Cooperative Game in Dynamic Spectrum Access with Unknown Model and Imperfect Sensing

Keqin Liu and Qing Zhao

***Abstract*—We consider dynamic spectrum access where distributed secondary users search for spectrum opportunities without knowing the primary traffic statistics. In each slot, a secondary transmitter chooses one channel to sense and subsequently transmit if the channel is sensed as idle. Sensing is imperfect, *i.e.,* an idle channel may be sensed as busy and vice versa. Without centralized control, each secondary user needs to independently identify the channels that offer the most opportunities while avoiding collisions with both primary and other secondary users. We address the problem within a cooperative game framework, where the objective is to maximize the throughput of the secondary network under a constraint on the collision with the primary system. The performance of a decentralized channel access policy is measured by the system regret, defined as the expected total performance loss with respect to the optimal performance in the ideal scenario where the traffic load of the primary system on each channel is known to all secondary users and collisions among secondary users are eliminated through centralized scheduling. By exploring the rich communication structure of the problem, we show that the optimal system regret has the same logarithmic order as in the centralized counterpart with perfect sensing. A decentralized policy is constructed to achieve the logarithmic order of the system regret. In a broader context, this work addresses imperfect reward observation in decentralized multi-armed bandit problems.**

***Index Terms*—Dynamic spectrum access, cognitive radio, cooperative game, distributed learning, imperfect sensing, system regret, decentralized multi-armed bandit.**

I. Introduction

We study a distributed learning problem in the context of dynamic spectrum access (DSA) under a noisy environment [1]. There are multiple secondary users independently searching for idle channels temporarily unused by the primary system. The traffic load of the primary system on each channel is *unknown* to the secondary users. At the beginning of each time slot, each secondary user chooses one channel to sense and subsequently transmit if the channel is sensed as idle. Due to noise and fading, sensing is *imperfect*: an idle channel can be sensed as busy and *vice versa*. As a consequence, a secondary user may transmit on a busy channel and causes a collision to the primary system (referred to as a *primary collision*). The secondary users are *decentralized*: they make channel access decisions solely based on local observations without information exchange or centralized control. A *secondary collision* happens when multiple secondary users transmit on the same idle channel. Under both primary collisions and secondary collisions, all transmissions involved fail. We address the problem within a cooperative game framework, where the objective is to maximize the long-term throughput of the secondary network under a constraint on the maximum allowable probability of primary collisions.

A. Learning under Competition and from Corrupted Data

In the case of a single secondary user, the above DSA problem can be formulated as a Multi-Armed Bandit (MAB) problem pioneered by Lai and Robbins in 1985 within a non-Bayesian framework [2]. In an MAB problem, a player selects one out of a given set of arms to play to accrue reward at each time. Each arm, when played, offers i.i.d. reward over time with unknown statistics. The player can improve its selection over time by learning from past reward observations which are assumed to be perfect. The performance of an arm selection policy is measured by *regret* defined as the total reward loss with respect to the case with known reward models. The essence of the problem is the well-known tradeoff between exploitation (*i.e.,* selecting the arm appearing to be the best based on past reward observations) and exploration (selecting an arm to learn its reward statistic to minimize future mistakes). It has been shown by Lai and Robbins in [2] that the optimal regret has a logarithmic order with time. An optimal policy was constructed under a general reward model to achieve the optimal regret[1]. In [3], Anantharam *et al.* extended Lai and Robbins's results to the case of multiple plays where the player chooses M arms to play at each time [3].

Even with imperfect sensing, the single-user DSA problem can be formulated as an MAB with a proper measure for the goodness of an arm. Specifically, the goodness of a channel is determined by how likely the secondary user can catch an opportunity (*i.e.,* the channel is idle and is correctly detected as such). Consequently, the reward offered from a channel can be measured by whether the user successfully transmits in the channel, which is perfectly observed. The problem thus falls into the general MAB model that considers perfect reward observations.

With multiple distributed secondary users, however, imperfect sensing significantly complicates the problem. The

Manuscript received August 17, 2011; revised October 28, 2011; accepted January 6, 2012. The associate editor coordinating the review of this paper and approving it for publication was N. Devroye.

This work was supported by the National Science Foundation under Grant CCF-0830685 and by the Army Research Office under Grant W911NF-08-1-0467. Part of this work was presented at the 44th Asilomar Conference on Signals, Systems, and Computers, November 2010.

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA (e-mail: {kqliu, qzhao}@ucdavis.edu).

Digital Object Identifier 10.1109/TWC.2012.020812.111547

[1]Note that the regret is a finer performance measure than the average reward. Any sub-linear regret leads to the same maximum average reward achieved in the case of known reward model.

main difficulty is that each secondary user cannot distinguish between secondary collisions caused by competition and primary collisions caused by sensing errors. A failed transmission due to secondary collisions does not reflect the channel quality. If a secondary user learns the channel quality from the history of successful transmissions (as in the single-user case), the best channels may not be correctly identified. In other words, collisions among secondary users affect not only the immediate reward but also the learning ability at each colliding user, which further degrades the system long-term throughput.

B. Main Results

In this paper, we formulate the multi-user DSA with imperfect sensing as a variant of decentralized MAB with multiple players to take into account the imperfect reward observation. The performance measure of a decentralized channel access policy is given by *system regret*, defined as the expected total throughput loss with respect to the optimal performance in the ideal case where the traffic load of the primary system on each channel is known to all secondary users and collisions among the secondary users are eliminated through centralized scheduling. Under the cooperative game framework, the objective of the secondary users is to minimize the rate that the system regret grows with time (*i.e.,* maximize the rate that the network throughput converges to the maximum). We show that the optimal system regret has the same logarithmic order as in the classic centralized MAB. Referred to as SLCD (Synchronized Learning under Corrupted Data), the proposed decentralized policy achieves the optimal logarithmic order of the system regret. Under this policy, the network throughput achieves the same maximum throughput attainable in the ideal case with known models and perfect scheduling. Furthermore, the policy ensures fairness among all secondary users, *i.e.,* each user achieves the same local throughput at the same rate.

The basic approach in the SLCD policy is to ensure that learning at each secondary user is carried out using only reliable information on the channel quality. This information is conveyed through the detection history of the primary traffic. The main challenge is that due to imperfect sensing, the detection outcomes at each secondary transmitter and receiver may disagree, *e.g.,* a channel may be detected as idle at the transmitter but busy at the receiver. If both the transmitter and receiver learn from their own detection outcomes, they may have different channel selections. Without a dedicated control channel between each transmitter and receiver, a natural but nontrivial question is how to achieve synchronized and efficient channel selection at each transmitter and receiver. While each transmitter and receiver can exploit idle channels to exchange control information to coordinate, achieving an efficient synchronization mechanism is nontrivial. Beyond the throughput sacrifice due to the control information exchange, the synchronization requirement also yields a constrained channel selection and observation sequence. Since the observation sequence determines the learning efficiency, the question here is whether the optimal tradeoff between exploitation and exploration under the unconstrained scenario can still be achieved. We show that under SLDC, the learning mistakes can be bounded within the same logarithmic order as in the unconstrained MAB. Meanwhile, the incurred control overhead is also bounded at the same order, leading to the optimal logarithmic system regret.

C. Related Work

This work builds upon our prior work on decentralized MAB with a perfect observation model [4], where the optimal system regret was shown to have the same logarithmic order as in the classic centralized MAB [2], [3]. With imperfect sensing, however, the multi-user DSA problem is significantly more complex as detailed in Sec. I-A. The result in this paper shows that for this class of decentralized MAB with imperfect observations, the system can still achieve the logarithmic order of the regret.

Under the assumption of perfect sensing, the multi-user DSA problem under unknown channel model was studied in [5]–[7]. In [5], a heuristic distributed policy based on histogram estimation of the unknown parameters was proposed to maximize the average reward. The system regret minimization was not addressed. In [6], [7], distributed policies that achieve the optimal logarithmic order of the system regret were developed based on UCB1 proposed in [8]. Specifically, a randomized strategy was proposed in [6] to orthogonalize users into the best channels without pre-agreement. In [7], UCB1 was extended to targeting at the mth $(1 < m < N)$ best channel and the distributed polices under both prioritized and fair access scenarios were proposed.

The above studies on multi-user DSA focus on the cooperative game framework where secondary users have a common global objective. In [9]–[12], a non-cooperative game framework was adopted where secondary users are considered *selfish*. In [9], a direct transmission model was considered where each secondary user transmits on the selected channel without sensing the primary traffic. Each user solely aims to maximize its local throughput. It was shown that the system converges to a Nash equilibrium when each user adopts the single-user policy proposed in [13]. Specifically, as time goes, users will be asymptotically orthogonalized to the M best channels and the system achieves the maximum long-term throughput without fairness. For the sensing-before-transmission model considered in this paper, each user can efficiently identify the best channel and severe collisions on the channel may happen when users are non-cooperative. Consequently, both the system and the individual performance suffer. In this paper, we show that if users are cooperative, the system can achieve an order-optimal and fair Nash equilibrium (in terms of regret minimization). In [10]–[12], transmission strategies for non-cooperative secondary users are analyzed under known channel interference and noise models, where the system Nash equilibria are characterized.

In this paper, we focus on a memoryless channel occupancy model commonly adopted in the literature of classic MAB [2], [3], [8]. In [14]–[18], a Markovian channel model with unknown transition probabilities was addressed under the perfect sensing scenario. Specifically, in [14], a single-user policy was constructed to achieve a regret with an order arbitrarily close to logarithmic when channels are governed by stochastically

identical two-state Markov chains. Under a weak definition of regret, single-user policies were proposed in [15], [16] to achieve a logarithmic order of the weak regret. The extension to the case of multiple users was addressed in [17], [18], where a distributed policy was constructed to achieve a logarithmic order of the weak regret. All these studies, however, assume a perfect observation model. The extension of the results in this paper to the Markovian model will be addressed in Sec. VI.

II. Network Model

Consider a spectrum consisting of N independent but nonidentical channels and M distributed secondary users. We consider the nontrivial scenario that the number of users is less than the number of channels[2]. This scenario is suitable for the cognitive radio network since the secondary users are not restricted to a particular frequency band and can search opportunities among a large set of channels. Furthermore, we only need to consider the group of secondary users that can interfere on the same set of channels. Let $\mathbf{S}(t) = [S_1(t), \cdots, S_N(t)] \in \{0,1\}^N$ $(t \geq 1)$ denote the system state, where $S_n(t)$ is the state of channel n in slot t. For simplicity, we assume that $S_n(t)$ evolves as an i.i.d. Bernoulli process[3] on the state space $\{0$ (busy), 1 (idle)$\}$ with unknown mean $\theta_n \in (0,1)$. The unknown mean $\theta_n \in (0,1)$ represents the unknown traffic load of the primary system on channel n, and the channel with a higher mean has a lighter traffic load.

In slot t, a secondary user (say *user* m $(1 \leq m \leq M)$) chooses a sensing action $a_m(t) \in \{1, \cdots, N\}$ that specifies the channel (say, *channel* n) to sense based on its observation and decision history. Based on the sensed signals, the user detects the channel state, which can be considered as a binary hypothesis test:

$$\mathcal{H}_0: \ S_n(t) = 1 \text{ (idle)} \ \ vs. \ \ \mathcal{H}_1: \ S_n(t) = 0 \text{ (busy)}.$$

The performance of channel state detection is characterized by the receiver operating characteristics (ROC) which relates the probability of false alarm ϵ to the probability of miss detection δ:

$$\epsilon \triangleq \Pr\{\text{decide } \mathcal{H}_1 | \mathcal{H}_0 \text{ is true}\}, \quad \delta \triangleq \Pr\{\text{decide } \mathcal{H}_0 | \mathcal{H}_1 \text{ is true}\}.$$

If the detection outcome is $\mathcal{H}_0$, the user accesses the channel for data transmission. The design should be subject to a constraint on the probability of accessing a busy channel, which causes interference to the primary system and also data loss of the user. Specifically, the probability $\mathcal{P}_n(t)$ of collision caused by the user and perceived by the primary system in any channel and slot is capped below a predetermined threshold ζ, *i.e.,*

$$\mathcal{P}_n(t) \triangleq \Pr(\text{decide } \mathcal{H}_0 | S_n(t) = 0) = \delta \leq \zeta, \quad \forall \ n, \ t.$$

[2]In the case of $M \geq N$, there is no longer an issue of learning and identifying the best channels since all channels will need to be utilized, and a zero system regret can be easily achieved by letting N users fully occupy the N channels.

[3]It is straightforward to extend the results to general i.i.d. processes.

We should set the miss detection probability $\delta = \zeta$ as the detector operating point to minimize the false alarm probability ϵ. If multiple secondary users decide to transmit over the same channel, they collide and no one can transmit successfully. In other words, a secondary user can transmit data successfully if and only if the chosen channel is idle, detected correctly, and no collision happens. Since failed transmissions may occur, acknowledgements (ACKs) are necessary to ensure guaranteed delivery. Specifically, when a secondary receiver successfully receives a packet over a channel, it sends an acknowledgement to the transmitter over the same channel at the end of the slot. Otherwise, the receiver does nothing, *i.e.,* a NAK is defined as the absence of an ACK. We assume that acknowledgements are received without error since acknowledgements are always transmitted over idle channels without collisions.

The DSA model considered in this paper and the associated results find applications in more general wireless communication networks including opportunistic transmission over fading channels, downlink scheduling in cellular systems, and resource-constrained jamming and anti-jamming.

III. A Decentralized MAB Formulation

We formulate the DSA problem as a decentralized MAB with imperfect observations. In a general decentralized MAB, there are M players independently playing N arms with unknown reward statistics. At each time, each player selects one arm to play and accrue certain amount of reward from this arm. Under a general observation model, the player may not be able to observe the actual reward offered by the selected arm. The DSA problem is a special class of decentralized MAB by considering secondary users as players and sensing a channel as playing an arm. The imperfect sensing scenario yields the imperfect observation of the actual channel state (*i.e.,* reward). A distinctive property of this class of decentralized MAB is that each user consists of one transmitter and receiver where they need to choose the same channel for data transmission at each time.

Under the synchronization constraint on each transmitter and receiver, we define a local policy π_m for user m as a sequence of functions $\pi_m = \{\pi_m(t)\}_{t \geq 1}$, where $\pi_m(t)$ maps user m's local information that is available to both its transmitter and receiver to the sensing action $a_m(t)$ in slot t. The decentralized policy π is thus given by the concatenation of the local policy for each user: $\pi = [\pi_1, \cdots, \pi_M]$. Define immediate reward $Y(t)$ as the total number of successful transmissions of the data (instead of the control information for synchronization) by all users in slot t:

$$Y(t) = \Sigma_{n=1}^{N} \mathbb{I}'_n(t) S_n(t),$$

where $\mathbb{I}'_n(t)$ is the indicator function that equals to 1 if channel n is accessed by only one user and used for transmitting the data (instead of the control information), and 0 otherwise.

Let $\Theta = (\theta_1, \theta_2, \cdots, \theta_N)$ be the unknown parameter set and σ a permutation such that[4] $\theta_{\sigma(1)} > \theta_{\sigma(2)} > \cdots > \theta_{\sigma(N)}$. The

[4]For the simplicity of the presentation, we assume that there is no tie in channel mean.

performance measure of a decentralized policy π is defined as the system regret

$$R_T^\pi(\Theta) = T\Sigma_{n=1}^M(1-\epsilon)\theta_{\sigma(n)} - \mathbb{E}_\pi[\Sigma_{t=1}^T Y(t)].$$

Note that $T\Sigma_{n=1}^M(1-\epsilon)\theta_{\sigma(n)}$ is the maximum expected total reward over T slots under the ideal scenario that the parameter set $\Theta = (\theta_1, \cdots, \theta_N)$ is known and users are centralized (thus the M best channels are sensed in each slot[5]).

We point out that the system regret is always growing with time since users can never perfectly identify the best channels. Under a cooperative game framework, the objective of secondary users is to minimize the rate at which $R_T(\Theta)$ grows with time T under any parameter set Θ by choosing the optimal decentralized policy π^*. Note that the system regret is a finer performance measure than the long-term throughput. All policies with a sub-linear system regret would achieve the maximum long-term throughput. However, the difference in their performance measured by the expected total bits of transmitted data over a time horizon of length T can be arbitrarily large as T grows. It is thus of great interest to characterize the minimum system regret and construct policies optimal under this finer performance measure.

Next, we show that the minimum system regret has the same logarithmic order with time as in the classic MAB with a single user and perfect sensing considered in [2], [3], [8].

Theorem 1: The optimal order of the system regret is logarithmic with time, *i.e.,* for an optimal decentralized policy π^*, we have, $\forall\ \Theta$,

$$L(\Theta) = \liminf_{T\to\infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \le \limsup_{T\to\infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} = U(\Theta) \quad (1)$$

for some constants $L(\Theta)$ and $U(\Theta)$ that depend on Θ.

Proof: To prove the lower bound, we consider a genie-aided system where secondary users are centralized and the synchronization constraint on each pair of transmitter and receiver is removed from consideration. Note that the channel parameters remain unknown to all users in the genie-aided system. It is easy to see that the problem is equivalent to the one with a single user that can sense M channels simultaneously in each slot. For simplicity, we focus on the latter one. In each slot, the user obtains two types of observations from each chosen channel: the detection outcome and the ACK/NAK. In Lemma 1, we show that the system regret in the genie-aided system is at least logarithmic with time. The proof is thus completed by noticing that the minimum system regret in the problem at hand is lower bounded by the one in the genie aided system.

Lemma 1: Let $\tilde{R}_T^\pi(\Theta)$ denote the system regret under a policy π in the genie-aided system over T slots. If $\tilde{R}_T^\pi(\Theta) = o(T^c)\ \forall\ \Theta$ and $\forall\ c > 0$, then, for any Θ,

$$\liminf_{T\to\infty} \frac{\tilde{R}_T^\pi(\Theta)}{\log T} \ge (1-\epsilon)\Sigma_{n:\ \theta_n<\theta_{\sigma(M)}} \frac{\theta_{\sigma(M)} - \theta_n}{G(\theta_n, \theta_{\sigma(M)})}, \quad (2)$$

where

$$\begin{aligned} G(\theta_i, \theta_j) &= (\epsilon\theta_i + (1-\delta)(1-\theta_i)) \log \frac{\epsilon\theta_i + (1-\delta)(1-\theta_i)}{\epsilon\theta_j + (1-\delta)(1-\theta_j)} \\ &\quad + \delta(1-\theta_i)\log\frac{\delta(1-\theta_i)}{\delta(1-\theta_j)} + (1-\epsilon)\theta_i \log\frac{(1-\epsilon)\theta_i}{(1-\epsilon)\theta_j} \end{aligned}$$

is the K-L distance between two joint distributions of the detection outcome and the ACK/NAK parameterized by θ_i and θ_j, respectively.

The proof of Lemma 1 follows a similar line to that of Theorem 3.1 in [3] by combining the detection outcome and ACK/NAK as a single observation vector of an arm.

For the upper bound, we show that their exists a decentralized policy that achieves logarithmic order of the system regret. See Sec. IV for details. ■

IV. An Order-Optimal Decentralized Policy

In this section, we establish a decentralized SLCD (Synchronized Learning under Corrupted Data) policy to achieve the optimal logarithmic order of the system regret while ensuring the fairness among all secondary users.

A. The General Structure

The general structure of the SLCD policy is based on a Time Division Fair Sharing (TDFS) of the M best channels among the M distributed users. The TDFS structure was first proposed in [4] under the scenario of perfect sensing. Due to the imperfect sensing of the channel state and the synchronization constraint, extending the TDFS framework to the problem at hand is nontrivial. Specifically, compared to perfect sensing, the channel sensing and observation sequence is constrained by the synchronization requirement and extracting reliable and sufficient information for efficient learning becomes more difficult (see Sec. IV-D for details).

Under the TDFS structure, the local policy of each user consists of disjoint rounds of sensing the M channels considered to be the best. Different users have different offsets in sensing the sets of the M channels. Consider, for example, user 1 has offset 0. In each round, the user successively senses the best, second best, $\cdots$, and the Mth best channels based on its local learning result. The offset in each user's round-robin schedule can be predetermined (*e.g.,* based on the user's ID).

To achieve the optimal order of the system regret, it is crucial that each user efficiently learn the correct ranking of the M best channels while ensuring the synchronization between the transmitter and the receiver without significant control overhead. We first propose a synchronization procedure for each transmitter and receiver, as given below.

B. Synchronization

Based on the symmetry among users, it is sufficient to consider one user, say, user 1. We assume that its transmitter and receiver have a simple initial setup for synchronization,

[5]Note that the benchmark performance of the ideal centralized case is given by orthogonalizing secondary users to the M best channels. We point out that when the false alarm probability is large, allowing multiple users to sense the same channel may lead to better exploitation of the idle slots. However, when the false alarm probability is relatively bounded (specifically, $\epsilon \le 0.5$) or when there are costs associated with secondary collisions (*e.g.,* energy consumption), orthogonalizing secondary users is desirable.

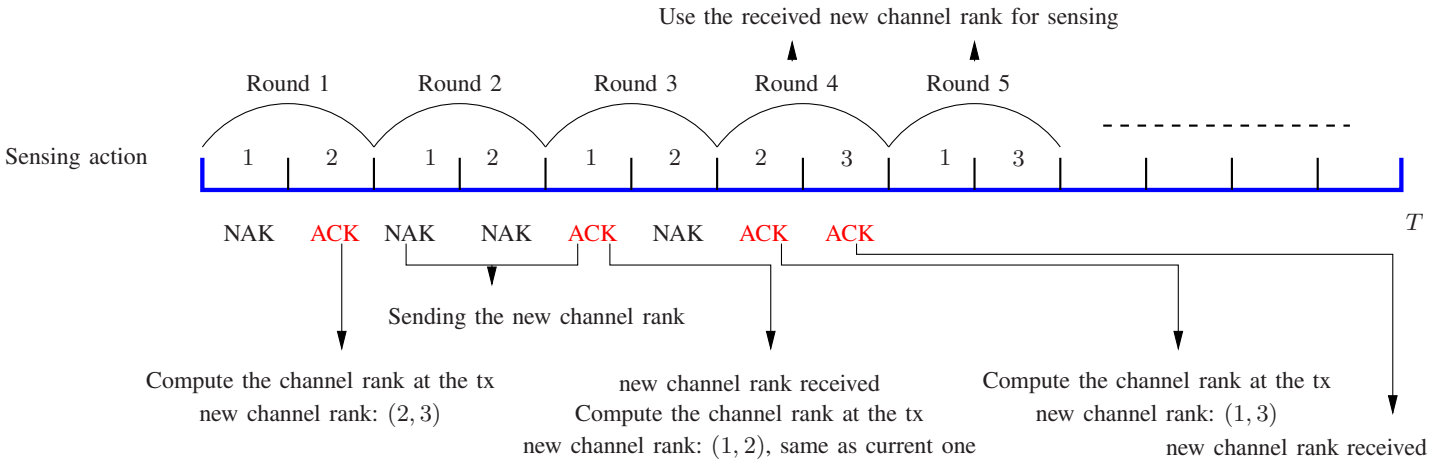


Fig. 1. An example of the structure of user 1's local policy under π_F^* ($M = 2$, $N = 3$, tx: transmitter).

e.g., in the first round[6], they will both tune to channel 1, 2, $\cdots$, M (*i.e.,* the initial rank of the M channels considered to be the best is $(1, 2, \cdots, M)$). As shown in Fig. 1, if an ACK is observed, the transmitter will compute a new channel rank based on all past detection outcomes (a detailed procedure is given in Sec. IV-C). If the new channel rank is different from the currently adopted one, the transmitter will keep sending the receiver the new rank until it is successfully received (*i.e.,* a new ACK is observed). Based upon a successful reception of the new channel rank, the transmitter and the receiver will adopt the new rank for channel sensing in the next round. We point out that, in each round, the transmitter only computes a new channel rank at most once based on the first ACK (if it exists) received in this round.

C. Efficient Learning of the Best Channels

Next, we consider the learning of the best channels at the transmitter when computing the channel rank. The basic approach is to let the transmitter learn from detection outcomes (instead of ACKs/NAKs) that represent the real channel quality. Specifically, the mean of detection outcomes from a channel (say, channel n) is equal to

$$\Pr\{S_n(t) = 1\} * (1 - \epsilon) + \Pr\{S_n(t) = 0\} * \delta = (1 - \epsilon - \delta)\theta_n + \delta,$$

the channel rank ordered by the state mean is thus the same as that ordered by the mean of detection outcomes[7]. By treating the detection outcome as the *new state* of each channel, we arrive at a perfect observation model at the transmitter. It is thus possible to extend the learning procedure for the perfect observation model addressed in [4] to the problem at hand. Basically, we let the transmitter first identify the best channel by applying a single-user policy (say, the Lai-Robbins policy [2]) for the classic MAB. To identify the kth ($1 \leq k \leq M$) best channel, the transmitter removes the $k - 1$ channels considered to have a higher rank than others and applies a parallel Lai-Robbins policy to the remaining $N - k + 1$ channels.

[6]We allow the case that users have different local channel indexes.

[7]Note that $1 - \epsilon - \delta$ is always nonnegative based on the concavity of the ROC curve [19].

A detailed implementation of the SLCD policy is given in Fig. 2.

D. Order-Optimality

To show the order-optimality of the SLCD policy, it is crucial to establish the the efficiency of the learning procedure given in Sec. IV-C. As mentioned before, compared to the perfect sensing model, the main difficulty is due to the synchronization constraint. Specifically, the transmitter cannot start sensing the channels considered as the best until the receiver has successfully received the channel rank information. The delayed channel sensing leads to different observation path and thus different channel learning result compared to the scenario of perfect sensing. Since the channel learning result further determines the future channel sensing sequence, this cascade effect needs to be carefully addressed. In the following theorem, we show that under the proposed synchronization procedure (see Sec. IV-B), the learning mistakes are bounded by the logarithmic order with time. By further bounding the control overhead for synchronization, we show that the SLCD policy achieves the logarithmic order of the system regret.

Theorem 2: Under the decentralized SLCD policy (denoted by π_F^*), we have

$$\limsup_{T\to\infty} \frac{R_T^{\pi_F^*}(\Theta)}{\log T} = C(\Theta) \tag{3}$$

for some constant $C(\Theta)$ that depends on Θ.

Proof: Note that the set of slots in which a reward loss occurs is a subset of slots in which there exists a user that does not sense the correct channel or a transmitter that sends the channel rank information instead of the data. It is thus sufficient to prove the expected number of slots that a user does not sense the M best channels in a correct order or its transmitter sends the channel rank information to the receiver has at most the logarithmic order with time.

Without loss of generality, consider user 1. We first present the following key lemma, which shows that the expected number of times that the transmitter does not compute the channel rank correctly has at most logarithmic order with time.

Lemma 2: Let $\bar{\tau}_u(T)$ denote the number of times that the channel rank is computed incorrectly at the transmitter, we

The Decentralized SLCD Policy

Without loss of generality, consider user m.

- Notations and Inputs: For two positive integers k and l, define $k \oslash l \triangleq ((k-1) \bmod l)+1$, which is an integer taking values from $1, 2, \cdots, l$. let $\bar{\theta}_n(t)$ denote the sample mean of detection outcomes obtained from channel n at the transmitter and $\tau_{n,t}$ the number of times that channel n has been sensed up to (but excluding) slot t. Let $I(\theta, \theta') = \theta \log(\theta/\theta') + (1-\theta)\log((1-\theta)/(1-\theta'))$ denote the K-L distance between the Bernoulli distributions parameterized by θ and θ', respectively. User m first senses each channel once in the first N slots to establish initial observations. Starting from slot $N+1$, user m's local policy consists of disjoint rounds of sensing the M channels considered to be the best. Let $\mathcal{Q}_k$ denote the channel sensing order in the kth round. Let $\mathcal{U}_k$ denote the number of computations of channel rank at the transmitter up to (and including) round k. Initially, $\mathcal{Q}_1 = (1, 2, \cdots, M)$ and $\mathcal{U}_0 = 0$. Select a b $(0 < b < 1/N)$.
- In the kth round, user m does the following.
  1. Both the transmitter and receiver sense the channels considered to be the M best in turn according to $\mathcal{Q}_k$. If an ACK is observed and this is the first ACK observed in this round, the transmitter sets $\mathcal{U}_k = \mathcal{U}_{k-1}+1$ and computes a new rank of the M channels considered to be the best according to step 2. If the new channel rank is different from $\mathcal{Q}_k$, the transmitter will send the receiver the new rank until the next ACK is observed. If the receiver received a new channel rank, then both the transmitter and receiver set $\mathcal{Q}_{k+1}$ equal to the new rank; otherwise $\mathcal{Q}_{k+1} = \mathcal{Q}_k$.
  2. First, the transmitter identifies the best channel. Let t denote the current time. The transmitter chooses between a leader l_t and a round-robin candidate $r_t = \mathcal{U}_k \oslash N$, where the leader l_t is the channel with the largest sample mean of detection outcomes among all channels that have been sensed for at least $(\mathcal{U}_k - 1)b$ times. The transmitter chooses the leader l_t as the best if $\tilde{\theta}_{l_t}(t) > \tilde{\theta}_{r_t}(t)$ and $I(\tilde{\theta}_{r_t}(t), \tilde{\theta}_{l_t}(t)) > \log(t-1)/\tau_{r_t,t}$; otherwise the transmitter chooses the round-robin candidate r_t as the best. To identify the kth $(k > 1)$ best channel, the transmitter removes the set of $k-1$ channels considered to have a higher rank than others from the channel set and then chooses between a leader and a round-robin candidate defined within the remaining channels. Specifically, let $m(t)$ denote the number of times that the same set of $k-1$ channels has been removed up to (and including) time t. Among all channels that have been sensed for at least $(m(t)-1)b$ times, let l_t denote the leader with the largest sample mean of detection outcomes. Let $r_t = m(t) \oslash (N-k+1)$ be the round-robin candidate where, for simplicity, we have assumed that the remaining channels are indexed by $1, \cdots, N-k+1$. The transmitter chooses the leader l_t as the kth best if $\tilde{\theta}_{l_t}(t) > \tilde{\theta}_{r_t}(t)$ and $I(\tilde{\theta}_{r_t}(t), \tilde{\theta}_{l_t}(t)) > \log(t-1)/\tau_{r_t,t}$; otherwise the user chooses the round-robin candidate r_t as the kth best.

Fig. 2. The decentralized SLCD policy.

have

$$\limsup_{T\to\infty} \frac{\bar{\tau}_u(T)}{\log T} = V(\Theta) \tag{4}$$

for some constant $V(\Theta)$ that depends on Θ.

Proof: See Appendix A for details. ∎

Now we show that the expected number of rounds that the user does not sense the M best channels in a correct order has at most the logarithmic order with time. Note that the expected number of slots between two successive computations of the channel rank at the transmitter is uniformly bounded by some constant. So the expected number of successive rounds that the user does not sense the M best channels in the correct order caused by the previous incorrect computation is uniformly bounded by some constant. Consequently, the expected number of rounds that the user does not sense the M best channels in a correct order has the same order as the incorrect computation of the channel rank at the transmitter, which has at most the logarithmic order with time based on Lemma 2.

Next, we bound the number of slots in which the transmitter sends the receiver the channel rank information instead of the data. Note that the transmitter only needs to send its receiver the information if the computed channel rank is different from the current one. Except that the channel rank is incorrectly computed, the channel ranks are all the same. By noticing that the expected number of times that the channel rank is incorrectly computed has at most the logarithmic order with time, the expected number of times that the transmitter needs to send its receiver the channel rank information has at most the logarithmic order with time. Since each sending duration till a successful reception is uniformly bounded in expectation, the expected number of slots that the transmitter sends its receiver the channel rank information has at most the logarithmic order with time.

We thus proved Theorem 2. ∎

Based on the symmetry among users' local policies, the SLCD policy achieves fairness among all users.

Theorem 3: Define the local regret for user m under the decentralized SLDC policy (denoted by π_F^*) as

$$R^{\pi^*_{F,m}}(\Theta) \triangleq \frac{1}{M} T \Sigma_{n=1}^{M} (1-\epsilon)\theta_{\sigma(n)} - \mathbb{E}_{\pi_F^*}[\Sigma_{t=1}^{T} Y_m(t)],$$

where $Y_m(t)$ is the immediate reward obtained by user m in

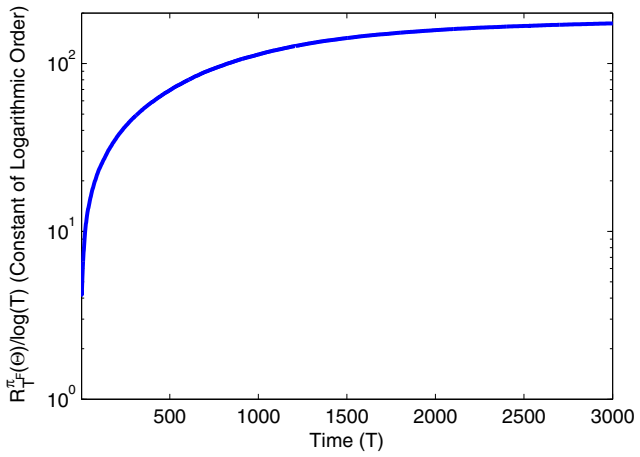


Fig. 3. The convergence of the regret ($M = 2$, $N = 9$, $\Theta = [0.1, 0.2, \cdots, 0.9]$, $\epsilon = 0.0854$, $\delta = 0.1$, (primary) signal to noise ratio=5db).

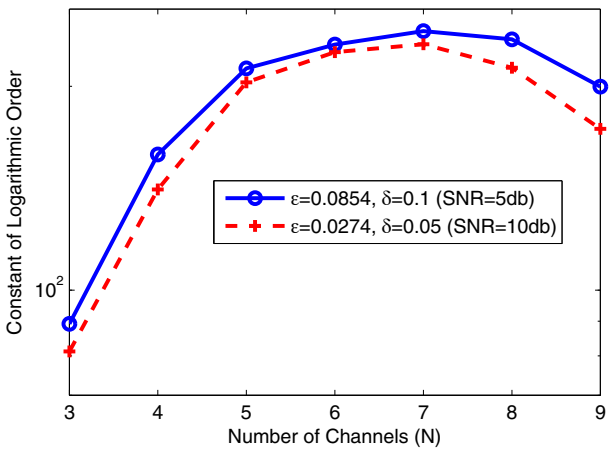


Fig. 4. The performance of SLCD ($T = 5000$, $M = 2$, $\Theta = [0.1, 0.2, \cdots, \frac{N}{10}]$, SNR: (primary) signal to noise ratio).

slot t. We have, for any $m \in \{1, \cdots, M\}$,

$$\limsup_{T\to\infty} \frac{R^{\pi^*_{F,m}}(\Theta)}{\log T} = \frac{1}{M} \limsup_{T\to\infty} \frac{R^{\pi^*_F}_T(\Theta)}{\log T}.$$

Based on Theorem 2 and 3, we arrive at the following corollary on the Nash Equilibrium of the system.

Corollary 1: Under the decentralized SLDC policy, the system achieves an order-optimal Nash equilibrium: each user cannot improve the local regret order by deviating from the local policy of SLDC.

V. Simulation Examples

In this section, we illustrate the performance of the decentralized SLCD policy. We consider the scenario that both the channel noise and the signal of the primary network are white Gaussian processes with zero mean but different power densities. The energy detector is adopted that is optimal under the Neyman-Pearson criterion [19]. In Fig. 3, we show the convergence of the system regret as a function of time. In Fig. 4, we plot the leading constant of the logarithmic order as a function of N. We observe that, from this example, the system performs better for smaller detection errors. Furthermore, the system performance is not monotonic as the number of channels increases. This is due to the tradeoff that as N increases, users are less likely to collide but learning the M best channels becomes more difficult.

VI. Extensions and Discussions

The results in this paper can be directly extended to a more general sensing model. Specifically, the probabilities (ϵ, δ) of sensing errors can vary across channels. It is only required that the probability of *detecting* an idle slot preserves the rank of the channels in terms of achievable throughput given by $\{(1-\epsilon_n)\theta_n\}_{n=1}^N$, i.e.,

$$(1-\epsilon_n)\theta_n \geq (1-\epsilon_m)\theta_m \implies (1-\epsilon_n)\theta_n + \delta_n(1-\theta_n) \geq (1-\epsilon_m)\theta_m + \delta_m(1-\theta_m).$$

Consequently, each user can efficiently learn the best channel (ranked by $\{(1-\epsilon_n)\theta_n\}_{n=1}^N$) based on the detection outcomes at the transmitter. For the general case that the error probabilities are also user-dependent, each channel offers different achievable throughput to different users. Efficient sharing among users thus becomes a complex issue. A similar problem with *centralized* users and perfect sensing was formulated as a combinatorial multi-armed bandit in [20] in which Auer *et al*'s UCB1 policy was extended to achieve a logarithmic regret. Extending the combinatorial bandit problem to the scenario of decentralized users and imperfect sensing is still open and requires a full investigation that is beyond the scope of this paper.

We further consider the generalization of the memoryless traffic model to a two-state Markovian model in which the channel state (busy or idle) transits as a Markov chain. Even with known system parameters (*i.e.,* transition probabilities) and a single user, the Markovian model yields a restless multi-armed bandit problem to which finding the optimal solution is PSPACE-hard in general [21]. For the case of unknown parameters, recent studies [15]–[18] have focused on a weaker objective: learning the arm with the highest stationary reward mean. The challenges arisen here are twofold. First, each user needs to observe a sufficient number of contiguous sample path segments to learn the stationary reward mean. Second, the user needs to bound the number of arm switchings to minimize the transient effect. Under a perfect observation/sensing model, a distributed policy was proposed in [17], [18] based on an epoch structure. Specifically, the policy consists of interleaving exploration and exploitation epochs with carefully controlled epoch lengths. During an exploration epoch, each user plays each of the N arms with even time portion to learn their reward statistics. During an exploitation epoch, each user plays the M arms locally learned as the best (ranked by the sample mean calculated from the observations obtained so far) under either a fair or a prioritized sharing scheme. The lengths of both the exploration and the exploitation epochs grow geometrically. The number of arm switchings at each user is thus at the logarithmic order with time. The tradeoff between exploration and exploitation at each user is balanced by choosing the cardinality of the sequence of the exploration

epochs. Specifically, it was shown that with an $O(\log T)$ cardinality of the exploration epochs, sufficiently accurate learning of the arm rank at each user can be obtained. For the case of imperfect sensing considered in this paper, we can incorporate the epoch structure into the SLCD policy, as detailed below.

1. Divide time into the exploration and exploitation epochs as in [18];
2. During the exploration epochs, each transmitter senses all channels in a round-robin fashion and identifies the M best arms based on the detection outcomes;
3. In the exploitation epochs, each transmitter first updates the receiver on the learned channel rank. The transmitter and the receiver then use the updated channel rank to choose and sense the best channels according to the process described in Sec. IV-B.

Based on Theorem 5 in [18] and Theorem 2, it is not difficult to show that the users can correctly learn and share the M best arms except for a logarithmic order of time, *i.e.,* the system achieves a logarithmic (weak) regret. We point out that the policy in [17], [18] and the above extended SLCD require certain knowledge on the system transition probabilities (although the knowledge can be eliminated by an arbitrarily small sacrifice of the regret order). Furthermore, all users need to adopt the same pre-determined exploration and exploitation epochs. A possible future direction is on relaxing these system constraints.

VII. Conclusion

In this paper, we addressed the dynamic spectrum access problem with distributed cooperative secondary users and imperfect spectrum sensing. Under a decentralized MAB approach, we showed that the optimal system regret has a logarithmic order with time. A decentralized channel access policy was proposed to achieve the logarithmic system regret and thus leads to a fast convergence to the same maximum throughput offered by the ideal scenario of known channel model and centralized users.

Appendix A. Proof of Lemma 2

We prove by induction on identifying the M best channels. Specifically, it is sufficient to show that, given that the $(i-1)$ best channels are correctly identified, the expected number of times that the ith best channel is not correctly identified has at most logarithmic order with time for all $1 \le i \le M$.

Let K denote the number of total computations of the channel rank over the horizon of T slots. Let $\mathcal{D}(K)$ denote the set of computations at which the $(i-1)$ best channels are correctly identified up to the Kth computation. Define function $f(x) \triangleq (1-\epsilon-\delta)x+\delta$. Consider channel n with $\theta_n < \theta_{\sigma(i)}$. For any $\alpha \in (0, f(\theta_{\sigma(i)}) - f(\theta_{\sigma(i+1)}))$, let $N_1(K)$ denote the number of computations in $\mathcal{D}(K)$ at which channel n is selected as the ith best when $l_t = \sigma(i)$ and $|\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| \le \alpha$ (t is the computation time), $N_2(K)$ the number of computations in $\mathcal{D}(K)$ at which channel n is selected as the ith best when $l_t = \sigma(i)$ and $|\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| > \alpha$, and $N_3(K)$ the number of computations in $\mathcal{D}(K)$ when $l_t \neq \sigma(i)$. It is sufficient to show that $\mathbb{E}[N_1(K)]$, $\mathbb{E}[N_2(K)]$, and $\mathbb{E}[N_3(K)]$ are all at most in the order of $\log T$.

Let $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$. Consider first $\mathbb{E}[N_1(T)]$. We have

$$\begin{aligned}
\mathbb{E}[N_1(k)] &= O(\mathbb{E}[|\{1 \le k \le K:\ k \in \{\mathcal{D}(K)\},\ \theta_{l_t} = \theta_{\sigma(i)},\\
&\qquad |\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| \le \alpha,\ \text{and channel } n \text{ is sensed}\}|])\\
&= O(\mathbb{E}[|\{1 \le j \le T-1:\ \tilde{\theta}_{n\ (j \text{ samples})} \ge f(\theta_{\sigma(i)}) - \alpha\\
&\qquad \text{or } I(\tilde{\theta}_{n\ (j \text{ samples})}, f(\theta_{\sigma(i)}) - \alpha) \le \log(T-1)/j\}|])\\
&= O(\log T), \qquad (5)
\end{aligned}$$

where the first equality is due to the fact that the probability that each computed channel rank will be executed for channel sensing is lower bounded by some constant non-zero probability, the second equality is due to the structure of the local policy of π_F^*, and the third equality follows the property of Bernoulli distributions established in [2].

Consider $\mathbb{E}[N_2(K)]$. Since the number of observations obtained from l_t at the sth ($\forall\ 1 \le s \le T$) computation is at least $(s-1)b$, we have that, $\forall\ 1 \le s \le T$,

$$\begin{aligned}
&\Pr\{\text{at the } s\text{th computation},\ \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| > \alpha\}\\
\le\ &\Pr\{\sup_{j \ge b(s-1)} |\tilde{\theta}_{l_t\ (j \text{ samples})} - f(\theta_{l_t}(t))| > \alpha\}\\
=\ &\Sigma_{i=0}^{\infty} b^i o(s^{-1})\\
=\ &o(s^{-1}), \qquad (6)
\end{aligned}$$

where the first equality is due to the property of Bernoulli distributions established in [2].

We thus have,

$$\begin{aligned}
\mathbb{E}[N_2(K)] &= \mathbb{E}(|\{1 \le k \le K:\ k \in \mathcal{D}(K),\ \theta_{l_t} = \theta_{\sigma(i)},\\
&\qquad |\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| > \alpha\}|)\\
&\le \Sigma_{s=1}^{T} \Pr\{\text{at the } s\text{th computation},\\
&\qquad \theta_{l_t} = \theta_{\sigma(i)},\ |\tilde{\theta}_{l_t}(t) - f(\theta_{l_t}(t))| > \alpha\}\\
&= o(\log T). \qquad (7)
\end{aligned}$$

Next, we show that $\mathbb{E}[N_3(K)] = o(\log T)$.

Choose $0 < \alpha_1 < (f(\theta_{\sigma(i)}) - f(\theta_{\sigma(i+1)}))/2$ and $c > (1 - Nb)^{-1}$. For $r = 0, 1, \cdots$, define the following events.

$$\begin{aligned}
A_r &\triangleq \cap_{i \le n \le N} \{\max_{\delta c^{r-1} \le s} |\tilde{\theta}_{\sigma(n)\ (s \text{ samples})} - f(\theta_{\sigma(n)})| \le \alpha_1\},\\
B_r &\triangleq \{\tilde{\theta}_{\sigma(i)\ (j \text{ samples})} \ge f(\theta_{\sigma(i)}) - \alpha_1\\
&\qquad \text{or } I(\tilde{\theta}_{\sigma(i)\ (j \text{ samples})}, f(\theta_{\sigma(i)}) - \alpha_1) \le \log(s_m - 1)/j\\
&\qquad \text{for all } 1 \le j \le bm,\ c^{r-1} \le m \le c^{r+1},\ \text{and } s_m > m\}.
\end{aligned}$$

By (6), we have $\Pr(\bar{A}_r) = o(c^{-r})$. Consider the following event:

$$\begin{aligned}
C_r &\triangleq \{\tilde{\theta}_{\sigma(i)\ (j \text{ samples})} \ge f(\theta_{\sigma(i)}) - \alpha_1\\
&\qquad \text{or } I(\tilde{\theta}_{\sigma(i)\ (j \text{ samples})}, f(\theta_{\sigma(i)}) - \alpha_1) \le \log(m)/j\\
&\qquad \text{for all } 1 \le j \le bm,\ c^{r-1} \le m \le c^{r+1}\}.
\end{aligned}$$

We have that $B_r \supset C_r$. From Lemma $1-(i)$ in [2], $\Pr(\bar{C}_r) = o(c^{-r})$. We thus have $\Pr(\bar{B}_r) = o(c^{-r})$.

Consider the sth computation where $c^{r-1} \le s-1 < c^{r+1}$. When the round-robin candidate $r_t = \sigma(i)$, we show that on the event $A_r \cap B_r$, $\sigma(i)$ must be identified as the ith best. It is sufficient to focus on the nontrivial case that $\theta_{l_t} < \theta_{\sigma(i)}$.

Since $\tau_{l_t,t} \geq (s-1)b$, on A_r, we have $\tilde{\theta}_{l_t}(t) < f(\theta_{\sigma(i)}) - \alpha_1$. We also have, on $A_r \cap B_r$,

$$\begin{aligned} \tilde{\theta}_{\sigma(i)}(t) &\geq f(\theta_{\sigma(i)}) - \alpha_1 \\ \text{or } I(\tilde{\theta}_{\sigma(i)}(t), f(\theta_{\sigma(i)}) - \alpha_1) &\leq \log(t-1)/\tau_{\sigma(i),t}. \end{aligned}$$

Channel $\sigma(i)$ is thus identified as the ith best on $A_r \cap B_r$. Since $(1-c^{-1})/N > b$, for any $c^r \leq s-1 \leq c^{r+1}$, there exists an r_0 such that on $A_r \cap B_r$, $\tau_{\sigma(i),t} \geq (1/N)(s - c^{r-1} - 2N) > bs$ for all $r > r_0$. It thus follows that on $A_r \cap B_r$, for any $c^r \leq s-1 \leq c^{r+1}$, we have $\tau_{\sigma(i),t} > (s-1)b$, and $\sigma(i)$ is thus the leader. We have, for all $r > r_0$,

$$\begin{aligned} &\Pr(\text{at the } s\text{th computation, } c^{r-1} \leq s-1 < c^{r+1}, l_t \neq \sigma(i)) \\ \leq\ &\Pr(\bar{A}_r) + \Pr(\bar{B}_r) = o(c^{-r}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[N_3(K)] &= \mathbb{E}[|\{1 \leq k \leq K : k \in \mathcal{D}(K), l_t \neq \sigma(i)\}|] \\ &\leq \Sigma_{s=1}^{T} \Pr(\text{at the } s\text{th computation, } l_t \neq \sigma(i)) \\ &\leq 1 + \Sigma_{r=0}^{\lceil \log_c T \rceil} \Sigma_{c^r \leq s-1 \leq c^{r+1}} \\ &\qquad \Pr(\text{at the } s\text{th computation, } l_t \neq \sigma(i)) \\ &= 1 + \Sigma_{r=0}^{\lceil \log_c T \rceil} o(1) \\ &= o(\log T). \end{aligned} \tag{8}$$

From (5), (7), (8), we arrive at Lemma 2.

REFERENCES

[1] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

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

[3] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—part I: IID rewards," *IEEE Trans. Auto. Control*, vol. 32, no. 11, pp. 968–976, 1987.

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

[5] L. Lai, H. El Gamal, H. Jiang, and H. Vincent Poor, "Cognitive medium access: exploration, exploitation and competition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 2, pp. 239–253, Feb. 2011.

[6] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.* , vol. 29, no. 4, pp. 781–745, Apr. 2011.

[7] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *Proc. 2011 IEEE Global Communications Conference*.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[9] G. Kasbekar and A. Proutiere, "Opportunistic medium access in multi-channel wireless systems: a learning approach," in *Proc. 2010 Allerton Conference on Communications, Control, and Computing*.

[10] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," *Mobile Networks and Applications*, vol. 11, no. 6, pp. 779–797, Dec. 2006.

[11] F. Wang, M. Krunz, and S. Cui, "Price-based spectrum management in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 74–87, Feb. 2008.

[12] J. W. Huang and V. Krishnamurthy, "Transmission control in cognitive radio as a Markovian dynamic game: structural result on randomized threshold policies," *IEEE Trans. Commun.*, vol. 58, no. 1, Jan. 2010.

[13] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[14] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: a case of near-logarithmic regret," in *Proc. 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[15] H. Liu, K. Liu, and Q. Zhao, "Logarithmic weak regret of non-Bayesian restless multi-armed bandit," in *Proc. 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[16] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: a restless bandit approach," in *Proc. 2011 IEEE INFOCOM*.

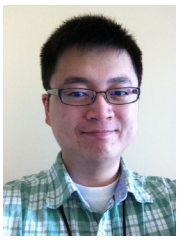
[17] H. Liu, K. Liu, and Q. Zhao, "Learning and sharing in a changing world: non-Bayesian restless bandit with multiple players," in *Proc. 2011 Information Theory and Applications Workshop*.

[18] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: restless multi-armed bandit with unknown dynamics," submitted to *IEEE Trans. Inf. Theory*, Nov. 2011. Available: http://arxiv.org/abs/1011.4969.

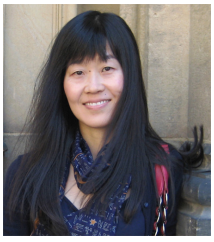
[19] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Springer, 2008.

[20] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation," *201 IEEE Symposium on International Dynamic Spectrum Access Networks*.

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



Keqin Liu (S'07-M'11) received the B.S. degree in Automation from Southeast University, China, in 2005 and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of California, Davis, USA, in 2008 and 2010, respectively. He is currently a Postdoctoral Scholar in the Department of Electrical and Computer Engineering, University of California, Davis, USA. His research interests are stochastic optimization in dynamic systems, distributed control and computing, and signal processing in wireless networks.



Qing Zhao (S'97-M'02-SM'08) received the Ph.D. degree in Electrical Engineering in 2001 from Cornell University, Ithaca, NY. In August 2004, she joined the Department of Electrical and Computer Engineering at University of California, Davis, where she is currently a Professor. Her research interests are in the general area of stochastic optimization, decision theory, and algorithmic theory in dynamic systems and communication and social networks.

She received the 2010 *IEEE Signal Processing Magazine* Best Paper Award and the 2000 Young Author Best Paper Award from the IEEE Signal Processing Society. She holds the title of UC Davis Chancellor's Fellow and received the 2008 Outstanding Junior Faculty Award from the UC Davis College of Engineering. She is also a co-author of two papers that received student paper awards at ICASSP 2006 and the IEEE Asilomar Conference 2006. She was a plenary speaker at the 11th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2010. She served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2006 to 2009 and an elected member of the IEEE Signal Processing Society SP-COM Technical Committee from 2006 to 2011.