Observation vs Statistics: Near Optimal Online Channel Access in Cognitive Radio Networks

Bowen Li*, Panlong Yang*, Jinlong Wang*, Qihui Wu*, Xiang-Yang Li†, Yunhao Liu‡
* Institute of Communication Engineering, PLAUST
† Department of Computer Science, Illinois Institute of Technology
‡ MOE Key Lab, School of Software, TNLIST, Tsinghua University

***Abstract*—We investigate efficient channel learning and opportunity utilization problem in cognitive radio networks (CRN). We find that the sensing order of multiple channels and channel accessing policy play a critical role in designing effective and efficient scheme to maximize the throughput. Leveraging this important finding, we propose a near optimal online channel access policy. We prove that, our policy can converge to an optimal point in a guaranteed probability. Further, we design a computational efficient channel access policy, integrating optimal stopping theory and multi-armed bandit policy effectively. The computational complexity is reduced from $O(KN^K)$ to $O(K)$, where N is the number of channels, and K is the maximum number of sensing/probing times in each procedure. Our simulation results validate our policy, showing at least 40% performance improvement over statistically optimal but fixed policy.**

I. Introduction

In cognitive radio networks, effective channel utilization plays an important role in improving system performance. Existing schemes can be classified into two categories based on whether channel statistics are known as a prior or not. On one hand, optimal stopping theory (OSP) [1]–[6] has been applied in dynamic spectrum access if the channel statistics information is available, making decision according to instantaneous observation. On the other hand, when there is no prior knowledge, the MAB (multi-armed bandit theory) [7]–[12] is used to tackle this problem, achieving a tradeoff between channel exploration and exploitation. It is naturally and widely believed that, these two theoretical paradigms are sufficient to tackle online channel access problem.

However, there is still a gap between these two paradigms, which makes the learning and utilizing processes difficult. Optimal stopping theory focuses on the current observation of channel states. Comparing the observations with some statistic results, the OSP theory will lead to an optimized threshold-based channel sequential sensing/probing and accessing (SSPA) strategy. What left the end user to do is selecting a time to stop sensing/probing and then accessing the channel. Different from optimal stopping theory, the MAB framework refines the statistical results using every instantaneous observation, and at every timeslot selects a channel to sense/access based on channel statistics. It is proved that MAB based approaches often lead to optimal fixed channel accessing in a large time horizon.

To tackle above issues, we formulate this as an observation V.S statistics problem in time scale. It is different from the conventional channel exploitation and exploration tradeoff problem. At each step, the observation is not only accounted for statistics, it might be also a direct stimulus for next step decision, i.e., whether to use current channel for data transmission with the observed channel quality or to further observe another channel. Moreover, the learning and utilization process can be seamless integrated together for efficient spectrum access. In other words, there is no fix borderline between exploration and exploitation processes in our scheme, we integrate both processes in a more intelligent and adaptive way. In this way, the time constraints in transmission and stochastic behavior in channel states can be solved in a unified theoretic framework.

Notice that, due to the time cost and resource constraints in the learning process, obtaining a complete channel statistics distribution is difficult. With only limited knowledge, the OSP theory framework is not workable if it is not revised, and the optimal decision might be difficult to reach. In this work, we first propose a myopic algorithm. Although the algorithm may not always converge to the optimal solution, we do have the following important insight when designing the algorithm. We note that sensing order indeed plays important role for online channel access policy. Order can also be leveraged for building an efficient channel learning and opportunity finding scheme. Leveraging this insight, we present a confidence interval estimation (CIE)-based learning policy, which achieves a near optimal balance for exploration and exploitation. Further, we build an OSP in MAB approach, seamlessly integrating the two paradigm in an efficient and effective channel learning and utilizing scheme. To the best of our knowledge, it is the first work on integrating OSP and MAB in one framework for solving the spectrum access problem. The computational overhead and time cost are considered in this investigation, which are important extensions for both MAB and OSP theoretical framework.

The contribution of this paper is three folds.

Firstly, we find that the sensing order of multiple channels and channel accessing policy play a critical role in designing effective and efficient scheme to maximize the overall throughput. With appropriate sensing order, current observations can be leveraged for opportunistic access and reducing the computational overhead.

Secondly, we present a near optimal learning policy using confidence interval estimation, which provide an efficient and

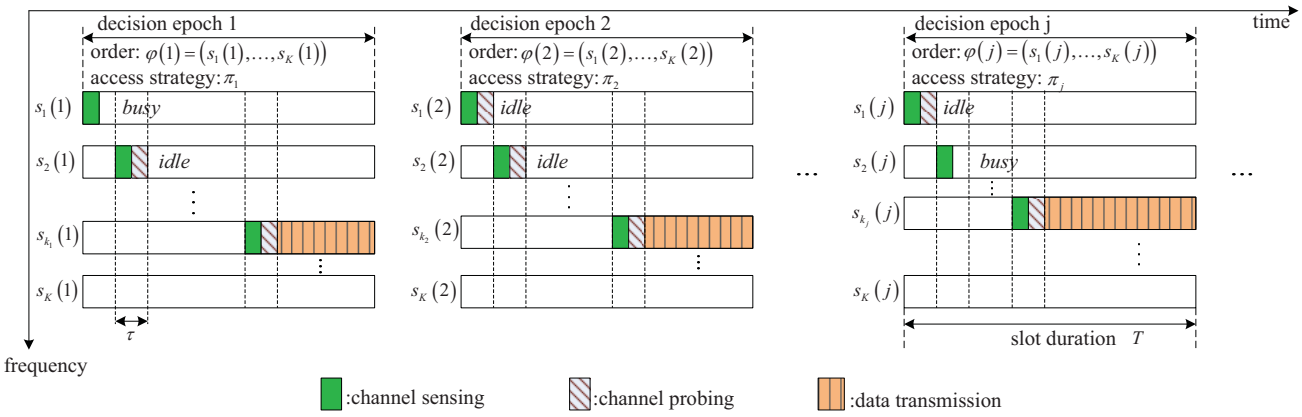

Fig. 1. Online learning of optimal sequential channel sensing, probing and accessing

effective balance between channel exploration and exploitation. We've proved that, our proposed policy converges to optimal SSPA strategy with guaranteed probability.

Thirdly, we present a computational efficient policy with slight performance loss, reducing the complexity from $O(KN^K)$ to $O(K)$, where N is the number of channels, and K is the maximum number of sensing/probing times in a SSPA process.

The rest of the paper is organized as follows. In Section II, we briefly present the system model and problem formulation. Section III describes our two algorithms separately. Our numerical simulation results are presented in Section IV. We conclude our paper in Section V.

II. System Model and Problem Formulation

Consider a cognitive radio network with potential channel set $\Omega = \{1, 2, \ldots, N\}$. Each cognitive user is operated in *constant access time* (CAT) mode [1]. The access time T for channel observation and data transmission is a constant. Such CAT scheme has been widely used for theoretical analysis in many wireless communication studies [2] [3] [10]. We denote each communication session as an *epoch*.

The channel state consists of two elements: availability and quality. Denote $a_i(j)$ as the availability of channel i in the j^{th} epoch. Availability state is $a_i(j) \in \{0, 1\}$, where $a_i(j) = 0$ indicates the primary user is transmitting over channel i in j^{th} epoch, and vice versa, $a_i(j) = 1$. We assume that the channel idle probability $\theta_i \in [0, 1]$ ($i \in \Omega$) is not known to user as a prior, but can be available through learning. The channel quality is characterized by the instantaneous received signal noise ratio (SNR) q, which corresponds to a transmit rate $\ln(1+q)$ $nats/s$ (1 nat is defined as $log_2 e (\approx 1.443) bits$). We consider a typical multipath propagation environment (i.e., Rayleigh fading), and thus the received instantaneous SNR is distributed exponentially [5] [13], and the p.d.f is given by

$$p(q) = \frac{1}{\gamma} e^{-\frac{q}{\gamma}}, \qquad q > 0$$

Where γ is the average received SNR. Denote $q_i(j)$ as the quality of channel i in j^{th} epoch, and $q_i(j)$ is under exponential distribution with mean value γ_i. Note that, the exact value of γ_i is also not known to user as a prior.

Naturally, the duration T is smaller than channel coherence time but much shorter than the sojourn time of primary user activities. It is reasonable to consider that the channel state is stable during T. As the interval time between epochs is relatively long in multi-user networks (as discussed in [1]), the channel states are independent in each epoch. This basic assumption is consistent with previous studies such as [3] [5] [8] [10].

The online learning process of SSPA is shown in Fig.1. In SSPA, user could sense/probe/access only one channel at a time. At the start of epoch j, user needs to determine a SSPA strategy $\langle \vec{\psi}(j), \vec{\pi}(j) \rangle$. The sensing order $\vec{\psi}(j) = (s_1(j), s_2(j), \ldots, s_K(j))$, which is a permutation of channels, determining the channel sensing/probing order in epoch j. While the accessing rule $\vec{\pi}(j) = (\Gamma_1(j), \Gamma_2(j), \ldots, \Gamma_K(j))$ is a sequence of SNR threshold, which helps determining whether to access the channel for transmission or not. According to strategy $\langle \vec{\psi}(j), \vec{\pi}(j) \rangle$, the SSPA procedure of epoch j proceeds step by step as follows. Note that, each channel sensing/probing process in an epoch means a step. First, user senses channel $s_1(j)$ to acquire channel availability $a_{s_1(j)}(j)$. If $a_{s_1(j)}(j) = 1$ (i.e., channel is idle), user further probes the channel, acquiring instantaneous received SNR $q_{s_1(j)}(j)$. After that, the user would compare $q_{s_1(j)}(j)$ with the first access threshold $\Gamma_1(j)$ in $\vec{\pi}(j)$ to determine whether to access the channel or go on SSPA process. If the channel is busy, user needs to wait for a constant channel probing time before switching to next channel. Such scheme is introduced for transceiver synchronization [4]. As a result, each step costs a constant time τ. Then, the maximum number of steps one could take in one epoch is $K = \min\left(N, \lfloor \frac{T}{\tau} \rfloor\right)$, where $\lfloor \cdot \rfloor$ represents round-down function. When user decides to access channel for data transmission after k^{th} channel sens-

ing/probing, the immediate throughput reward is

$$r(j) = c_k \ln\left(1 + q_{s_k(j)}(j)\right) = (1 - k\beta)\ln\left(1 + q_{s_k(j)}(j)\right) \tag{1}$$

Where $\beta = \frac{\tau}{T}$ is the normalized observation cost, and $c_k = 1 - k\beta$ denotes the normalized remaining transmission time at step k. The actual throughput can be easily obtained by scaling our reward with a constant $\frac{T}{\ln 2}$.

We define the deterministic learning policy χ to be a map from observation history $\mathcal{F}$ to SSPA strategy $\langle\vec{\psi}, \vec{\pi}\rangle$. Determining a SSPA strategy $\langle\vec{\psi}, \vec{\pi}\rangle$ in each epoch includes: 1) selecting K channels from channel set Ω, 2) arranging the order of the selected K channels for sequential channel sensing/probing, and 3) obtaining the accessing rule for channel accessing. Our main goal is to devise a learning policy guiding the system converging to the throughput-optimal SSPA strategy. Meanwhile, we need the cost on learning as small as possible. In the rest of this paper, we denote χ the learning policy and $\langle\vec{\psi}, \vec{\pi}\rangle$ as the joint SSPA strategy, in which $\vec{\psi}$ is the sensing order and $\vec{\pi}$ is the accessing rule.

III. Computational Efficient Policy: An OSP in MAB Approach

To reduce the computational complexity, we consider a decoupling approach where in each epoch, the joint decision-making process is separated into two phases: sensing order selection and accessing rule derivation. We formulate the sensing order selection across epochs as a multi-armed bandit problem, and obtain the accessing rule using optimal stopping theory. We call this decoupling approach as 'OSP in MAB'. In this learning policy, the user just needs to calculate the accessing rule of the selected order in each epoch, thus the computational complexity is greatly decreased.

A. Algorithm Description

We consider each sensing order as an arm, and the order selection problem across epochs is formulated as a multi-armed bandit problem. At each epoch, user chooses an arm according to the historical reward statistics, obtaining the immediate throughput reward as well as refining the statistics. During an epoch, finding the optimal SSPA strategy under the chosen sensing order $\vec{\psi}$ is formulated as an optimal stopping problem. The accessing rule $\vec{\pi}\left(\vec{\psi}, \left\{\vec{\theta}, \vec{\gamma}\right\}\right)$ that maximizes the immediate throughout in current epoch is derived by backward deduction.

1) Order Selection Across Epochs: In each epoch, user selects a sensing order and proceeds SSPA according to the corresponding optimal accessing rule (the acquisition of accessing rule is introduced in Section III-A2). The reward is recorded for achieving optimal sensing order. Always, we need to select the currently best sensing order to maximize immediate reward. Note that, there is still a need to carefully explore other suboptimal orders to improve overall throughput.

We leverage the UCB1 [14] approach in order to achieve a proper balance between exploitation and exploration. Two variables are used for each order $\vec{\psi}_m$ $(1 \le m \le M)$: $\hat{\mu}_m(j)$ is the average value of all the observed rewards of order $\vec{\psi}_m$ up to the epoch j, and $n_m^o(j)$ is the number of times that $\vec{\psi}_m$ having been chosen up to epoch j. They are both initialized to zero and updated according to the following rules:

$$\hat{\mu}_m(j) = \begin{cases} \frac{\hat{\mu}_m(j-1)n_m^o(j-1) + r_m(j)}{n_m^o(j-1)+1}, & \text{if order } \vec{\psi}_m \text{ is selected} \\ \hat{\mu}_m(j-1), & \text{else} \end{cases} \tag{2}$$

$$n_m^o(j) = \begin{cases} n_m^o(j-1) + 1, & \text{if order } \vec{\psi}_m \text{ is selected} \\ n_m^o(j-1), & \text{else} \end{cases} \tag{3}$$

At the very beginning, each order is chosen only once. As the progress goes on, one would always choose the order with highest index $\hat{\mu}_m^u(j)$ in the j^{th} epoch. Where

$$\hat{\mu}_m^u(j) = \hat{\mu}_m(j) + r_{max}\sqrt{\frac{2\ln j}{n_m^o(j)}} \tag{4}$$

is composed of two items defining the exploration vs. exploitation trade-off [14]. The maximum achievable immediate reward in one epoch is given by $r_{max} = (1-\beta)\log(1+q_{max})$. The first item in Equ.(4) is the average throughput $\hat{\mu}_m(j)$ up to epoch j. The second item is related to the size of the one sided confidence interval (according to Chernoff-Hoeffding Bounds) for the average reward. In summary, the sensing order with higher average reward $\hat{\mu}_m(j)$ as well as smaller $n_m^o(j)$ has the higher priority to be selected. As the enumerator $2\ln j$ increases sub-linearly with epoch j, there is a tendency for user in favor of the sensing order with the highest average reward as time goes by.

2) Accessing Rule in One Epoch: As the sensing order $\vec{\psi}(j) = (s_1(j), s_2(j), \ldots, s_K(j))$ has been determined by Equ.(4), given the current statistics $\left\{\vec{\theta}, \vec{\gamma}\right\}$, the accessing rule $\vec{\pi}(j) = (\Gamma_1(j), \Gamma_2(j), \ldots, \Gamma_K(j))$ that maximizes immediate throughput reward can then be derived by backward deduction. Then, the SSPA in current epoch is carried out as follows: sequentially sense/probe channels according to channel sequence $(s_1(j), s_2(j), \ldots, s_K(j))$, and access channel $s_k(j)$ $(1 \le k \le K)$ when the observed channel quality $q_{s_k(j)}(j) \ge \Gamma_k(j)$.

The complete procedure of *OSP in MAB* approach is then listed in Fig.2.

B. Complexity Analysis

As shown in Fig.2, since user needs only to calculate the optimal stopping rule for the given sensing order, the computational complexity is $O(K)$. Comparing with CIE-based learning policy, the computational complexity is greatly reduced.

However, such computational benefit comes at the cost of higher storage overhead. The OSP in MAB learning policy needs to record two variables for each possible sensing order and four variables for each channel statistics, thus the storage overhead is $O\left(N^K\right)$.

Moreover, as we stated before, the OSP in MAB policy is a compromised approach that decouples the joint optimization into two phases, and the two phases are optimized separately.

Algorithm OSP in MAB

1: $j = 0$; for all $1 \leq m \leq M$: $\hat{\mu}_m = 0$, $n_m^o = 0$; for all $1 \leq i \leq N$: $\hat{\theta}_i = 0$, $n_i^s = 0$, $\hat{\gamma}_i = 0$, $n_i^p = 0$
2: Sense and probe channels sequentially, guarantee that all channels are probed at least one time
3: Update $\hat{\theta}_i$, n_i^s, $\hat{\gamma}_i$, n_i^p accordingly
4: **for** $j = 1 : M$ **do**
5: Select order $\vec{\psi}_j$
6: Proceed SSPA with $\langle \vec{\psi}_j, \vec{\pi}\left(\vec{\psi}_j, \left\{\vec{\hat{\theta}}, \vec{\hat{\gamma}}\right\}\right)\rangle$
7: Update $\hat{\mu}_m$, n_m^o, $\hat{\theta}_i$, n_i^s , $\hat{\gamma}_i$, n_i^p accordingly
8: **end for**
9: **for** $j = M + 1 : L$ **do**
10: Select order $\vec{\psi(j)} = \vec{\psi}_m$ that maximizes $\hat{\mu}_m + r_{max}\sqrt{\frac{2\log j}{n_m^o}}$
11: Proceed SSPA with $\langle \vec{\psi}_m, \vec{\pi}\left(\vec{\psi}_m, \left\{\vec{\hat{\theta}}, \vec{\hat{\gamma}}\right\}\right)\rangle$
12: Update $\hat{\mu}_m$, n_m^o, $\hat{\theta}_i$, n_i^s , $\hat{\gamma}_i$, n_i^p accordingly
13: **end for**

Fig. 2. Algorithm on OSP in MAB Policy

It is hard to prove the optimality of the approach. However, although strict proof on the optimality is not available, we have done extensive simulation, finding that the performance of the OSP in MAB learning policy is close to that of CIE-based learning policy.

IV. Simulation and Performance Analysis

In this section, we evaluate the proposed algorithm via simulation and make performance analysis on the achieved results.

A. Throughput Gain of Diversity Exploitation

Before investigating the performance of the proposed learning policy, we show first the throughput gap between the mechanisms with diversity exploitation and without diversity exploitation when channel statistics are known. For scheme without diversity exploitation, user chooses only one channel in each slot. Specifically, user senses/probes a chosen channel in each epoch. If the channel is idle then the user will transmit data with the maximum achievable rate. Otherwise, wait until the next epoch. Such communication model is considered in [7]–[12]. We now consider that user could always choose the channel with the highest expected capacity in each slot, thus leading to the maximum achievable throughput without diversity exploitation. We call this method *statically optimal* scheme. For diversity exploitation, we consider the SSPA strategy. User sequentially senses/probes multiple channels in each epoch. Each sensing/probing costs a normalized time, $\beta = \frac{\tau}{T}$. Note that, the SSPA always proceeds with optimal sensing order and accessing rule, which is derived by appealing to optimal stopping theory.

The throughput gain is defined as the ratio of the throughput in our *optimal SSPA strategy* and the *statically optimal*scheme, which is shown in Fig.3. The result is derived from 100 groups of independent parameters. In each group, the channel idle probability is randomly generated in the range of $[0, 1]$ and the average received SNR is generated in the range of $[5, 15]dB$. Comparing with the static scheme, the SSPA strategy could achieve appreciable throughput gain, fully exploiting channel diversity. It is clear that such diversity gain increases when N increases or β decreases. This is reasonable, since more channels would lead more instantaneous channel quality diversity, meanwhile, a higher β value indicates a higher cost for diversity exploitation. It is shown that even when β is very high, i.e., $\beta = 0.05$ (i.e. $T = 1s$ when $\tau = 50ms$), the SSPA strategy could outperform static scheme about 40% in throughput when $N = 10$. In the following subsection, we evaluate the proposed policies that attain channel diversity by learning when channel statistics are unknown.

B. Performance of Online Learning of SSPA

In this subsection, we consider the channel statistics are unknown, and evaluate the performance of our proposed learning policies. Five policies are considered for performance comparison. They are *myopic policy*, *CIE-based online learning policy*, *OSP in MAB policy*, *genie-based policy* and *order optimal single index policy*. The genie-based policy is a reference, which uses the optimal SSPA strategy derived from full channel statistical information, and obtain the maximum throughput. The *order optimal single index policy* is presented by Lai *et al.* in [7]. Such learning policy has been proved to be order optimal when without considering diversity exploitation, i.e., user only senses/probes one channel in each slot.

Our experiment settings are as follows. We first run the experiment 100 rounds independently. Each round consists of 5000 decision epochs. Similarly, the channel idle probability is randomly generated in range $[0, 1]$ and the average received SNR is generated in range $[5, 15]dB$. At the very begining of each decision epoch, the channel availability as well as channel instantaneous quality (i.e. SNR) are generated independently according to statistical parameters in current round. We run the five policies under the same environment respectively. The parameters in the simulations are as follows: the number of channel $N = 5$, normalized sensing/probing cost $\beta = 0.05$, δ in CIE policy is set to 0.01.

Then, we derive the regret of all the four policies (except genie-based policy) by comparing their obtained accumulated throughput with genie-based policy. The regret is shown in Fig.4. Moreover, the averaged regret per epoch which defined as $\frac{\rho(L)}{L}$ is also depicted in Fig.5. It is clearly shown in these two figures that the *order optimal single index policy* performs poor in respect of throughput per unit time. As shown in Fig.5, even when $L = 5000$, there exists a constant throughput gap between the optimal SSPA strategy and order optimal single index policy. Similar to the regret of *order optimal single index policy*, the regret of *myopic policy* is also approximately linearly increased with epoch number L. The main reason is that, the myopic policy converges to a sub-optimal SSPA strategy, leading to a constant gap on throughput. It is obvious

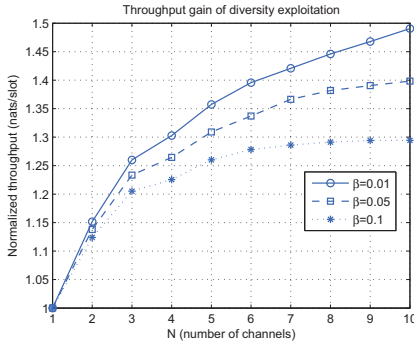

Fig. 3. Throughput gain of *SSPA* over *always the best one*

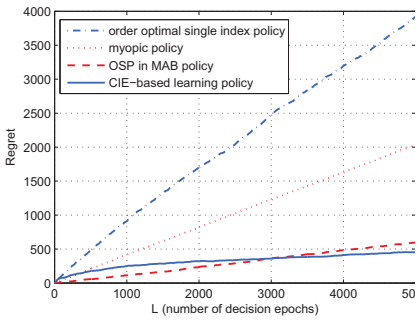

Fig. 4. Regret of the four policies

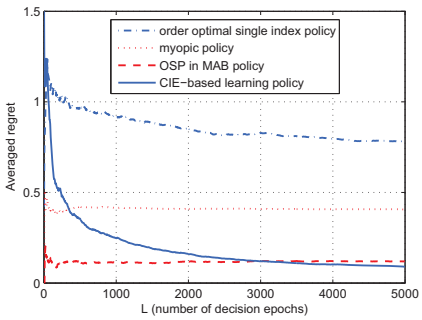

Fig. 5. Averaged regret: $\rho(L)/L$

that our proposed two policies: *CIE-based online learning policy* and *OSP in MAB policy*, perform well. The regret is sub-linearly increased with L. Generally speaking, the *CIE-based online learning policy* slightly outperforms *OSP in MAB policy* in respect of system throughput. However, the *OSP in MAB policy* would be more favorable in practical design for lower computational complexity.

V. Conclusion

In this work, channel learning and opportunity utilization problem is considered with the resource constraints and timing cost. We find that the channel sensing order and accessing rule is important for maximizing overall throughput. We leverage it to design a low computation complexity algorithm. With appropriate sensing order, observations can be leveraged for opportunistic access and reduce the computational overhead. The CIE-based method can achieve an efficient and effective balance between channel exploration and exploitation. It converges to optimal SSPA strategy with guaranteed probability. However, the CIE-based method is in high computational complexity. The OSP-MAB based method can significantly reduce the complexity in computation, and there is slight performance loss. Also, the storage complexity increases, but it is acceptable.

In future work, we are to improve the overall network performance in presence of multiple access contention from secondary users should also be seriously considered. Also, we are to implement our policy to cognitive radio platform, such as USRP [15], and provide a working system for validation.

Acknowledgment

This research is partially supported by NSF China under Grants No. 60932002, 61003277, 61170216, 61172062, NSF CNS-0832120, NSF CNS-1035894, China 973 Program under Grants No. 2009CB320400, 2010CB328100, 2010CB334707, 2011CB302705, program for Zhejiang Provincial Key Innovative Research Team, and program for Zhejiang Provincial Overseas High-Level Talents.

References

[1] A. Sabharwal, A. Khoshnevis, and E. Knightly, "Opportunistic spectral usage: bounds and a multi-band csma/ca protocol," *IEEE/ACM Trans. Netw.*, vol. 15, pp. 533–545, June 2007.

[2] P. Chaporkar and A. Proutiére, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1546–1555, 2008.

[3] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 17, pp. 1805–1818, 2009.

[4] T. Shu and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," in *Proceedings of the 15th annual international conference on Mobile computing and networking*, ser. MobiCom '09. New York, NY, USA: ACM, 2009, pp. 37–48.

[5] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 297–307, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1109/T-WC.2009.071363

[6] B. Li, P. Yang, J. Wang, Q. Wu, and X. yang Li, "Finding optimal action point for multi-stage spectrum access in cognitive radio networks," in *ICC*, 2011.

[7] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation and competition," *CoRR*, vol. abs/0710.1385, 2007.

[8] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players." *IEEE Transactions on Signal Processing*, pp. 5667–5681, 2010.

[9] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *INFOCOM*, 2010, pp. 803–811.

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

[11] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *DYSPAN*, 2010.

[12] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *INFOCOM*, 2011.

[13] Q. Zhang and S. A. Kassam, "Finite-state markov model for rayleigh fading channels," *IEEE Transactions on Communications*, vol. 47, pp. 1688–1692, 1999.

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

[15] R. Dhar, G. George, and A. Malani, "Supporting integrated mac and phy software development for the usrp sdr." VA, USA: USENIX Association: Networking Technologies for Software Defined Radio Networks, SDR06, Mar. 2006.