Performance and Convergence of Multi-user Online Learning

Cem Tekin, Mingyan Liu
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, Michigan, 48109-2122
Email: {cmtkn, mingyan}@umich.edu

Abstract

We study the problem of allocating multiple users to a set of wireless channels in a decentralized manner when the channel qualities are time-varying and unknown to the users, and accessing the same channel by multiple users leads to reduced quality (e.g., data rates) received by the users due to interference. In such a setting the users not only need to learn the inherent channel quality and at the same time the best allocations of users to channels so at to maximize the social welfare. Assuming that the users adopt a certain online learning algorithm, we investigate under what conditions the socially optimal allocation is achievable. In particular we examine the effect of different levels of knowledge the users may have and the amount of communications and cooperation. The general conclusion is that when the cooperation of users decreases and the uncertainty about channel payoffs increases it becomes harder to achieve the socially optimal allocation. Specifically, we consider three cases. In the first case, channel rates are generated by an *iid* process. The users do not know this process or the interference function, and there is no information exchange among users. We show that by using a randomized learning algorithm users converge to the pure Nash equilibria of an equivalent congestion game. In the second case, a user is assumed to know the total number of users, and the number of users on the channel it is using. We show that a sample-mean based algorithm can achieve the socially optimal allocation with a sub-linear regret in time. In the third case, we show that if the channel rates are constant but unknown, if a user knows the total number of users, then the socially optimal allocation is achieved in finite time with a randomization learning algorithm.

I. Introduction

In this paper we study the dynamic spectrum access and spectrum sharing problem in a learning context. Specifically, we consider a set of N common channels shared by a set of M users. A channel has time varying rate $r(t)$, and its statistics are not completely known by the users. Thus each user needs to employ some type of learning to figure out which channels are of better quality, e.g., in terms of their average achievable rates. At the same time, simultaneous use of the same channel by multiple users will result in reduced rate due to interference or collision. The precise form of this performance degradation may or may not be known to the user. Thus the users also need to use learning to avoid excess interference or congestion. Furthermore, each user may have private information that is not shared, e.g., users may perceive channel quality differently due to difference in location as well as individual modulation/coding schemes.

Without a central agent, and in the presence of information decentralization described above, we are interested in the following questions: (1) for a given common learning algorithm, does the multiuser (asynchronous) learning process converge, and (2) if it does, what is the quality of the equilibrium point with respect to a globally optimal spectrum allocation scheme, one that could be computed for a global objective function with full knowledge of channel statistics as well as user's private information.

A few recent studies have addressed these questions in some special cases. For instance, in [1] it was shown that learning using a sample-mean based index policy leads to a socially optimal (sum of individual utilities) allocation when channels evolve as iid processes and colliding players get zero reward provided that this optimal allocation is such that each user occupies one of the M best channels (in terms of average rates). This precludes the possibility that not all users may have the same set of M best channels, and that in some cases the best option is for multiple users to share a common channel, e.g., when $N < M$.

In this study we investigate under what conditions the socially optimal allocation is achievable by considering different levels of communication (or cooperation) allowed among users, and different levels of uncertainty on the channel statistics.

The general conclusion, as intuition would suggest, is that when the cooperation of users increases and the channel uncertainty decreases it becomes easier to achieve the socially optimal welfare. Specifically, we assume that the rate (or reward) user i gets from channel j at time t is of the form $r_j(t)g_j(n_j(t))$ where $r_j(t)$ is the rate of channel j at time t, $n_j(t)$ is the number of users using channel j at time t, and g_j is the user independent channel interference function (CIF) for channel j. This model is richer than the previously used models [1]–[3] since $r_j(t)$ can represent environmental effects such as fading or primary user activity, while g_j captures interactions between users. We consider the following three cases.

In the first case (C1), each channel evolves as an i.i.d random process in time, the users do not know the channel statistics, nor the form of the interference, nor the total number of users present in the system, and no direct communication is allowed among users. A user can measure the the overall rate it gets from using a channel but cannot tell how much of it is due to the dynamically changing channel quality (i.e., what it would get if it were the only user) vs. interference from other users. In this case, we show that if all users follow the Exp3 algorithm [4] then the channel allocation converges to a set of pure Nash equilibria (PNE) of a congestion game defined by the CIFs and mean channel rates. In this case a socially optimal allocation cannot be ensured, as the set of PNE are of different quality, and in some cases the socially optimal allocation may not be a PNE.

In the second case (C2), each channel again evolves as an i.i.d random process in time, whose statistics are unknown to the user. However, the users now know the total number of users in the system, as well as the fact that the quantitative impact of interference is common to all users (i.e., user independent), though the actual form of the interference function is unknown. In other words the rate of channel j at time t is perceived by user i as $h(t, n_j(t))$ so user i cannot distinguish between components $r_j(t)$ and $g_j(n_j(t))$. Furthermore, users are now allowed minimal amount of communication when they happen to be in the same channel, specifically to find out the total number of simultaneous users of that channel. In this case we present a sample-mean based randomized learning policy that achieves socially optimal allocation as time goes to infinity, with a sub-linear regret over the time horizon with respect to the socially optimal.

In the third case (C3), as in case (C2) the users know the total number of users in the system, as well as the fact that the interference function is user independent and decreasing without knowing the actual form of the interference function. However, the channels are assumed to have constant, albeit unknown, rates. We show that even without any communication among users, there is a randomized learning algorithm that achieves the socially optimal allocation in finite time.

It's worth pointing out that in the settings outlined above, the users are *non-strategic*, i.e., each user simply follow a pre-set learning rule rather than playing a game. In this context it is reasonable to introduce minimal amount of communication among users and assume they may cooperate. It is possible that even in this case the users may not know their interference function but only the total rate they get for lack of better detecting capabilities (e.g., they may only be able to detect the total received SNR as a result of channel rate and user interference).

Online learning by a single user was studied by [5]–[8], in which sample-mean based index policies were shown to achieve logarithmic regret with respect to the best single-action policy without a priori knowledge of the statistics, and are order-optimal, when the rewards are given by an iid process. In [9]–[11] Markovian rewards are considered, with [11] focusing on *restless* reward processes, where a process continues to evolved according to a Markov chain regardless of the users' actions. In all these studies learning algorithms were developed to achieve logarithmic regret. Multi-user learning with iid reward processes have been studied in a dynamic spectrum context by [1], [2], [12], with a combinatorial structure adopted in [12], and two simple collision models in [1], [2]. In the first model, when there is more than one user on a channel all get zero reward, whereas in the second model one of them, selected randomly, gets all the reward while others get zero reward. These collision models do not capture more sophisticated communication schemes where the rate a user gets is a function of the received SNR of the form $g_j(n+1) = f_j(\frac{P_t}{N_o+(n-1)P_t}) =$ where P_t is the nominal transmit power of all users and N_o the noise. Moreover, in the above studies the socially optimal allocation is a rather simple one: it is the orthogonal allocation of users to the first M channels with the highest mean rewards. By contrast, we model a more general interference relationship among users, and the socially optimal allocation in this case may include sharing the same channels. In this case additional mechanisms may be needed for the learning algorithms to converge.

All of the above mentioned work assumes some level of communication between the users either at the beginning or during the learning. If we assume no communication between the users then achieving the socially optimal allocation seems very challenging in general. Then one may ask if it is possible to achieve some kind of equilibrium allocation. Kleinberg et. al. [3] showed that it is possible for the case when the channel rates are constant and the users do not know the interference functions. They show that when the users use *aggregate monotonic selection dynamics*, a variant of Hedge algorithm [13], the allocation converges to *weakly stable equilibria* which is a subset of Nash equilibria (NE) of the congestion game defined by g_j. They show that for almost all congestion games weakly stable equilibria is the same as PNE.

Other than the work described above [14] considers *spatial congestion games*, a generalization of congestion games and gives conditions under which there exists a PNE and best-response play converges to PNE. A mechanism design approach for socially optimal power allocation when users are strategic is considered in [15].

The organization of the remainder of this paper is as follows. In section II we present the notations and definitions that will be used throughout the paper. In sections III, IV, V we analyze the cases stated in (C1), (C2), (C3) and derive the results respectively.

II. Preliminaries

Denote the set of users by $\mathcal{M} = \{1, 2, \ldots, M\}$, and the set of channels $\mathcal{N} = \{1, 2, \ldots, N\}$. Time is slotted and indexed by $t = 1, 2, \ldots$ and a user can select a single channel at each time step t. Without loss of generality let $r_j(t) \in [0, 1]$ be the rate of channel j at time t, $g_j : \mathbb{N} \to [0, 1]$ be the interference function (CIF) on channel j where $g_j(n)$ represents the interference when there are n users on channel j. We express the rate of channel j seen by a user as $h_j(t) = r_j(t)g_j(n_j(t))$ when a user does not know the total number of users $n_j(t)$ using channel j at time t as in cases (C1) and (C3). When a user knows $n_j(t)$ at time t then we express the rate of channel j at time t as $h_{j,n_j(t)}(t) = r_j(t)g_j(n_j(t))$ as in case (C2). $\{r_j(t)\}_{t=1,2,\ldots}$ is generated by a non-negative iid process with mean $\mu_j \in [0, 1]$. Let $\mathcal{S}_i = \mathcal{N}$ be the set of feasible actions of user i and $\sigma_i \in \mathcal{S}_i$ be the action, i.e., channel selected by user i. Let $\mathcal{S} = \mathcal{S}_1 \bigotimes \mathcal{S}_2 \bigotimes \ldots \bigotimes \mathcal{S}_N = \mathcal{N}^M$ be the set of feasible action profiles and $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_M\} \in \mathcal{S}$ be the action profile of the users. We use the notation $\sigma(t)$ whenever the action profile is time dependent. Let $K_j(\sigma)$ be the set of users on channel j when the action profile is σ. Let $\mathcal{A}^* = \arg\max_{\sigma \in \mathcal{S}} \sum_{i=1}^{M} \mu_{\sigma_i} g_{\sigma_i}(K_{\sigma_i}(\sigma)) = \arg\max_{\sigma \in \mathcal{S}} \sum_{j=1}^{N} \mu_j K_j(\sigma) g_j(K_j(\sigma))$ be the set of socially optimal allocations and denote by σ^* any action profile that is in the set $\mathcal{A}^*$. Let v^* denote the socially optimal welfare, i.e., $v^* = \sum_{i=1}^{M} \mu_{\sigma_i^*} g_{\sigma_i^*}(K_{\sigma_i^*}(\sigma^*))$ and v_j^* denotes the payoff of channel j under the socially optimal allocation, i.e., $v_j^* = \mu_j g_j(K_j(\sigma^*))$ if $K_j(\sigma^*) \neq 0$ and $v_j^* = 1$ otherwise. Note that any permutation of actions in σ^* is also a socially optimal allocation since CIFs are user-independent.

Let π_i be the policy of user i. When π_i is deterministic, $\pi_i(t) = \sigma_i(t)$ is in general a function from all past observations and decisions of user i to the set of actions $\mathcal{S}_i$. When π_i is randomized, $\pi_i(t)$ generates a probability distribution over the set of actions $\mathcal{S}_i$ according to all past observations and decisions of user i from which the action at time t is sampled.

For any policy π, the regret at time n is $R(n) = nv^* - E\left[\sum_{t=1}^{n} \sum_{i=1}^{M} r_{\sigma_i(t)}(t) g_{\sigma_i(t)}(K_{\sigma_i(t)}(\sigma(t)))\right]$, where expectation is taken with respect to the random nature of the rates and the randomization of the policy. Note that for a deterministic policy expectation is only taken with respect to the random nature of the rates. For any randomized policy π_i, let $p_i(t) = [p_{i1}(t), p_{i2}(t), \ldots, p_{iN}(t)]$ be the mixed strategy of user i at time t, i.e., a probability distribution on $\{1, 2, \ldots, N\}$. For a profile of policies $\pi = [\pi_1, \pi_2, \ldots, \pi_M]$ for the users let $p(t) = [p_1(t)^T, p_2(t)^T, \ldots p_M(t)^T]$ be the profile of mixed strategies at time t, where $p_i(t)^T$ is the transpose of $p_i(t)$. Then $\sigma_i(t)$ is the action sampled from the probability distribution $p_i(t)$. The dependence of p to π is trivial and not shown in the notation.

III. Allocations Achievable with Exp3 Algorithm (Case 1)

We start by defining a congestion game. A congestion game [16], [17] is given by the tuple $(\mathcal{M}, \mathcal{N}, (\Sigma_i)_{i \in \mathcal{M}}, (h_j)_{j \in \mathcal{N}})$, where $\mathcal{M}$ denotes a set of players (users) , $\mathcal{N}$ a set of resources (channels), $\Sigma_i \subset 2^{\mathcal{N}}$ the strategy space of player i, and $h_j : \mathbb{N} \to \mathbb{Z}$ a payoff function associated with resource j, which is a function of the number of players using that resource. It

is well known that a congestion game has a potential function and the local maxima of the potential function corresponds to PNE, and every sequence of asynchronous improvement steps is finite and converges to PNE.

In this section we relate the strategy update rule of Exp3 [4] under assumptions (C1) to a congestion game. Exp3 as given in Figure 1 is a randomized algorithm consisting of an exploration parameter γ and weights w_{ij} that depend exponentially on the past observations where i denotes the user and j denotes the channel. Each user runs Exp3 independently but we explicitly note the user dependence because a user's action affects other users' updates.

Exp3 (for user i)

Initialize: $\gamma \in (0,1)$, $w_{ij}(t) = 1, \forall j \in \mathcal{N}$, $t = 1$
while $t > 0$ **do**

$$p_{ij}(t) = (1-\gamma)\frac{w_{ij}(t)}{\sum_{l=1}^{N} w_{il}(t)} + \frac{\gamma}{N}$$

  Sample $\sigma_i(t)$ from the distribution on $p_i(t) = [p_{i1}(t), p_{i2}(t), \ldots, p_{iN}(t)]$
  Play channel $\sigma_i(t)$ and receive reward $h_{\sigma_i(t)}(t)$
  for $j = 1, 2, \ldots, N$ **do**
    if $j = \sigma_i(t)$ **then**
      Set $w_{ij}(t+1) = w_{ij}(t) \exp\left(\frac{\gamma h_{\sigma_i(t)}(t)}{p_{ij}(t)N}\right)$
    else
      Set $w_{ij}(t+1) = w_{ij}(t)$
    end if
  end for
  $t = t + 1$
end while

Fig. 1. pseudocode of Exp3

At any time step before the channel rate and user actions are drawn from the corresponding distributions, let R_j denote the random variable corresponding to the reward of the jth channel. Let $G_{ij} = g_j(1 + K'_j(i))$ be the random variable representing the payoff user i gets from channel j where $K'_j(i)$ is the random variable representing the number of users on channel j other than user i. Let $U_{ij} = R_j G_{ij}$ and $\bar{u}_{ij} = E_j[E_{-i}[U_{ij}]]$ be the expected payoff to user i by using channel j where E_{-i} represents the expectation taken with respect to the randomization of players other than i, E_j represents the expectation taken with respect to the randomization of the rate of channel j. Since the channel rate is independent of users' actions $\bar{u}_{ij} = \mu_j \bar{g}_{ij}$ where $\bar{g}_{ij} = E_{-i}[G_{ij}]$.

Lemma 1: Under (C1) when all players use Exp3, the derivative of the continuous-time limit of Exp3 is the replicator equation given by

$$\xi_{ij} = \frac{1}{N}(\mu_j p_{ij}) \sum_{l=1}^{N} p_{il}(\bar{g}_{ij} - \bar{g}_{il})$$

The proof of this lemma is not given due to limited space. Lemma 1 shows that the dynamics of a user's probability distribution over the actions is a replicator equation which is commonly studied in evolutionary game theory [18], [19]. With this lemma we can establish the following theorem.

Theorem 1: For all but a measure zero subset of $[0,1]^{2N}$ from which μ_js and g_js are selected, when players use arbitrarily small γ in Exp3, the action profile converges to the set of PNE of the congestion game $(\mathcal{M}, \mathcal{N}, (\mathcal{S}_i)_{i\in\mathcal{M}}, (\mu_j g_j)_{j\in\mathcal{N}})$.

Proof: Here we will briefly explain the proof. For the complete proof see [3]. Defining the expected potential function to be the expected value of the potential function ϕ where expectation is taken with respect to the user's randomization one can show that the solutions of the replicator equation converges to the set of fixed points. Then the stability analysis using the Jacobian matrix yields that every stable fixed point corresponds to a Nash equilibrium. Then one can prove that for any stable fixed point the eigenvalues of the Jacobian must be zero. This implies that every stable fixed point corresponds to a *weakly*

stable Nash equilibrium strategy in the game theoretic sense. Then using tools from algebraic geometry one can show that every weakly stable Nash equilibrium is a pure Nash equilibrium of the congestion game.

We also need to investigate the error introduced by treating the discrete time update rule as a continuous time process. However, by taking γ infinitesimal we can approximate the discrete time process by the continuous time process. For a discussion when γ is not infinitesimal one can define *approximately stable equilibria* [3]. ■

The main difference between Exp3 and Hedge [3] is that in Exp3 players do not need to observe the payoffs from the channels that they do not play, whereas Hedge assumes complete observation. In addition to that, we considered the dynamic channel rates which is not considered in [3].

IV. An Algorithm for Socially Optimal Allocation With Sub-Linear Regret (Case 2)

In this section we propose an algorithm such that when users use this algorithm, the regret with respect to the socially optimal allocation is $O(n^{\frac{2M-1+2\gamma}{2M}})$ for $\gamma > 0$ arbitrarily small. Clearly this regret is sublinear and approaches linear as the number of users M increases. This means that the time average of the sum of the utilities of the players converges to the social optimal. Let $\mathcal{K} = \{k = (k_1, k_2, \ldots, k_N) : k_j \geq 0, \forall j \in \mathcal{N}, k_1 + k_2 + \ldots + K_N = M\}$ denote an allocation of M users to N channels. Note that this allocation gives only the number of users on each channel. It does not say anything about which user uses which channel. We assume the socially optimal allocation is unique up to permutations so $k^* = \arg\max_{k\in\mathcal{K}} \sum_{j=1}^{N} \mu_j k_j g_j(k_j)$ is unique. We also assume the following stability condition of the socially optimal allocation. Let $v_j(k_j) = \mu_j g_j(k_j)$. Then the stability condition says that $\arg\max_{k\in\mathcal{K}} \sum_{j=1}^{N} \hat{v}_j(k_j) = k^*$ if $|\hat{v}_j(k) - v_j(k)| \leq \epsilon, \forall k \in \{1, 2, \ldots, M\}, \forall j \in \mathcal{N}$, for some $\epsilon > 0$. Let $T^i_{j,k}(t)$ be the number of times user i used channel j and observed k users on it up to time t. We refer to the tuple (j, k) as an arm. Let $n^i_{j,k}(t')$ be the time of the t'th observation of user i from arm (j, k). Let $u^i_{j,k}(n)$ be the sample mean of the rewards from arm (j, k) seen by user i at the end of the nth play of arm (j, k) by user i, i.e., $u^i_{j,k}(n) = \frac{h_{j,k}(n^i_{j,k}(1))+\ldots+h_{j,k}(n^i_{j,k}(n))}{n}$.

The following will be useful in the proof of the main theorem of this section.

Lemma 2: Let $X_i, i = 1, 2, \ldots$ be a sequence of independent Bernoulli random variables such that X_i has mean q_i with $0 \leq q_i \leq 1$. Let $\bar{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i$, $\bar{q}_k = \frac{1}{k}\sum_{i=1}^{k} q_i$. Then for any constant $\epsilon \geq 0$ and any integer $n \geq 0$,

$$P\left(\bar{X}_n - \bar{q}_n \leq -\epsilon\right) \leq e^{-2n\epsilon^2}. \tag{1}$$

Proof: The result follows from symmetry and [20]. ■

Lemma 3: For $p > 0, p \neq 1$

$$\frac{(n+1)^{1-p} - 1}{1-p} < \sum_{t=1}^{n} \frac{1}{t^p} < 1 + \frac{n^{1-p} - 1}{1-p} \tag{2}$$

Proof: See [21]. ■

We introduce a randomized algorithm RLA in Figure 2.

Theorem 2: When all players use RLA the regret with respect to the socially optimal allocation is $O(n^{\frac{2M-1+2\gamma}{2M}})$ where $\gamma > 0$ can be arbitrarily small.

Proof: (Sketch) Let $H(t)$ be the event that at time t there exists at least one user that computed the socially optimal allocation wrongly. Let ω be a sample path. Let $\epsilon^i_{j,l}(t) = \sqrt{\frac{a \ln t}{T^i_{j,l}(t)}}$. Then for $a > 0$.

$$E\left[\sum_{t=1}^{n} I(\omega \in H(t))\right] \leq \sum_{t=1}^{n}\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{l=1}^{M}\left(P\left(|u^i_{j,l}(T^i_{j,l}(t)) - v_j(l)| \geq \epsilon^i_{j,l}(t)\right) + P\left(T^i_{j,l}(t) < \frac{a \ln t}{\epsilon^2}\right)\right). \tag{3}$$

We have

$$P\left(|u^i_{j,l}(T^i_{j,l}(t)) - v_j(l)| \geq \epsilon^i_{j,l}(t)\right) = \frac{2}{t^{2a}}, \tag{4}$$

where (4) follows from the Chernoff-Hoeffding inequality.

RLA (for user i)

Initialize: $0 < \gamma << 1$, $\hat{h}^i_{j,k}(1) = 0, T^i_{j,k}(1) = 0, \forall j \in \mathcal{N}, k \in \mathcal{M}$, $t = 1$, sample $\sigma_i(1)$ uniformly from $\mathcal{N}$.
while $t > 0$ **do**
  play channel $\sigma_i(t)$, observe $l(t)$ the total number of players using channel $\sigma_i(t)$ and reward $h_{\sigma_i(t),l(t)}(t)$.
  Set $T^i_{\sigma_i(t),l(t)}(t+1) = T^i_{\sigma_i(t),l(t)}(t) + 1$ and $T^i_{j,l}(t+1) = T^i_{j,l}(t)$ for $(j,l) \neq (\sigma_i(t), l(t))$.
  Set $\hat{h}^i_{\sigma_i(t),l(t)}(t+1) = \frac{T^i_{\sigma_i(t),l(t)}(t)\hat{h}^i_{\sigma_i(t),l(t)}(t) + h_{\sigma_i(t),l(t)}(t)}{T^i_{\sigma_i(t),l(t)}(t+1)}$ and $\hat{h}^i_{j,l}(t+1) = \hat{h}^i_{j,l}(t)$ for $(j,l) \neq (\sigma_i(t), l(t))$.
  Calculate the socially optimal allocation $k^{i*}(t+1) = \arg\max_{k \in \mathcal{K}} \sum_{j=1}^{N} k_j u^i_{j,k_j}(t+1)$.
  Let $\theta^{*i}(t+1)$ be the set of channels used by at least one user in $k^{*i}(t+1)$.
  Draw i_t randomly from Bernoulli distribution with $P(i_t = 1) = \frac{1}{t^{(1/2M)-\gamma/M}}$
  if i_t=0 **then**
    if $\sigma_i(t) \in \theta^*(t+1)$ and $l(t) = k^{i*}_j(t+1)$ **then**
      $\sigma_i(t+1) = \sigma_i(t)$
    else
      $\sigma_i(t+1)$ is selected uniformly at random from the channels in $\theta^*(t+1)$.
    end if
  else
    Draw $\sigma_i(t+1)$ uniformly at random from $\mathcal{N}$
  end if
  $t = t + 1$
end while

Fig. 2. Pseudocode of RLA

Now we will bound $P\left(T^i_{j,l}(t) < \frac{a \ln t}{\epsilon^2}\right)$. Let $TR^i_{j,l}(t)$ be the number of time steps in which player i played channel j and observed l users on channel j in the time steps where all players randomized up to time t. Then

$$P\left(T^i_{j,l}(t) < \frac{a \ln t}{\epsilon^2}\right) \leq P\left(TR^i_{j,l}(t) < \frac{a \ln t}{\epsilon^2}\right) \tag{5}$$

Then for $t > \tau(M, N, \epsilon, \gamma, \gamma', a)$ where $0 < \gamma' < \gamma$

$$P\left(TR^i_{j,l}(t) < \frac{a \ln t}{\epsilon^2}\right) \leq \frac{1}{t^2} \tag{6}$$

Where (6) follows from Lemma 2 and 3. Let $a = 1$. Then continuing form (3) by substituting (4) and (6)

$$E\left[\sum_{t=1}^{n} I(\omega \in H(t))\right] \leq M^2 N \left(\tau(M, N, \epsilon, \gamma, \gamma', 1) + 3\sum_{t=1}^{n} \frac{1}{t^2}\right). \tag{7}$$

Thus we proved that the expected number of time steps in which there exists at least one user that computed the socially optimal allocation incorrectly is finite. Now we consider only the randomizations that comes from i_t. Because the users randomize with probability $\frac{1}{t^{1/2M-\gamma/M}}$ at each t, the expected number of time steps in which there exists some player randomizing up to time n is

$$O(n^{\frac{2M-1+2\gamma}{2M}}). \tag{8}$$

Note that players can choose γ arbitrarily small with a tradeoff, i.e., increasing the regret due to $\tau(M, N, \epsilon, \gamma, \gamma', 1)$.

Now consider the following worst case analysis. We classify the time steps into two types: *good* time steps in which all the players know the socially optimal allocation correctly and none of the players randomize except for the randomizations done for settling down to the socially optimal allocation, and *bad* time steps in which there exists a player that does not know the socially optimal allocation correctly or there is a player that randomizes except for the randomizations done for settling down to the socially optimal allocation. The number of bad time steps in which there exists a player that does not know the socially optimal allocation correctly is finite while the number of time steps in which there is a player that randomizes excluding the randomizations done for settling down to the socially optimal allocation is $O(n^{\frac{2M-1+2\gamma}{2M}})$. The worst case is when each bad

step is followed by a good step. Then from this good step the expected number of times it takes to settle down to the socially optimal allocation is $\left(1-\frac{1}{\binom{M+z^*-1}{z^*-1}}\right)/\left(\frac{1}{\binom{M+z^*-1}{z^*-1}}\right)$ where z^* is the number of channels which has at least one user in the socially optimal allocation. Assuming in the worst case the sum of the utilities of the players is 0 when they are not playing the socially optimal allocation we have

$$R(n) \leq \frac{1-\frac{1}{\binom{M+z^*-1}{z^*-1}}}{\frac{1}{\binom{M+z^*-1}{z^*-1}}}\left(M^2N\left(\tau(M,N,\epsilon,\gamma,\gamma',1)+3\sum_{t=1}^{n}\frac{1}{t^2}\right)+O(n^{\frac{2M-1+2\gamma}{2M}})\right)=O(n^{\frac{2M-1+2\gamma}{2M}})$$

■

Note that we mentioned earlier, under a classical multi-armed bandit problem approach as cited before [1], [2], [5], [6], [9]–[11], a logarithmic regret $O(\log n)$ is achievable. The fundamental difference between these studies and the problem in the present paper is the following. Assume that at time t user i selects channel j. This means that i selects to observe an arm from the set $\{(j,k): k \in \mathcal{M}\}$ but the arm assigned to i is selected from this set depending on the choices of other players.

Also note that in RLA a user computes the socially optimal allocation according to its estimates at each time step. This could pose significant computational effort since integer programming is NP-hard in general. However, by exploiting the stability condition on the socially optimal allocation a user may reduce the number of computations; this is a subject of future research.

V. An Algorithm for Socially Optimal Allocation (Case 3)

In this section we assume that $g_j(n)$ is decreasing in n for all $j \in \mathcal{N}$. For simplicity we assume that the socially optimal allocation is unique up to the permutations of σ^*. When this uniqueness assumption does not hold we need a more complicated algorithm to achieve the socially optimal allocation. All users use the Random Selection (RS) algorithm defined in Figure 3. RS consists of two phases. Phase 1 is the learning phase where the user randomizes to learn the interference functions. Let $B_j(t)$ be the set of distinct payoffs observed from channel j up to time t. Then the payoffs in set $B_j(t)$ can be ordered in a decreasing way with the associated indices $\{1,2,\ldots,|B_j(t)|\}$. Let $O(B_j(t))$ denote this ordering. Since the CIFs are decreasing, at the time $|B_j(t)|=M$, the user has learned g_j. At the time $|\cup_{j=1}^{N} B_j(t)|=MN$, the user has learned all CIFs. Then, the user computes $\mathcal{A}^*$ and phase 2 of RS starts where the user randomizes to converge to the socially optimal allocation.

Random Selection (RS)

Initialize: $t=1$, $b=0$, $B_j(1)=\emptyset, \forall j \in \mathcal{N}$, sample $\sigma_i(1)$ from the uniform distribution on $\mathcal{N}$
Phase 1
while $b < MN$ **do**
  if $h_{\sigma_i(t)}(t) \notin B_{\sigma_i(t)}(t)$ **then**
    $B_{\sigma_i(t+1)(t+1)} \leftarrow O(B_{\sigma_i(t)}(t) \cup h_{\sigma_i(t)}(t))$
    $b = b+1$
  end if
  Sample $\sigma_i(t+1)$ from the uniform distribution on $\mathcal{N}$
  $t = t+1$
end while
find the socially optimal allocation σ^*
Phase 2
while $b \geq MN$ **do**
  if $h_{\sigma_i(t)}(t) < v^*_{\sigma_i(t)}$ **then**
    Sample $\sigma_i(t+1)$ from the uniform distribution on $\mathcal{N}$
  else
    $\sigma_i(t+1) = \sigma_i(t)$
  end if
  $t = t+1$
end while

Fig. 3. pseudocode of RS

Theorem 3: Under the assumptions of (C3) if all players use RS algorithm to choose their actions, then the expected time to converge to the socially optimal allocation is finite.

Proof: Let T_{OPT} denote the time the socially optimal allocation is achieved, T_L be the time when all users learn all the CIFs, T_F be the time it takes to reach the socially optimal allocation after all users learn all the CIFs. Then $T_{OPT} = T_L + T_F$ and $E[T_{OPT}] = E[T_L] + E[T_F]$. We will bound $E[T_L]$ and $E[T_F]$. Let T_i be the first time that i users have learned the CIFs. Let $\tau_i = T_i - T_{i-1}, i = 1, 2, \ldots, M$ and $T_0 = 0$. Then $T_L = \tau_1 + \ldots + \tau_M$. Define a Markov chain over all N^M possible configurations of M users over N channels based on the randomization of the algorithm. This Markov chain has a time dependent stochastic matrix which changes at times $T_1, T_2, \ldots, T_M$. Let $P_{T_0}, P_{T_1}, \ldots, P_{T_M}$ denote the stochastic matrices after the times $T_0, T_1, \ldots, T_M$ respectively. This Markov chain is irreducible at all times up to T_M and is reducible with absorbing states corresponding to the socially optimal allocations after T_M. Let $\hat{T}_1, \hat{T}_2, \ldots \hat{T}_M$ be the times that all configurations are visited when the Markov chain has stochastic matrices $P_{T_0}, P_{T_1}, \ldots, P_{T_{M-1}}$ respectively. Then because of irreducibility and finite states $E[\hat{T}_i] < z_1, i = 1, \ldots, M$ for some constant $z_1 > 0$. Since $\tau_i \leq \hat{T}_i, i = 1, \ldots, M$ a.s. we have $E[T_L] < M z_1$. For the Markov chain with stochastic matrix P_{T_M} all the configurations that do not correspond to the socially optimal allocation are transient states. Since starting from any transient state the mean time to absorbtion is finite $E[T_F] < z_2$, for some constant $z_2 > 0$. ■

REFERENCES

[1] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple players: Learning under competition," in *Proc. of IEEE INFOCOM*, March 2010.

[2] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players, http://arxiv.org/abs/0910.2065."

[3] R. Kleinberg, G. Piliouras, and E. Tardos, "Multiplicative updates outperform generic no-regret learning in congestion games," in *Annual ACM Symposium on Theory of Computing (STOC)*, 2009.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, pp. 48–77, 2002.

[5] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[6] V. Anantharam, P. Varaiya, and J. . Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Trans. Automat. Contr.*, pp. 968–975, November 1987.

[7] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, December 1995.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, p. 235256, 2002.

[9] V. Anantharam, P. Varaiya, and J. . Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards," *IEEE Trans. Automat. Contr.*, pp. 977–982, November 1987.

[10] C. Tekin and M. Liu, "Online algortihms for the multi-armed bandit problem with markovian rewards," in *Allerton Conference*, 2010.

[11] ——, "Online learning in opportunistic spectrum access: A restless bandit approach," in *30th IEEE International Conference on Computer Communications (INFOCOM)*, April 2011.

[12] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation," in *IEEE Symp. on Dynamic Spectrum Access Networks (DySPAN)*, April 2010.

[13] Y. Freund and R. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behaviour*, vol. 29, pp. 79–103, 1999.

[14] S. Ahmad, C. Tekin, M. Liu, R. Southwell, and J. Huang, "Spectrum sharing as spatial congestion games," *http://arxiv.org/abs/1011.5384*.

[15] A. Kakhbod and D. Teneketzis, "Power allocation and spectrum sharing in cognitive radio networks with strategic users," in *49th IEEE Conference on Decision and Control (CDC)*, December 2010.

[16] R. Rosenthal, "A class of games possessing pure-strategy nash equilibria," *International Journal of Game Theory*, vol. 2, pp. 65–67, 1973.

[17] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 0044, pp. 124–143, 1996.

[18] J. M. Smith, "Evolution and the theory of games," *Cambridge University Press*, 1982.

[19] W. H. Sandholm, "Population games and evolutionary dynamics," *Manuscript*, 2008.

[20] J. S. D.W. Turner, D.M. Young, "A kolmogorov inequality for the sum of independent bernoulli random variables with unequal means," *Statistics and Probability Letters*, vol. 23, pp. 243–245, 1995.

[21] E. Chlebus, "An approximate formula for a partial sum of the divergent p-series," *Applied Mathematics Letters*, vol. 22, pp. 732–737, 2009.