. Veeravalli, "Multi-user multi-armed bandits for uncoordinated spectrum access," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, Feb 2019, pp. 653–657.

[4] E. Boursier and V. Perchet, "SIC-MMAB: Synchronisation Involves Communication in Multiplayer Multi-Armed Bandits," *arXiv e-prints*, p. arXiv:1809.08151, Sep 2018.

[5] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multi-player multi-armed bandits," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, Dec 2012, pp. 3960–3965.

[6] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, "Distributed learning and optimal assignment in multiplayer heterogeneous networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, April 2019, pp. 1693–1701.

[7] I. Bistritz and A. Leshem, "Distributed multi-player bandits - a game of thrones approach," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 7222–7232.

[8] E. Boursier, V. Perchet, E. Kaufmann, and A. Mehrabian, "A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players," *arXiv e-prints*, p. arXiv:1902.01239, Feb 2019.

[9] A. Magesh and V. V. Veeravalli, "Multi-player Multi-Armed Bandits with non-zero rewards on collisions for uncoordinated spectrum access," *arXiv e-prints*, p. arXiv:1910.09089, Oct 2019.

Authorized licensed use limited to: University of Illinois. Downloaded on January 22,2021 at 20:12:30 UTC from IEEE Xplore. Restrictions apply.