ADVERSARIAL MULTI-USER BANDITS FOR UNCOORDINATED SPECTRUM ACCESS

Meghana Bande *Venugopal V. Veeravalli*

ECE Department and Coordinated Science Laboratory,
University of Illinois at Urbana-Champaign.
Email: {mbande2,vvv}@illinois.edu

ABSTRACT

An adversarial multi-user multi-armed bandit framework is used to develop algorithms for uncoordinated spectrum access. It is assumed that the number of users is unknown, and that users receive zero reward on collision. The users do not coordinate with each other, and an adversary chooses different rewards for different users on the same channel. The proposed algorithm combines the Exp3.P algorithm developed in prior work for single user adversarial bandits with a collision resolution mechanism to achieve sub-linear regret. It is shown that if every user employs the proposed algorithm, the system wide regret is of the order $O(T^{\frac{3}{4}})$ over a horizon of time T. The algorithm is then extended to the dynamic case where the number of users in the system evolves over time, and it is shown to lead to sub-linear regret.

***Index Terms*—** Cognitive radio, multi-armed bandits, dynamic spectrum access.

I. INTRODUCTION

The existing spectrum management paradigm treats frequency spectrum as a fixed commodity, which leads to spectrum under-utilization. Cognitive radio has emerged as a useful strategy to increase spectrum utilization. The existing literature on cognitive radio has largely been focused on the primary/secondary user paradigm, where secondary users need to detect vacant spectrum when available and vacate the occupied spectrum when a primary user wants to transmit.

We focus on a different type of spectrum sharing system in which there is no distinction between users, and in which there is no coordination among the users. The collective performance across all users is more important than that of individual users. This is in contrast to the typical primary/secondary user paradigm in which secondary users bear the responsibility for ensuring priority-based spectrum sharing. We model this system using an adversarial multi-user multi-armed bandit framework [1]. Our goal is to design an efficient channel access mechanism by managing interference in the system through a decentralized policy across the users.

Multi-armed bandit problems have been studied in the context of cognitive radio using different formulations. However, all the existing approaches with multiple users have focussed on Markovian or stochastic multi-user multi-armed bandits (MABs). A Markovian channel model for a two-user two-channel system was considered in [2], where the probability transitions were assumed to be known. Coordination between users was considered in the schemes of [3], [4], [5]. The stochastic MAB model with no communication between the users was considered in [6], [7], [8], [9] and [10]. However, it was assumed that the channel is same for all the users. A stochastic multi-user MAB with user dependent rewards on channel was considered in [11]. However, the algorithm considers coordination and communication between users via an auction algorithm.

The adversarial bandit problem is an important variation of the multi-armed bandit problem, where no stochastic assumption is made on the generation of rewards. The term "adversarial" refers to the mechanism choosing the sequence of rewards on each arm. If this mechanism is independent of the users actions, then it is an *oblivious* adversary. If the mechanism may adapt to the users' past behaviors, then it is a non-oblivious adversary [1]. The existing literature on adversarial MABs is focused on the single user case, and a detailed overview of the proposed solutions for the adversarial MAB formulation can be found in [1]. The proposed algorithms in the single user adversarial setting achieve a sub-linear regret of the order of $O(\sqrt{T})$ over a time horizon T.

We consider multi-user dynamic spectrum allocation without any coordination among the users. We also assume that the rewards on each channel are user dependent and may vary with time. Such a system is captured through a multi-user adversarial MAB model, particularly when the reward distribution for each channel and user may change over time. We assume that the number of users is unknown and that there is no communication between the users. However, we make the mild assumption that the users have access to a shared clock for time synchronization (see also, [9], [12], [13]). We propose an algorithm, and show that if each user employs the algorithm, the system wide regret is $O(T^{\frac{3}{4}})$ over a time horizon T. To the best of our knowledge, we are the first to consider the multi-user setting for adversarial MABs and to provide sub-linear regret guarantees. We extend our algorithm to the dynamic case, and show that, with minor restrictions on the rate at which users enter the system, we can achieve sub-linear regret.

II. SYSTEM MODEL AND NOTATION

Let K be the number of users in the system and M the number of channels. We assume that there are more channels than users in the system i.e., $K \leq M$. We also assume initially that the users have unlimited data to transmit. In the dynamic setting, we lift this assumption, and allow the users to become inactive based on their data needs and new users to join the system. We assume that the users have knowledge of M but not of K. The assumption of known M is reasonable if the spectrum partition is enforced and fixed. On the other hand, it is not realistic to assume the knowledge of K in an uncoordinated network.

We model the system as an adversarial multi-user MAB with K users and M channels. We assume that each user chooses a

This work was supported in part by the NSF SpecEES program by grant number CNS1730882, through the University of Illinois.

channel according to the same algorithm. For user $k \in [K]$, let $p_t^k = (p_{1,t+1}^k, ..., p_{M,t+1}^k)$ denote the probability vector across the arms, where $p_{m,t}^k$ is the probability of choosing arm m at time t. Let $a_t^k \in [M]$ denote the channel chosen by user k at time t based on the previous reward history of the user according to p_t^k. We assume that if more than one user chooses the same channel, they all receive zero reward. In other words, the users observe zero reward on collision. If there is no collision on the channel, the user observes a reward that is chosen by an adversary. We assume that the adversary chooses different reward for different users for the same channel. For example, the reward could be the rate achieved by the user on the channel which depends on the channel gain of the specific user. Let $g_{a_t^k,t}^k$ denote the reward observed by user k on choosing channel a_t^k at time t. We assume that $g_{i,t}^k \in [0,1]$.

We adopt the standard notion of pseudo-regret used for adversarial bandits in [1]. The expected total regret in the system until time T is defined as

$$\mathbb{E}[R(T)] = \max_{\mathcal{K}:\mathcal{K}\subseteq[M],|\mathcal{K}|=K} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{K}} g_{i,t}^k - \sum_{t=1}^{T}\sum_{k=1}^{K} g_{a_t^k,t}^k\right].$$

III. SINGLE USER MAB

We consider the Exp3.P algorithm described in [1] for a single user MAB in an adversarial setting. We modify the algorithm so that the user chooses an arm and updates the probability vector only in a few time-slots. This modification is useful in the multi-user case, where the users may not choose an arm in each time-slot due to possible collisions. We now present a modified version of the Exp3.P algorithm, in which a new arm is chosen and the probability is updated at time-slots $t_1, t_2, \ldots, t_n$ such that $n \leq T$ and $t_{j+1} - t_j = \frac{T}{n}$. For each $j \in [n]$, we consider the reward over the time-period $t_{j+1} - t_j$, with the reward being normalized to lie between 0 and 1. We drop the superscript denoting the user in the notation for the single user case.

Modified Exp3.P

1: $\beta = \sqrt{\frac{\ln M}{Mn}}$, $\eta = 0.95\sqrt{\frac{\ln M}{Mn}}$ and $\gamma = 1.05\sqrt{\frac{M \ln M}{n}}$.
2: Initial probability distribution $p_0 = (\frac{1}{M}, \ldots, \frac{1}{M})$.
3: **for** $j = 1, \ldots, n$ **do**
4: $a_j \sim p_j$, remain on arm for next $t_{j+1} - t_j$ time-slots
5: Compute reward as $g'_{i,j} = \frac{\sum_{t_j \leq t \leq t_{j+1}} g_{i,t}}{t_{j+1} - t_j}$ and the estimated gain for each arm as

$$\tilde{g}_{i,j} = \frac{g'_{i,j}\mathbb{1}_{a_j=i} + \beta}{p_{i,j}}$$

and update the cumulative gain $\tilde{G}_{i,j} = \sum_{s=1}^{j} \tilde{g}_{i,s}$
6: Calculate the new probability distribution over the arms $p_{j+1} = (p_{1,j+1}, ..., p_{M,j+1})$ where

$$p_{i,j+1} = (1-\gamma)\frac{\exp(\eta\tilde{G}_{i,j})}{\sum_{m=1}^{M}\exp(\eta\tilde{G}_{m,j})} + \frac{\gamma}{M}$$

7: **end for**

Theorem 1: The expected regret of Modified Exp3.P algorithm until time T is given by

$$\mathbb{E}\left[\sum_{t=1}^{T}(g_{m,t} - g_{a_t,t})\right] \leq \max_{m\in[M]} \mathbb{E}\left[\sum_{t=1}^{T}(g_{m,t} - g_{a_t,t})\right] \leq \frac{T}{\sqrt{n}}h(M) \quad (1)$$

where $h(M) = 5.15\sqrt{M \ln M} + \sqrt{\frac{M}{\ln M}}$, and does not depend on T and $n \leq T$.

Proof: We have

$$\mathbb{E}\left[\sum_{t=1}^{T}(g_{m,t} - g_{a_t,t})\right] = (T/n)\mathbb{E}\left[\sum_{j=1}^{n}(g'_{m,j} - g'_{a_j,j})\right], \quad (2)$$

where $g'_{m,j} = \frac{\sum_{t_j \leq t \leq t_{j+1}} g_{m,t}}{t_{j+1} - t_j}$. Using (2) and noting that until time T we consider n time-slots, the proof follows from the regret bound of Exp3.P in [1]. ∎

IV. MULTI-USER MAB: ALGORITHM

We now consider the multi-user adversarial bandits under a known finite horizon T and propose an algorithm which when employed by all users independently leads to regret of order $O(T^{\frac{3}{4}})$.

In a multi-user adversarial system, every time t that a user k chooses an arm according to a certain probability distribution p_t^k to randomize against the adversary, there is a possibility for collision with other users. Hence there is a need for a collision resolution mechanism, so that the regret does not grow linearly with time. Instead of choosing an arm every time-slot, a user chooses an arm only a sub-linear number of times until T (e.g., T^x where $x < 1$). The goal is to randomize sufficient number of times so as to counteract the adversary while making sure that the regret due to collisions does not become large.

We propose an algorithm (Algorithm 1) that combines the modified Exp3.P algorithm (Section III) with a collision resolution mechanism. In the algorithm $x < 1$. In the analysis in Section V, we pick $x = \frac{1}{2}$ which is large enough to maintain the sub-linear regret achieved by the modified Exp3.P algorithm but small enough so that the regret due to collisions is sub-linear as well.

In every time-interval of length T^{1-x}, we first have a collision resolution phase. Each user chooses a channel with probability p_t^k. A user settles or fixes on a channel if at any time the user finds a channel without collision. Once a user settles on a channel, the user keeps transmitting on the channel until the end of the time-interval of length T^{1-x}. The system incurs regret until all K users have settled on K channels, and we call this duration the *fixing time* or the collision resolution phase. The remaining part of the algorithm corresponds to each of the K users employing the modified Exp3.P algorithm, where they choose a channel once every T^x time-slots.

V. MULTI-USER MAB: ANALYSIS

In this section, we first consider the regret due to the collision resolution phase, then the regret due to the modified Exp3.P part of Algorithm 1, and then combine them to find an upper bound on the system-wide regret incurred when each user independently employs Algorithm 1.

V-A. Regret during collision resolution

Theorem 2: The expected regret accumulated by the system during a collision resolution phase is upper bounded by

$$\frac{K^2M^K}{\gamma} \leq \frac{K^2M^KT^{\frac{x}{2}}}{\sqrt{M \ln M}}.$$

Proof: We first note from equation (3) that the probability of choosing any channel by any user is at least $\frac{\gamma}{M}$. Let $\rho_t^k = \max_m p_{m,t}^k$, which implies that $\rho_t^k \geq \frac{1}{M}$. Let "maximal" channel

Algorithm 1

$\beta = \sqrt{\frac{\ln M}{MT^x}}$, $\eta = 0.95\sqrt{\frac{\ln M}{MT^x}}$ and $\gamma = 1.05\sqrt{\frac{M \ln M}{T^x}}$.
The initial probability distribution $p_0^k = (\frac{1}{M}, \ldots, \frac{1}{M})$
for $t =$ multiples of $\frac{T}{T^x}$ **do**
  for $t' = 1$ to T^{1-x} **do**
    $a_{t'}^k \sim p_t^k$
    if no collision **then**
      break
    end if
  end for
  Choose action $a_{t'}^k$ for the remaining $T^{1-x} - t'$ time-slots
  Compute reward as $g_{i,t}'^k = \frac{\sum g_{i,t}^k}{T^{1-x}-t'}$ and the estimated gain for each arm as

$$\tilde{g}_{i,t}^k = \frac{g_{i,t}'^k \mathbb{1}_{a_{t'}^k = i} + \beta}{p_{i,t}^k}$$

  and update the cumulative gain $\tilde{G}_{i,t}^k = \sum_{s=1}^t \tilde{g}_{i,s}^k$
  Calculate the new probability distribution over the arms $p_{t+1}^k = (p_{1,t+1}^k, ..., p_{M,t+1}^k)$ where

$$p_{i,t+1}^k = (1-\gamma)\frac{\exp(\eta \tilde{G}_{i,t}^k)}{\sum_{m=1}^M \exp(\eta \tilde{G}_{m,t}^k)} + \frac{\gamma}{M} \tag{3}$$

end for

for a user refer to the channel that has the highest probability of being chosen by that particular user. Thus, each user can be associated with one channel such that probability of choosing it is greater than $\frac{1}{M}$. Since $K \leq M$, for each user, there exists at least one channel such that it not the maximal channel for any of the remaining $K-1$ users. Note that even when some users fix or settle on a channel, and there are both unfixed channels and unfixed users in the system, we can still find an unfixed channel such that it is not the maximal channel for the remaining unfixed users.

Based on the above discussion, we define the event B_k to be the event where all unfixed users except user k choose their maximal arm, and user k chooses an unfixed arm that is not the maximal arm for any other unfixed users.

Let $\mathcal{M}_{u,t}$ denote the set of unfixed arms at time t. The probability of any player k being fixed at time t is given by,

$$\begin{aligned}
&\Pr\{\text{User } k \text{ being fixed}\} \\
&= \sum_{m \in \mathcal{M}_{u,t}} \Pr\{\text{User } k \text{ is the only unfixed user on arm } m\} \\
&\geq \Pr(B_k) \\
&\geq (\Pi_{i\in[K], i\neq k} \rho_t^i) \min_{m \in \mathcal{M}_{u,t}} p_{m,t}^k \\
&\geq \frac{\gamma}{M}(\frac{1}{M})^{K-1} = \frac{\gamma}{M^K}.
\end{aligned}$$

For any player k, the expected time to get fixed is given by

$$\mathbb{E}[t_f^k] = \frac{1}{\Pr\{\text{User } k \text{ being fixed}\}} \leq \frac{M^K}{\gamma}$$

and the regret during the collision resolution phase is given by

$$\mathbb{E}\left[\sum_k \sum_{t=1}^{\max_k t_f^k} R_{k,t}\right] \leq \mathbb{E}[K \max\, t_f^k] \leq \mathbb{E}[K \sum_{k=1}^K t_f^k] \leq K^2 \mathbb{E}[t_f^k],$$

where $R_{k,t}$ denotes the regret incurred by player k at time t and we have $R_{k,t} \leq 1$ by our assumption that rewards lie between zero and one. ∎

V-B. Regret due to Modified Exp3.P

We now bound the regret incurred by the users using Algorithm 1 during the time the users are not in the collision resolution phase. This corresponds to each of the K users independently employing the modified Exp3.P algorithm introduced in Section III.

In Algorithm 1, when the users are not in the collision resolution phase, each user employs modified Exp3.P with $n = T^x$. Using the result of Theorem 1 for K users, for any distinct set $\mathcal{K} \subseteq [M]$ consisting of K arms,

$$\mathbb{E}\left[\sum_{t \notin \text{collision phase}} \left(\sum_{i\in\mathcal{K}} g_{i,t}^k - \sum_{k=1}^K g_{a_t^k,t}^k\right)\right] \leq KT^{1-\frac{x}{2}} h(M).$$

Thus,

$$\max_{\mathcal{K}} \mathbb{E}\left[\sum_{t \notin \text{collision phase}} \left(\sum_{i\in\mathcal{K}} g_{i,t}^k - \sum_{k=1}^K g_{a_t^k,t}^k\right)\right] \leq KT^{1-\frac{x}{2}} h(M) \tag{4}$$

where $h(M) = 5.15\sqrt{M \ln M} + \sqrt{\frac{M}{\ln M}}$, and does not depend on T.

V-C. Main Result

We now present the upper bound on the expected regret incurred by the users employing Algorithm 1.

Theorem 3: The expected regret of K users using Algorithm 1 with M arms for T time-slots, is given by

$$\mathbb{E}[R(T)] \leq T^{\frac{3}{4}} h'(M, K)$$

where $h'(M,K) = K\left(5.15\sqrt{M \ln M} + \sqrt{\frac{M}{\ln M}} + \frac{KM^K}{\sqrt{M \ln M}}\right)$, and does not depend on T. Thus, $\mathbb{E}[R(T)] \sim O(T^{\frac{3}{4}})$.

Proof: The expected regret is due to collision resolution as well as the modified Exp3.P algorithm which is played a sub-linear number of times. Let T_f denote the time taken for collision resolution.

$$\begin{aligned}
\mathbb{E}[R(T)] &\leq T^x K \mathbb{E}[T_f] + (T^{1-x} - \mathbb{E}[T_f]) h(M) T^{\frac{x}{2}} \\
&\leq \frac{K^2 M^K}{\sqrt{M \ln M}} T^{\frac{3x}{2}} + KT^{1-\frac{x}{2}} h(M) \\
&\sim O(T^{\frac{3x}{2}} + T^{1-\frac{x}{2}})
\end{aligned}$$

where the inequalities follow from Theorem 2 and equation (4), and $h(M) = 5.15\sqrt{M \ln M} + \sqrt{\frac{M}{\ln M}}$. If we choose x such that $\frac{3x}{2} = 1 - \frac{x}{2}$, we have $x = \frac{1}{2}$ which gives us

$$\mathbb{E}[R(T)] \leq T^{\frac{3}{4}} K\left(\frac{KM^K}{\sqrt{M \ln M}} + h(M)\right).$$

∎

VI. UNKNOWN TIME HORIZON

In this section, we extend the results to the case of unknown time horizon. Each user considers some known time τ greater than the expected fixing time for the system and runs Algorithm 1. Once the user reaches the end of time τ, the user continues to use Algorithm 1 with a time-period of length 2τ. In this way when the user reaches the end of the previous time-period, the user

doubles it and continues with Algorithm 1. Let T be such that $\tau + 2\tau + \ldots + 2^r\tau \leq T \leq \tau + 2\tau + \ldots + 2^{(r+1)}\tau$. Note that this is same as $2^{(r+1)}\tau \leq T + \tau < 2^{(r+2)}\tau$.

Algorithm 2

1: **for** $(2^{(r+1)} - 1)\tau \leq T < (2^{(r+2)} - 1)\tau$ **do**
2: Run Algorithm 1 with time-period $2^{r+1}\tau$
3: **end for**

Theorem 4: The expected regret from using Algorithm 2 for T time-slots where $(2^{(r+1)} - 1)\tau \leq T < (2^{(r+2)} - 1)\tau$ is given by

$$\mathbb{E}[R(T)] \leq h'(M,K)\frac{(2(T+\tau))^{\frac{3}{4}}}{2^{\frac{3}{4}} - 1}$$

where $h'(M,K) = K\left(5.15\sqrt{M\ln M} + \sqrt{\frac{M}{\ln M}} + \frac{KM^K}{\sqrt{M\ln M}}\right)$ and does not depend on T. Thus, $\mathbb{E}[R(T)] \sim O(T^{\frac{3}{4}})$.

Proof: We have $(2^{(r+1)} - 1)\tau \leq T < (2^{(r+2)} - 1)\tau$ which gives us $2^{(r+1)}\tau \leq T + \tau$.

Using Theorem 3, the regret up to time T bounded as follows:

$$\begin{aligned}\mathbb{E}[R(T)] &\leq h'(M,K)(\tau^{\frac{3}{4}} + (2\tau)^{\frac{3}{4}} + \ldots + (2^{r+1}\tau)^{\frac{3}{4}}) \\ &= h'(M,K)\tau^{\frac{3}{4}}\frac{(2^{(r+2)\frac{3}{4}} - 1)}{2^{\frac{3}{4}} - 1} \\ &\leq h'(M,K)\frac{(2(T+\tau))^{\frac{3}{4}} - \tau^{\frac{3}{4}}}{2^{\frac{3}{4}} - 1}.\end{aligned}$$

■

Note that each user only needs knowledge of K in order to fix on an initial τ such that $\tau \geq \mathbb{E}T_f$, where T_f is the time taken for collision resolution. Furthermore, τ can be chosen even without the knowledge of K by simply replacing K by M, and the analysis follows because $K \leq M$.

VII. DYNAMIC CASE

In this section, we extend the results to a dynamic system with a changing number of users. Consider a system which starts with K users, and in which users leave the system once they are done with their transmission. It is easy to see that Algorithm 2 in this case leads to system-wide regret of the order $O(T^{\frac{3}{4}})$ over a time horizon T.

Let us now consider a dynamic system where users enter and leave the system over time. In order to use Algorithm 2 to obtain a sub-linear regret bound, we need to impose some restrictions on the number of users that have entered the system until time t, which we denote by κ_t. It is easy to see that the number of epochs in which users enter the system must be sub-linear in time to have sub-linear regret in the system. We restrict the number of users entering the system κ_t to be $O(t^{\zeta})$ where $\zeta < \frac{1}{2}$. We note that this is similar to the dynamic case in [14] where there is a restriction on the number of users entering and leaving the system.

Let K_t denote the number of active users at time t. Note that even in the dynamic scenario, we still retain the assumption of having $K_t \leq M$ in the system.

Theorem 5: The expected system-wide regret from using Algorithm 2 for T time-slots where $(2^{(r+1)} - 1)\tau \leq T < (2^{(r+2)} - 1)\tau$ with the number of users entering the system $\kappa_T \sim O(T^{\zeta})$, with $\zeta < \frac{1}{2}$, is given by

$$\mathbb{E}[R(T)] \leq h'(M,M)\frac{(2(\tau+T))^{\frac{3}{4}}}{2^{\frac{3}{4}} - 1} + M\kappa_T T^{\frac{1}{2}}$$

where $h'(M,M) = M\left(5.15\sqrt{M\ln M} + \sqrt{\frac{M}{\ln M}} + \frac{M^{M+1}}{\sqrt{M\ln M}}\right)$ and does not depend on T. Thus, $\mathbb{E}[R(T)] \sim O(T^{\frac{3}{4}} + \kappa_T T^{\frac{1}{2}})$.

Proof: We have $(2^{(r+1)} - 1)\tau \leq T < (2^{(r+2)} - 1)\tau$ which gives us $2^{(r+1)}\tau \leq \tau + T$. In epochs where no users enter the system, the regret can be bound by Theorem 4, and in epochs with new users, the regret accumulates through the entire epoch. The epoch length is upper bounded by $(2^{(r+1)}\tau)^{\frac{1}{2}}$, since $x = \frac{1}{2}$ from Theorem 3. The regret up to time T bounded as follows:

$$\begin{aligned}\mathbb{E}[R(T)] &\leq \text{Static case regret} + K_t\sum^{\kappa_T}\text{Epoch length} \\ &\leq h'(M,M)\frac{(2(\tau+T))^{\frac{3}{4}}}{2^{\frac{3}{4}} - 1} + M\kappa_T(2^{r+1}\tau)^{\frac{1}{2}} \\ &\leq h'(M,M)\frac{(2(\tau+T))^{\frac{3}{4}}}{2^{\frac{3}{4}} - 1} + M\kappa_T(\tau+T)^{\frac{1}{2}}.\end{aligned}$$

Thus, $\mathbb{E}[R(T)] \sim O(T^{\frac{3}{4}} + \kappa_T T^{\frac{1}{2}})$, and if κ_T is $O(T^{\zeta})$, with $\zeta < \frac{1}{2}$, we have sub-linear regret. ■

VIII. EXPERIMENTS

In this section, we illustrate the performance of our algorithm in a simple adversarial setting. We consider a non-oblivious adversary, i.e., an adversary whose rewards do not depend on the users' reward history.

We consider a system with known time-horizon T and fixed number of users K. We choose $K = 4$ users and $M = 7$ channels. We set $T = 160000$, which gives us $T^{\frac{1}{2}} = 400$ time-slots, $\beta = 0.026$, $\eta = 0.025$ and $\gamma = 0.194$ in Algorithm 1. The reward distributions for the channels are drawn i.i.d from the uniform distribution $[a, 1]$ where a for each channel at each time-slot is drawn i.i.d from the uniform distribution $[0.2, 1]$.

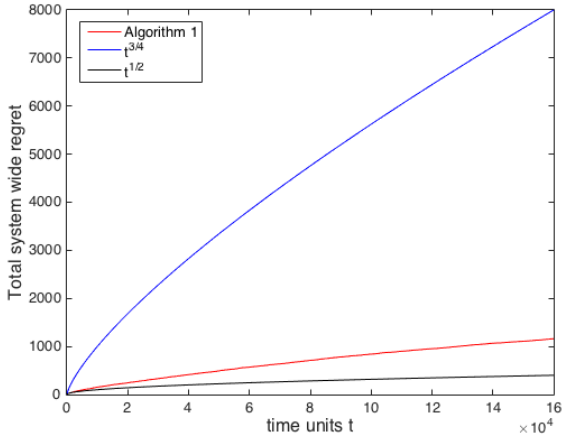

Fig. 1: Accumulated regret as a function of time.

We repeat the experiment 100 times and consider the average accumulated regret with time. From Figure 1, we see that the regret grows with time at a rate much lower than $T^{\frac{3}{4}}$, but higher than $T^{\frac{1}{2}}$, which is the expected regret in the single user case.

IX. REFERENCES

[1] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[2] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *IEEE International Conference on Communications (ICC) Workshops*, 2008, pp. 487–492.

[3] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 1904–1919, 2009.

[4] H. Gang, Z. Qian, and X. Ming, "Contention-aware spectrum sensing and access algorithm of cognitive network," in *IEEE International Conference on Cognitive Radio Oriented Wireless Networks and Communications*, 2008, pp. 1–8.

[5] H. Liu, L. Huang, B. Krishnamachari, and Q. Zhao, "A negotiation game for multichannel access in cognitive radio networks," in *Proceedings of the 4th Annual International Conference on Wireless Internet*, 2008.

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

[7] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[8] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 66–81.

[9] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits–a musical chairs approach," in *International Conference on Machine Learning*, 2016, pp. 155–163.

[10] M. Bande and V. V. Veeravalli, "Multi-user multi-armed bandits for uncoordinated spectrum access," in *Proc. IEEE International Conference on Computing, Networking and Communications (ICNC)*, 2019.

[11] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[12] J. Nieminen, R. Jantti, and L. Qian, "Time synchronization of cognitive radio networks," in *Global Telecommunications Conference, GLOBECOM*. IEEE, 2009, pp. 1–6.

[13] O. Avner and S. Mannor, "Learning to coordinate without communication in multi-user multi-armed bandit problems." *arXiv preprint arXiv:1504.08167*, 2015.

[14] M. Bande and V. V. Veeravalli, "Multi-user multi-armed bandits for uncoordinated spectrum access," *arXiv preprint arXiv:1807.00867*, 2018.