

Distributed Learning under Imperfect Sensing in Cognitive Radio Networks

Keqin Liu*, Qing Zhao*, Bhaskar Krishnamachari[◊]

*University of California, Davis, CA, 95616, USA

{kqliu, qzhao}@ucdavis.edu

[◊]University of Southern California, Los Angeles, CA, 90089, USA

bkrishna@usc.edu

Abstract—We consider a cognitive radio network, where M distributed secondary users search for spectrum opportunities among N independent channels without information exchange. The occupancy of each channel by the primary network is modeled as a Bernoulli process with unknown mean which represents the unknown traffic load of the primary network. In each slot, a secondary transmitter chooses one channel to sense and subsequently transmit if the channel is sensed as idle. Sensing is considered to be imperfect, *i.e.*, an idle channel can be sensed as busy and vice versa. Users transmitting on the same channel collide and none of them can transmit successfully. The objective is to maximize the system throughput under the collision constraint imposed by the primary network while ensuring synchronized channel selection between each secondary transmitter and its receiver. The performance of a channel selection policy is measured by the system regret, defined as the expected total performance loss with respect to the optimal performance under the ideal scenario where all channel means are known to all users and collisions among users are eliminated through perfect scheduling. We show that the optimal system regret has the same logarithmic order as the *centralized* counterpart with *perfect sensing*. An order-optimal decentralized policy is constructed to achieve the logarithmic order of the system regret while ensuring fairness among all users.

Index Terms—Cognitive radio, distributed learning, regret, decentralized multi-armed bandit, imperfect observation

I. INTRODUCTION

We consider a distributed learning problem arisen in the context of cognitive radio networks. There are multiple distributed secondary users searching for idle channels temporarily unused by the primary network. We assume that the state—1 (idle) or 0 (busy)—of each channel evolves as an i.i.d. Bernoulli process across time slots with an unknown mean which represents the unknown traffic load of the primary network. At the beginning of each slot, each secondary transmitter chooses one channel to sense and subsequently transmits to its receiver if the channel is sensed as idle. Sensing is subject to errors: an idle channel may be sensed as busy and *vice versa*. If the transmission is successful, the secondary receiver sends back an acknowledgement (ACK) to the transmitter over the same channel at the end of the slot. The secondary users do not exchange information on their decisions and

observations. There are two types of collisions that may occur: a *primary collision* happens when a secondary user transmits in a busy channel and a *secondary collision* happens when multiple secondary users transmit in the same channel. In either case, the transmission fails. The objective is to design a decentralized channel selection policy for optimal long-term network throughput under a constraint on the maximum probability of primary collisions.

Another important design constraint is the synchronized channel selection between each secondary transmitter and its receiver. We do not assume any dedicated control channel to coordinate each secondary transmitter and receiver pair. To ensure synchronization, they can either make a decision based on the common observation history (*i.e.*, number of ACKs observed from each channel) or exploit the idle channels to exchange control information to coordinate. The tradeoff involved here is that the information from ACKs may not be sufficient for learning the channel rank (ordered by channel means) due to collisions while additional communications between a secondary transmitter and its receiver sacrifices the throughput.

We measure the performance of a decentralized policy by the system regret, which is defined as the expected total data loss with respect to the optimal performance under the ideal scenario where all channel means are known to all users and collisions among users are eliminated through perfect scheduling. The objective is to minimize the rate at which the regret grows with time. Note that the system regret is a finer performance measure than the long-term throughput. All policies with a sublinear 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 regret and construct policies optimal under this finer performance measure.

The above problem involves a complicated dilemma of exploitation, exploration, and competition. Specifically, each user needs to learn the channel rank efficiently in order to choose the best channels while avoiding significant collisions to other users. Compared to the scenario of perfect sensing, learning the channel rank under imperfect sensing is substantially more

⁰This work was supported by the Army Research Laboratory under the NS-CTA Grant W911NF-09-2-0053, and by the Army Research Office under Grant W911NF-08-1-0467.

challenging due to the imperfect observation of channel states and the synchronization constraint between each secondary receiver and its transmitter.

In this paper, we show that the minimum system regret has the same logarithmic order as the *centralized* counterpart with *perfect sensing*. A decentralized policy is constructed to achieve this optimal order. Under this policy, the system throughput quickly converges to the maximum throughput in the ideal scenario of known channel model and centralized scheduling. The proposed policy further achieves fairness among users, *i.e.*, all users converge to the same local throughput at the same rate as time goes to infinity.

Related Work Under perfect sensing, the cognitive radio network with unknown Bernoulli channel model and multiple distributed users was considered in [1]–[3]. In [1], a heuristic policy based on histogram estimation of the unknown parameters was proposed. This policy provides a linear order of the system regret, and thus cannot achieve the maximum throughput. In [2], the problem is formulated as a decentralized Multi-Armed Bandit (MAB), which generalizes the classic MAB with a single user [4], [5]. A time division fair sharing (TDFS) framework for constructing order-optimal and fair decentralized policies is proposed in [2] under general reward, observation, and collision models. In [3], order-optimal distributed policies were established based on the single-user policies proposed in [6]. Compared to the TDFS policies developed in [2], the policies proposed in [3] are limited to Bernoulli reward models and cannot achieve fairness among users. In [7], a more general channel model that allows each channel to have different means for different users is considered under perfect sensing. A centralized policy that assumes full information exchange and cooperation among users is proposed which achieves logarithmic order of the regret. In [8], the decentralized MAB proposed in [2] was extended to an imperfect observation model, where a policy was constructed to achieve $O(\sqrt{T})$ regret with time T . We point out that the results in this paper provide a non-trivial class of decentralized MAB where the optimal logarithmic regret can be achieved under imperfect observations.

Notation Let $|\mathcal{A}|$ denote the cardinality of set \mathcal{A} . For two positive integers k and l , define $k \oslash l \triangleq ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \dots, l$.

II. NETWORK MODEL

Consider the spectrum consisting of N independent but nonidentical channels and M distributed secondary users. Each user consists of one transmitter and one receiver. Let $\mathbf{S}(t) = [S_1(t), \dots, S_N(t)] \in \{0, 1\}^N$ ($t \geq 1$) denote the system state, where $S_n(t) \in \{0 \text{ (busy)}, 1 \text{ (idle)}\}$ is the state of channel n in slot t that evolves as an i.i.d. Bernoulli process with unknown mean $\theta_n \in (0, 1)$. We assume that the M largest means are distinct.

In slot t , a secondary user (say *user* m ($1 \leq m \leq M$)) chooses a sensing action $a_m(t) \in \{1, \dots, N\}$ 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:

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

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 network and also data loss of the user. Specifically, the probability of collision $\mathcal{P}_n(t)$ perceived by the primary network in any channel and slot is capped below a predetermined threshold ζ , *i.e.*,

$$\mathcal{P}_n(t) \triangleq \Pr(\text{decide } \mathcal{H}_1 | S_n(t) = 0) = \delta \leq \zeta, \quad \forall n, t.$$

We should set the miss detection probability $\delta = \zeta$ as the detector operating point to minimize the false alarm probability ϵ . If multiple 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 the 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.

III. A DECENTRALIZED MAB FORMULATION

We formulate the dynamic spectrum access problem as a decentralized Multi-Armed Bandit (MAB) with an imperfect observation model. 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 an imperfect observation model, the player may not be able to observe the actual reward offered by the selected arm. The dynamic spectrum access 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 feature of this class of decentralized MAB is the synchronization constraint on each transmitter and receiver. Specifically, in each slot, each secondary transmitter and its receiver need to select the same channel for data transmission without a dedicated control channel. One natural way is that the transmitter and its receiver use the common

local observation history (ACKs/NAKs) in learning and decision making. However, due to the collisions among secondary users, the information included in previously observed ACKs/NAKs may not be sufficient to learn the unknown channel model efficiently. An alternative approach is to let each transmitter decide whether or not to send its receiver the control information (instead of the data) for synchronization. We consider the worst scenario that each transmission of the control information occupies an entire idle slot. Since sending the control information causes a sacrifice in the immediate throughput, it should be avoided as much as possible in order to maximize the number of opportunities for transmitting the data.

We define a local policy π_i for user i as a sequence of functions $\pi_i = \{\pi_i(t)\}_{t \geq 1}$, where $\pi_i(t)$ maps user i 's local information that is common to its transmitter and receiver to the sensing action $a_i(t)$ in slot t . The decentralized policy π is thus given by the concatenation of the local policy for each user: $\pi = [\pi_1, \dots, \pi_M]$. 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 :

$$Y(t) = \sum_{j=1}^N \mathbb{I}'_j(t) S_j(t),$$

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

Let $\Theta = (\theta_1, \theta_2, \dots, \theta_N)$ be the unknown parameter set and σ a permutation such that $\theta_{\sigma(1)} > \theta_{\sigma(2)} > \dots > \theta_{\sigma(M)} \geq \theta_{\sigma(M+1)} \geq \dots \geq \theta_{\sigma(N)}$. The performance measure of a decentralized policy π is defined as the system regret

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

Note that $T \sum_{j=1}^M (1 - \epsilon) \theta_{\sigma(j)}$ is the maximum expected total reward over T slots under the ideal scenario that the parameter set $\Theta = (\theta_1, \dots, \theta_N)$ is known and users are centralized¹.

Note that the regret is always growing with time since users can never identify the channel parameters perfectly. The objective 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^* .

IV. OPTIMAL ORDER OF THE SYSTEM REGRET

In this section, we show that the minimum system regret has logarithmic order with time, which implies that the system can achieve the maximum throughput at a significantly fast rate.

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

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

¹Note that the benchmark performance of the ideal centralized case is given by orthogonalizing secondary users to the M best channels. We point out that if the false alarm probability is too large, allowing multiple users to sense the same channel may lead to better exploitation of idle slots. However, when the false alarm probability is properly bounded (specifically, $\epsilon \leq 0.5$) or considering certain cost incurred by secondary collisions (*e.g.*, energy waste), orthogonalizing secondary users is desirable.

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

Proof: See [9] for details. ■

V. THE ORDER-OPTIMAL DECENTRALIZED POLICY

In this section, we establish an order-optimal and fair decentralized policy π_F^* to achieve the logarithmic order of the system regret. The general structure of the policy is based on the time division fair sharing (TDFS) of the M best channels among the M distributed users. The TDFS structure was first proposed in [2] in the scenario of perfect sensing. Due to the imperfect observation of channel state and the synchronization constraint, extending the TDFS framework to the problem at hand is highly nontrivial.

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 M channels. Consider, for example, user 1 has offset 0. In each round, the user successively senses the best, second best, \dots , and the M th best channels according to its own ranking. 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 learns and senses the M best channels in the correct order while ensuring synchronization between each transmitter and its receiver without significant communication overhead. We first propose a synchronization mechanism for each transmitter and its receiver. 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, *e.g.*, in the first round, they will both tune to channel 1, 2, \dots , M (*i.e.*, the initial channel rank of the M channels considered to be the best is $(1, 2, \dots, M)$). As shown in Fig. 1, if an ACK is observed, the transmitter will update the channel rank according to its sensing and detection history. If the updated channel rank is different from the current one, the transmitter will keep sending its receiver the updated channel rank (instead of the data) until the update is successfully received (*i.e.*, a new ACK is observed). For simplicity of presentation, we assume that the channel capacity is large enough to send the channel rank in one slot when it is idle². Based upon a successful reception of the updated channel rank, the transmitter and receiver will use this new channel rank for channel sensing in the next round. We point out that in each round the transmitter only updates the channel rank once based on the first ACK (if it exists) received in that round.

Next, we consider the learning of the channel rank at the transmitter whenever an update is required. The basic approach is to reduce the problem to the one with the perfect observation model as considered in [2]. Note that the transmitter only uses detection outcomes (not ACKs/NAKs) to learn the channel order at each update. Since the mean of detection outcomes from a channel (say, channel n) is equal to $(1 - \epsilon - \delta)\theta_n + \delta$,

²If the channel capacity is not large enough to send the channel rank in one slot, the transmitter will send the channel rank in multiple slots.

the channel rank ordered by the state means is the same as that ordered by the means of detection outcomes. We can thus treat the detection outcome as the *new state* of each channel in learning the channel rank. Consequently, the observation of the new state becomes perfect. The user then adopts a procedure analogous to that in [2] to identify the set of the M best channels. Basically, the user first identifies the best channel by applying a single-user policy (say, the Lai-Robbins policy established in [4]) for the classic MAB. To identify the k th ($1 \leq k \leq M$) best channel, the user removes the $k - 1$ channels considered to have a higher rank than other channels and applies the Lai-Robbins policy to the remaining $N - k + 1$ channels. The main differences here from that in [2] are twofold. First, the user needs to identify the entire rank of the M best channels in one shot (as the first ACK is observed in the current round). Second, an updated channel rank cannot be realized for channel sensing without a successful reception of the rank at the receiver. Establishing the efficiency of learning the channel rank is thus more challenging compared to the scenario addressed in [2].

A detailed implementation of the decentralized policy π_F^* is given in Fig. 2.

Theorem 2: Under the decentralized policy π_F^* , we have

$$\limsup_{T \rightarrow \infty} \frac{R_T^{\pi_F^*}(\Theta)}{\log T} = C(\Theta) \quad (2)$$

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

Proof: See [9] for details. ■

Based on the symmetry among users' local policies, we show that π_F^* achieves fairness among all users.

Theorem 3: Define the local regret for user i under π_F^* as

$$R^{\pi_F^*, i}(\Theta) \triangleq \frac{1}{M} T \sum_{j=1}^M (1 - \epsilon) \mu(\theta_{\sigma(j)}) - \mathbb{E}_{\pi_F^*} [\sum_{t=1}^T Y_i(t)],$$

where $Y_i(t)$ is the immediate reward obtained by user i in slot t . We have

$$\limsup_{T \rightarrow \infty} \frac{R^{\pi_F^*, i}(\Theta)}{\log T} = \frac{1}{M} \limsup_{T \rightarrow \infty} \frac{R_T^{\pi_F^*}(\Theta)}{\log T} \quad \forall i \in \{1, \dots, M\}.$$

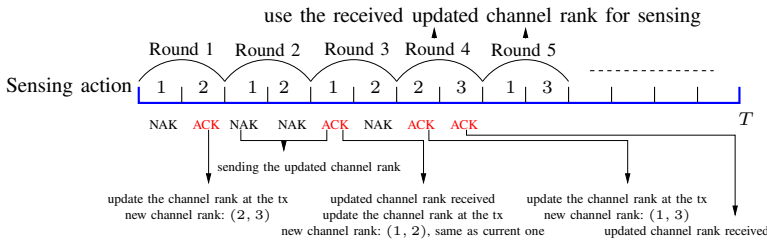


Fig. 1. An example of the structure of user 1's local policy under π_F^* ($M = 2$, $N = 3$, tx: transmitter).

VI. SIMULATION EXAMPLES

In this section, we study the performance of the order-optimal policy π_F^* for the cognitive radio network.

We consider a scenario in which both the channel noise and the signal of the primary network are white Gaussian

The Decentralized Policy π_F^*

Without loss of generality, consider user m .

- Notations and Inputs: let $\tilde{\theta}_n(t)$ 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 is 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 k th round. Let \mathcal{U}_k denote the number of updates of channel rank at the transmitter up to (and including) round k . Initially, $\mathcal{Q}_1 = (1, 2, \dots, M)$ and $\mathcal{U}_0 = 0$. Select a b ($0 < b < 1/N$).
- In the k th 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 updates the rank of the M channels considered to be the best according to step 2. The transmitter sends the receiver the updated channel rank if it is different from \mathcal{Q}_k until the next ACK observed. If the receiver received a packet consisting of the updated channel rank previously sent by the transmitter, the receiver sends back an ACK and both the transmitter and receiver set \mathcal{Q}_{k+1} equal to the updated channel rank; otherwise $\mathcal{Q}_{k+1} = \mathcal{Q}_k$.
 2. First, the transmitter identifies the best channel. Let t denote the current time. The user chooses between a leader l_t and a round-robin candidate $r_t = \mathcal{U}_k \odot N$, 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 user 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 user chooses the round-robin candidate r_t as the best. To identify the k th ($k > 1$) best channel, the user 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 is removed. 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 $r_t = m(t) \odot (N - k + 1)$ be the round-robin candidate where, for simplicity, we have assumed that the remaining channels are indexed by $1, \dots, N - k + 1$. The user chooses the leader l_t as the k th 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 k th best.

Fig. 2. The decentralized policy π_F^* .

processes with zero mean but different power densities. The energy detector is adopted since it is optimal under the Neyman-Pearson criterion [10]. In Fig. 3, we illustrate the convergence of the regret as a function of time. In Fig. 4, we plot the 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. In Fig. 5, we plot the constant of the logarithmic order as a function of M . We observe that the system performance degrades as M increases. This is due to the increased competitions and learning load encountered by all users.

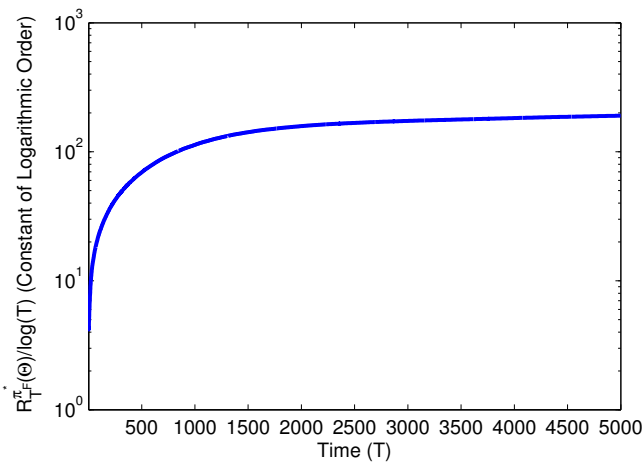


Fig. 3. The Convergence of the regret ($M = 2$, $N = 9$, $\Theta = [0.1, 0.2, \dots, 0.9]$, $\epsilon = 0.0854$, $\delta = 0.1$, (primary) signal to noise ratio=5dB).

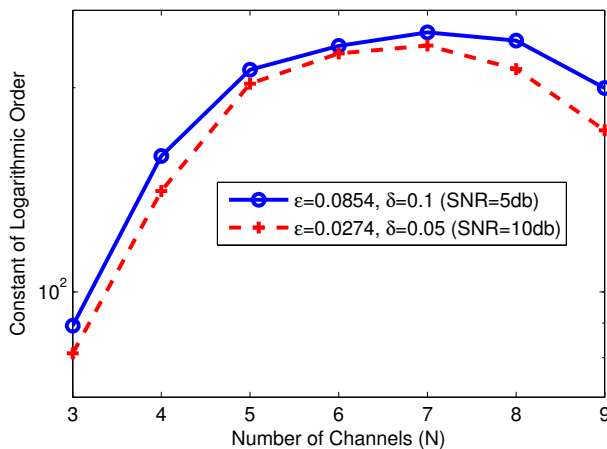


Fig. 4. The performance of π_F^* ($T = 5000$, $M = 2$, $\Theta = [0.1, 0.2, \dots, \frac{N}{10}]$, SNR: (primary) signal to noise ratio).

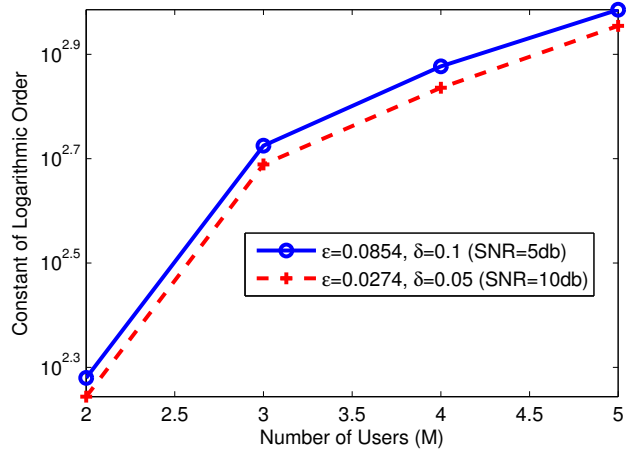


Fig. 5. The performance of π_F^* ($T = 5000$, $N = 9$, $\Theta = [0.1, 0.2, \dots, 0.9]$, SNR: (primary) signal to noise ratio).

VII. CONCLUSION

In this paper, we formulated the distributed learning problem in cognitive radio networks as a decentralized MAB, where the channel state observation is imperfect. The optimal system regret was shown to have a logarithmic order. An order-optimal and fair decentralized policy was proposed to achieve the logarithmic order of the regret, leading to a fast convergence to the same maximum network throughput as in the ideal scenario of known channel model and centralized users.

REFERENCES

- [1] L. Lai, H. Jiang and H. Vincent Poor, "Medium Access in Cognitive Radio Networks: A Competitive Multi-armed Bandit Framework," in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008.
- [2] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Distributed Multiple Players," in *Proc. of Information Theory and Applications Workshop (ITA)*, San Diego, January, 2010.
- [3] A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Players: Learning under Competition," in *Proc. of IEEE INFOCOM*, San Diego, Mar., 2010.
- [4] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.
- [5] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID Rewards," *IEEE Tran. on Auto. Control*, vol. 32, no. 11, pp. 968-976, 1987.
- [6] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, Vol. 47, pp. 235-256, 2002.
- [7] Y. Gai, B. Krishnamachari, and R. Jain, "Learning Multiplayer Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-Armed Bandit Formulation," in *Proc. of IEEE DySPAN*, Singapore, April, 2010.
- [8] K. Liu, Q. Zhao, and B. Krishnamachari, "Decentralized Multi-Armed Bandit with Imperfect Observations," in *Proc. of Allerton Conference on Communications, Control, and Computing*, September, 2010.
- [9] K. Liu, Q. Zhao, and B. Krishnamachari, "Distributed Learning under Imperfect Sensing in Cognitive Radio Networks," Technical Report, Univ. of California, Davis, June, 2010, <http://www.ece.ucdavis.edu/~qzhao/TR-10-01.pdf>.
- [10] B. C. Levy, "Principles of Signal Detection and Parameter Estimation," Springer, July, 2008.