Online Sequential Channel Accessing Control: A Double Exploration vs. Exploitation Problem

Panlong Yang, *Member, IEEE*, Bowen Li, *Student Member, IEEE*, Jinlong Wang, Xiang-Yang Li, *Fellow, IEEE*, Zhiyong Du, *Student Member, IEEE*, Yubo Yan, *Student Member, IEEE*, and Yan Xiong

Abstract—**In opportunistic channel access, the user needs to make real time decisions on when and which channel to access with uncertainty. Assuming perfect channel statistics, several studies have applied optimal stopping theory to derive control strategy for sequential sensing/probing based opportunistically accessing (*s*-SPA), exploiting temporary opportunities among multiple channels. Meanwhile, numerous multi-arm bandit (MAB)-based approaches have been proposed for online learning of channel selection in periodical sensing/accessing system, however, these schemes fail to exploit the opportunistic diversity in short term. In this paper, we investigate online learning of optimal control in *s*-SPA systems, where both statistics learning and temporary opportunity utilization are jointly considered. An effective and efficient online policy, so called IE-OSP, is proposed, which theoretically guarantees system converges to the optimal *s*-SPA strategy with bounded probability. Experimental results further show that, the regret of IE-OSP is almost in optimal logarithmic increasing rate over time, and is sub-linear with the increasing number of channels. Compared with existing solutions, our proposed algorithm achieves 25 $\sim$ 30% throughput gain in typical scenarios.**

Index Terms—**Opportunistic spectrum access, sequential sensing and accessing, online learning, diversity exploitation.**

I. Introduction

OPPORTUNISTIC channel access (OSA), due to its flexibility and efficiency in spectrum utilization, has become a well established concept in designing wireless systems [1], [2]. With the success of OSA-based standards such as IEEE 802.11h [3], 802.22 [4], and 802.11af [5], more and more organizations are considering to adopt OSA in future communication standards. In achieving perfect opportunistic channel utilization, the key challenge comes from the unpredictable channel status. Specifically, to acquire the exact channel state, user needs to detect whether the channel is available with spectrum sensing [6], and evaluate the link quality with probing [7]. Online accessing control, i.e., making real time decisions on when and which channel to access, plays a critical role in improving system performance as well as avoiding interference to primary users.

Based on sequential channel sensing and probing, user could opportunistically access a good channel for communication, so as to exploit diversity of temporary channel status among channels. The sequential accessing control problem is firstly studied in multiple i.i.d Rayleigh channels scenario [8], where a multichannel opportunistic auto rate protocol is proposed. Further, more generalized scenarios allowing users to recall pre-probed channels [9], [10] or considering the activities of primary users [11], [12] are further studied. The major concern in these studies is to balance exploration and exploitation on temporary channel status. Corresponding control strategies are constructed on the ideal assumption that the user has perfect knowledge of channel statistics. Since channel statistics are usually unavailable in advance, obtaining complete channel statistics before a communication session will be costly, and would also result in unacceptable delay and overhead.

Our work aims to achieve more throughput gain under the rule of MAB. The reason is, the short-term statistical results could be leveraged for such improvement. We find that, even when no recall action is allowed, the optimal stopping rule could still be applied, where users could opportunistically select the temporary 'good' channel to access, if the user could sense more channels. This motivation relies on two basic facts. First, most of the channels are slow fading, especially for indoor WiFi transmissions. Second, with the advances of wireless communication technology, the channel probing efficiency could be improved in relatively smaller time. Motivated by the aforementioned two conditions, we believe that, the statistical channel knowledge accumulated in the probing process could be leveraged for performance improvements.

To this end, this paper attempts to combine the following two models that have each been quite extensively studied in recent literature: (1) using online learning methods to make sequential channel access decisions when the average channel qualities are unknown a priori (which involves exploration and exploitation); and (2) optimal stopping time methods to determine whether to

Manuscript received June 26, 2014; revised December 4, 2014; accepted April 3, 2015. This research is partially supported by NSF China under Grants No 61232018, 61272487, 61170216, 61172062, 61228202, 61273210, NSF CNS-0832120, NSF CNS-1035894, NSF ECCS-1247944, and NSF CMMI 1436786. The associate editor coordinating the review of this paper and approving it for publication was C. Ghosh.

P. Yang is with the Institute of Communication Engineering, People's Liberation Army University of Science and Technology (PLAUST), Nanjing 210007, China, and also with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China (e-mail: panlongyang@gmail.com).

B. Li is with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China.

J. Wang, Z. Du, and Y. Yan are with the Institute of Communication Engineering, People's Liberation Army University of Science and Technology (PLAUST), Nanjing 210007, China.

X.-Y. Li is with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China, and also with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616-3793 USA.

Y. Xiong is with the Department of Computer Science and Technology, University of Science and Technology, Hefei 230026, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2424413

continue sensing the qualities of a given sequence of channels or stop and use the channel for data transmission.

We first analyze the property of optimal sequential sensing, probing and accessing strategy with perfect channel statistics, and then propose an intuitive solution, i.e., myopic learning policy, to help understanding the online accessing control problem. After analyzing the convergence of the myopic learning policy, we find that properly exploring the inaccurately estimated channels is critical for guaranteeing the convergence property. Inspired by this observation, we develop an online policy referred to as IE-OSP, which achieves nearly optimal balance between exploration and exploitation. The main contribution of this paper is two-folds:

First, the brand new double exploration vs. exploitation problem is well studied under the myopic learning policy. We show that, such learning policy with greedy exploitation is non-zero-regret, which indicates that, optimizing opportunity exploitation during a slot is incompatible with that of statistics exploration. Thus, a tradeoff between them is needed for maximizing overall system throughput. Moreover, both the sensing order and accessing rule play critical roles in designing effective and efficient online learning policy.

Secondly, we present a statistical learning based online policy referred to as IE-OSP, which integrates confidence interval estimation into the optimal stopping analytical framework. We've proved that, using the IE-OSP policy, system is guaranteed to converge to the optimal *s*-SPA strategy with bounded probability. Extensive simulation results show that, the expected regret of the IE-OSP policy achieves near optimal logarithmic increasing rate over time, and is sub-linear increasing with the number of channels. Comparing with existing solutions, our proposed scheme achieves 25∼30% throughput gain in most scenarios.

The rest of the paper is organized as follows. The related work is introduced in Section II and in Section III, we briefly present the system model and problem formulation. Further, we analyze the online sequential channel accessing control problem with an intuitive learning policy in Section IV. In Section V, the proposed IE-OSP algorithm and corresponding analysis are presented. Our evaluation results are presented in Section VI. Finally, we conclude our paper in Section VII.

II. Related Work

Opportunistic spectrum accessing control have received much attention recently. Online decisions are made under channel uncertainty, maximizing the system throughput by flexibly exploiting communication opportunities. The most relevant studies to our work can be classified to the following two broad categories:

A. Optimal Control for Sequential Sensing, Probing, and Accessing

To efficiently explore and exploit *diversity on temporary channel status* among multiple channels, optimal control algorithms for sequential channel sensing, probing and accessing scheme have been widely studied. The real time decisions, i.e., whether to access channel or continue to observe another channel immediately, are made on the observed temporary channel status.

Considering i.i.d. Rayleigh fading channels, Sabharwal *et al.* [8] firstly analyze the gains from opportunistic band selection. To obtain such gain, sequential probing based opportunistic channel accessing scheme is proposed, and optimal skipping rule is derived by finite-horizon optimal stopping formulation. More generalized scenarios, e.g., with arbitrary number of channels, statistically non-identical channels, and possibly different probing costs, are studied in seminar work [9], [10], [13]. Moreover, recalling a pre-probed channel as well as accessing an unobserved channel are allowed in their considered communication model.[1] The corresponding optimal strategies are derived by comprehensive theoretic proofs. In [11], Shu and Krunz consider an OSA network with primary users, and thus channel quality as well as availability are considered when making accessing decisions. States of different channels are considered to be i.i.d. to each other, and an infinite-horizon optimal stopping model is leveraged to formulate the online control problem during the *s*-SPA process. For scenarios with non-identical channels, sensing order plays a critical role in achieving maximum throughput. Jiang *et al.* firstly considered the problem of acquiring the optimal sensing/probing order for a single user case in [12]. A computational efficient algorithm is constructed by appealing to dynamic program. Later, Fan *et al.* [14] extends sensing order selection to a two-user case, where a coordinator in the network to determine the sensing orders for each of the two users is required. Recently, Zhao *et al.* [15] propose a novel sensing metric that integrate the channel availability, link quality and access collisions, to guide the sensing order selection. A dynamic programming algorithm is presented, which allows each node to efficiently determine its sensing order in coordination with neighboring nodes. More recently, Pei *et al.* [16] extend the sequential channel sensing and accessing control to a new area, where energy-efficiency is mainly concerned. In their work, sensing order, accessing strategy and transmit power are jointly optimized with dynamic programming. Unlike assuming time-independent channels, i.e., channel states are considered to be independent across slots, Li *et al.* [17] consider Markovian channels and investigate the sequential probing based opportunistic channel accessing and releasing scheme, where a two-dimension optimal stopping framework is proposed for achieving optimal action point under Rayleigh fading. Wang *et al.* [18] exploit constructive interference for scalable flooding. Reference [19]–[21] propose schedule schemes to optimize throughput. Other works [22]–[24] are proposed to exploit the frequency diversity.

The major difference between our work and the above-mentioned studies can be explained as follows. In all the above-mentioned studies, the optimal control strategies are constructed on the assumption of perfect channel statistics. In contrast, we consider more practical scenarios that channel

[1] "Recalling a channel" means revisit the previous probed channel. Such that, the reward could be increased if the user found the previously probed channel is better. Comparing with scheme without recalling, such scheme could achieve lower regret value.

statistics are unknown in the beginning, and focus on investigating online learning method to achieve optimal control of sequential sensing, probing and accessing.

B. Online Learning of Dynamic Channel Selection

Online learning framework for opportunistic spectrum access when channel statistics is unknown a priori, especially formulated as multi-armed bandit (MAB) problems [25], has been fully investigated for periodical sensing/accessing system. The main concern in these studies is to explore and exploit *diversity on channel statistics* among multiple channels efficiently. Specifically, the dynamic selection process is expected to converge to choosing the statistically optimal channel, i.e., the channel with maximum expected reward, thus to achieve diversity gain over channel statistics.

Lai *et al.* [26] firstly apply multi-arm bandit formulations to user-channel selection problems in OSA networks. Especially for the single user case, the UCB1 [27] algorithm is proposed, which is order-optimal with respect to regret. And for decentralized multiple users, a randomized access policy is presented for learning the unknown parameters efficiently. Liu and Zhao [28] formulate the secondary user channel selection to a decentralized multi-armed bandit problem, where contentions among multiple users are considered. A policy achieving asymptotically logarithmic regret is proposed in their work. Anandkumar in [29] and [30] proposed two policies for distributed learning and accessing rule, lead to order-optimal throughput. In addition to learning the channel availability, the secondary users also learn others' strategies, even the total number of users, through channel level feedback. Tekin and Liu [31] modeled each channel as a restless Markov chain rather than time-independent channels as studied before, and multiple channel states rather than binary states are considered. They present a sample-mean based index policy, showing that, under mild conditions, it could achieve logarithmic regret uniformly over time. For the multiuser-multichannel matching problem, Gai *et al.* [32] develop a combinatorial multi-armed bandits (MAB) formulation to address the channel allocation problem under centralized setting. An online learning algorithm that achieves $O(\log T)$ regret uniformly over time is derived. Later, Kalathil *et al.* [33] consider a decentralized setting where there is no dedicated communication channel for coordination among the users. An online index-based distributed learning policy called the dUCB4 algorithm is developed, which achieves the expected regret growing at most as $near-O(\log^2 T)$. Huang *et al.* [34] study the scaling problem of general cognitive radio networks, Dong *et al.* [35] propose a auction scheme.

The main difference between our work and existing online learning frameworks can be explained as follows. All existing studies are focused on periodical sensing/accessing system, where the user only needs to select one channel at a slot. While we consider online learning of optimal control in sequential sensing, probing and accessing systems, where a series of decisions are needed to be made in each slot.

Remark: To the best of our knowledge, it is the first work on integrating OSP and MAB in one unified theoretic framework, making a good balance between statistical exploration across slots and opportunity exploitation during a slot.

III. System Model and Problem Formulation

Considering an OSA network with potential channel set $\Omega = \{1, 2, \ldots, N\}$, each cognitive user could sense/probe/access only one channel at a time, and is operated in *constant access time* (CAT) mode [8], i.e., users could have a constant duration T for channel observation and data transmission, once they would win a communication chance. The communication chances of users come from wining competition with the control channel in distributed wireless system [8], or assigned by a center node as in one hop access system [36]. We denote the duration of each access time as a slot.

The channel state consists of two elements: channel availability and link quality. Denote $a_i(j)$ as the availability of channel i in the j^{th} slot, and availability state $a_i(j) \in \{0, 1\}$, where $a_i(j) = 0$ indicates that the primary user is transmitting over channel i in the j^{th} slot, and $a_i(j) = 1$, otherwise. The channel quality is characterized by the temporary 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). Denote $q_i(j)$ as the quality of channel i in the j^{th} slot. We consider slow-varying Rayleigh fading channels, which is typical for multipath propagation environment [11], [17]. Thus the received temporary SNR is distributed exponentially [12], [37], 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. Both the channel idle probability vector $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ and the SNR mean vector $\Upsilon = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ are unknown to user at the beginning, but can be available through learning. Channel state is considered to be stable during T, as slot duration in OSA system is set to be much shorter than channel coherence time, as well as the sojourn time of primary user activities. Moreover, as the interval time between consecutive communication chances is relatively long in multi-user networks (as discussed in [8]), the channel states in different slots are commonly treated to be independent of each other. This assumption is consistent with previous studies [8]–[12], [26], [28]–[30], [32]. Also, there is another concern that, since the channel states are assumed i.i.d over time, there is no need to assume constant channel quality during T, and allowing the recall process could improve the results. The main reason is to protect primary users' communication. Since there is contention among users, and the primary users could use the licensed channel anytime, we need to set the duration T short enough for this concern. Thus, there is no chance to recall back the previous probed channels.

We depict the online accessing control process in Fig. 1. The s-SPA proceeds slot by slot. For a given slot, says slot j, s-SPA process can be described as follows. Firstly, user senses a channel $\phi_1(j)$ to acquire the channel availability $a_{\phi_1(j)}(j)$. If $a_{\phi_1(j)}(j) = 1$ (i.e., the sensed channel is idle), user further probes the channel via physical layer measurement mechanism (which also has been applied in [17]), acquiring temporary link

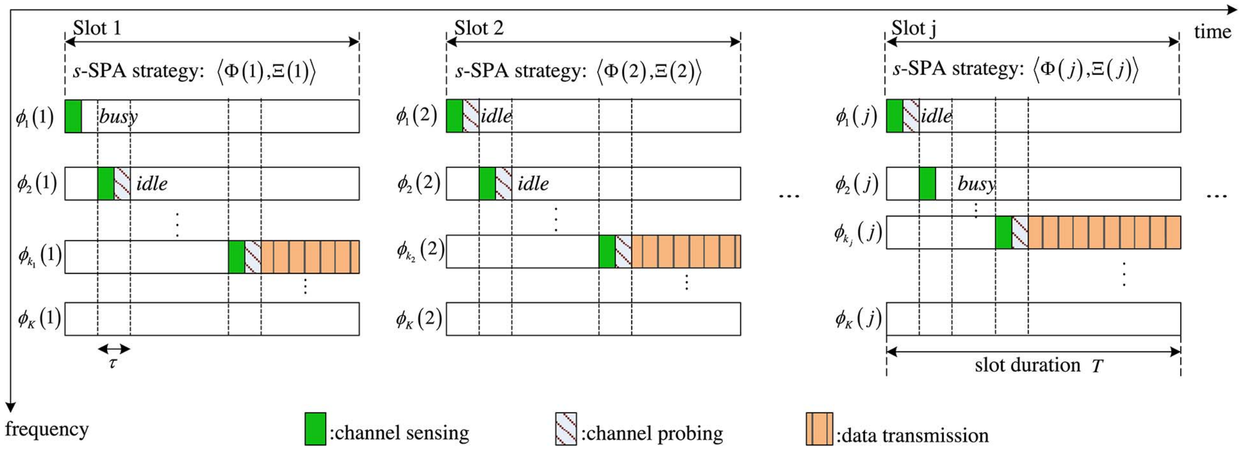


Fig. 1. Online sequential sensing, probing and accessing (*s*-SPA) control.

quality $q_{\phi_1(j)}(j)$. With the observed result, user needs to make a real time decision on whether to access the channel $\phi_1(j)$, or go on *s*-SPA process by switching to another channel, says $\phi_2(j)$. During the *s*-SPA process, if a channel is sensed to be busy, the user is forbidden to send measurement packet for primary user protection. However, the user still needs to wait for a constant channel probing time before switching to next channel. Such scheme is introduced for transceiver synchronization under the case that the channel availability of transmitter and receiver is different [11]. As a result, each sensing/probing step costs a constant time τ. Correspondingly, the maximum number of steps one could take in one slot is $K = \min\left(N, \left\lfloor \frac{T}{\tau} \right\rfloor\right)$, where $\lfloor \cdot \rfloor$ represents round-down function.

When user decides to access channel for data transmission after the k^{th} channel sensing/probing step, the immediate normalized throughput is given by

$$\begin{aligned} r(j) &= c_k \ln\left(1 + q_{\phi_k(j)}(j)\right) \\ &= (1 - k\beta)\ln\left(1 + q_{\phi_k(j)}(j)\right) \end{aligned} \tag{1}$$

where $\beta = \frac{\tau}{T}$ is a normalized observation cost, which is a factor to show the fraction of time a probing duration occupies the whole time slot. As we know, in evaluating the probing time overhead, the normalized β factor is used to evaluate this overhead. In our work, we use $c_k = 1 - k\beta$ to evaluate the pure data transmission time in each slot. The actual throughput can be easily obtained by scaling our reward[2] with a constant $\frac{T}{\ln 2}$.

We define the deterministic learning policy χ, mapping from the observation history $\mathcal{F}_{j-1}$ to a *s*-SPA strategy $\langle \Phi(j), \Xi(j) \rangle$ at each slot j, where $\Phi(j) = (\phi_1(j), \phi_2(j), \ldots, \phi_K(j))$ is a permutation of channels that determines the channel sensing/probing order in a slot, and $\Xi(j)$ is the corresponding accessing rule determining when to access which channel. For notation convenience, we define Ψ as the set of all possible sensing orders, and denote the m^{th} element in it as $\Phi_m = (\phi_1^m, \phi_2^m, \ldots, \phi_K^m)$. Correspondingly, the number of all possible sensing orders $|\Psi| = M = \binom{N}{K} K!$. Then, deriving a *s*-SPA strategy $\langle \Phi, \Xi \rangle$ in a slot includes:

1) selecting K channels from channel set Ω;
2) arranging the order of the selected K channels for sequential channel sensing/probing;
3) deriving an accessing rule for opportunistic channel accessing.

Our main goal is to devise a learning policy guiding the system converging to the throughput-optimal *s*-SPA strategy. Meanwhile, the accumulated throughput loss during the learning process should be as small as possible. We use *regret* value to characterize the accumulated throughput loss, which is defined as the gap between the accumulated reward gained by always using the perfect *s*-SPA strategy, and using the *s*-SPA strategy proposed by learning policy in each slot. Mathematically, the regret of learning policy χ up to slot L is

$$\rho_\chi(L) = LV^*_{\{\Theta,\Upsilon\}} - \sum_{j=1}^{L} {}_\chi V^{\langle \Phi(j), \Xi(j) \rangle}_{\{\Theta,\Upsilon\}} \tag{2}$$

Here, $V^*_{\{\Theta,\Upsilon\}}$ is the maximum expected throughput one could obtain in one slot under the environment $\{\Theta, \Upsilon\}$, which is achieved by user applying the ideal *s*-SPA strategy $\langle \Phi^*, \Xi^* \rangle$ derived with perfect statistical knowledge. $V^{\langle \Phi(j), \Xi(j) \rangle}_{\{\Theta,\Upsilon\}}$ is the corresponding reward user obtains with the strategy $\langle \Phi(j), \Xi(j) \rangle$ derived by learning policy χ.

The main notations and definitions of this paper are summarized in Table I.

IV. Understanding Sequential Accessing Control in *s*-SPA

In this section, we are aiming to demonstrate the fundamental tradeoff problem behind the sequential accessing control in *s*-SPA. We first propose a preliminary on the throughput-optimal sequential sensing, probing and accessing strategy with perfect statistics. After that, an intuitive strategy referred to as myopic learning policy is studied, and several observations are derived from the convergence analysis of this learning policy.

[2] The reward is directly related with the throughput. The difference is, when we use the reward for denotation, it mainly focuses on the regret analysis, where the reward value is evaluated with expectation value in the long run. On the other hand, when the term 'throughput' is used, it mainly focuses on the achievable data transmission rate, which is an instant value for evaluation.

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Description
N:	total number of channels
K:	number of maximum observing steps in one slot, where $K = \min\left(N, \frac{T}{\tau}\right)$
M:	total number of possible sensing orders
i:	channel index, $1 \le i \le N$
j:	slot index, $1 \le j \le L$
k:	step index during a slot, $1 \le k \le K$
δ:	a tunable parameter in IE-OSP, where $1-\delta$ is the confidence coefficient of the estimations
c_k:	normalized remaining time for data transmission after k^{th}: $c_k = 1 - k\beta$
Φ_m:	the m^{th} sensing order in sensing order set Ψ: $\Phi_m = \left(\phi_1^m, \ldots, \phi_K^m\right)$
ϕ_k^m:	ID of the k^{th} channel in sensing order Φ_m
Ξ:	accessing rule, described by a sequence of SNR thresholds, i.e.,$\Xi = (\Gamma_1, \Gamma_2, \ldots, \Gamma_K)$
Λ_1^m:	maximum expected reward user could obtain with sensing order Φ_m
$a_i(j)$:	availability of channel i in epoch j
$q_i(j)$:	quality of channel i in epoch j
$\rho_\chi(L)$:	regret of learning policy χ up to slot L
$\{\Theta, \Upsilon\}$:	channel statistics, where $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, and $\Upsilon = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$
$\left\{\hat{\Theta}, \hat{\Upsilon}\right\}$:	estimated channel statistics, where $\hat{\Theta} = \left\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N\right\}$ and $\hat{\Upsilon} = \{\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_N\}$
$\left\{\hat{\Theta}^u, \hat{\Upsilon}^u\right\}$:	upper confidence bound of estimated channel statistics, where $\hat{\Theta}^u = \left\{\hat{\theta}_1^u, \hat{\theta}_2^u, \ldots, \hat{\theta}_N^u\right\}$ and $\hat{\Upsilon}^u = \left\{\hat{\gamma}_1^u, \hat{\gamma}_2^u, \ldots, \hat{\gamma}_N^u\right\}$
$V_{\{\Theta', \Upsilon'\}}^{\langle\Phi', \Xi'\rangle}$:	expected throughput obtained by strategy $\langle\Phi', \Xi'\rangle$ when statistics is $\{\Theta', \Upsilon'\}$
$\langle\Phi(j), \Xi(j)\rangle$:	*s*-SPA strategy in the j^{th} slot

A. Optimal s-SPA Strategy Under Perfect Statistics

Given a channel sensing order Φ_m and the channel statistics $\{\Theta, \Upsilon\}$, obtaining the optimal *s*-SPA strategy can be formulated as an optimal stopping problem (OSP) [38]: during the sequential sensing/probing process, user makes a real time decision on when to stop channel sensing by accessing an observed channel. We formulate the problem as follows.

After sensing/probing channel ϕ_k^m, if the observed channel is idle with channel quality $q_{\phi_k^m}$, the achievable reward in step k is given by:

$$r_k^m = \begin{cases} c_k \ln\left(1 + q_{\phi_k^m}\right), c_k \ln\left(1 + q_{\phi_k^m}\right) > \Lambda_{k+1}^m \\ \Lambda_{k+1}^m, \qquad\qquad \text{else} \end{cases} \tag{3}$$

where $\Lambda_{k+1}^m = E[r_{k+1}^m]$ is the expected reward when user decides to skip the current channel under sensing order Φ_m.

Since in the last step K, the optimal choice is always to access the channel if it is available. Therefore,

$$\Lambda_K^m = E\left[r_K^m\right] = c_K E\left[\theta_{\phi_K^m} \ln\left(1 + q_{\phi_K^m}\right)\right]$$

Then, the expected reward in each step $\Lambda_{K-1}^m, \Lambda_{K-2}^m, \ldots, \Lambda_1^m$ can be obtained using backward deduction according to Eqn. (3).

Specifically, with the channel statistics $\{\Theta, \Upsilon\}$, the expected reward Λ_K^m is given by

$$\begin{aligned} \Lambda_K^m &= c_K \theta_{\phi_K^m} \int_0^\infty \log(1+q) \frac{1}{\gamma_{\phi_K^m}} e^{-\frac{q}{\gamma_{\phi_K^m}}} dq \\ &= c_K \theta_{\phi_N^m} e^{\frac{1}{\gamma_{\phi_K^m}}} \mathrm{Ei}\left(1, \frac{1}{\gamma_{\phi_K^m}}\right) \end{aligned} \tag{4}$$

where function Ei is the exponential integral function defined as $\mathrm{Ei}(1, x) = \int_x^\infty \frac{e^{-t}}{t} dt$ for $x > 0$.

For $1 \le k < K$, the Λ_k^m can be computed using the following recursion [8], [12], [38].

$$\begin{aligned} \Lambda_k^m &= \left(1 - \theta_{\phi_k^m}\right) \Lambda_{k+1}^m \\ &\quad + \theta_{\phi_k^m} \Lambda_{k+1}^m \int_0^{c_k \log(1+q) \le \Lambda_{k+1}^m} \frac{1}{\gamma_{\phi_k^m}} e^{-\frac{q}{\gamma_{\phi_k^m}}} dq \\ &\quad + c_k \theta_{\phi_k^m} \int_{c_k \log(1+q) > \Lambda_{k+1}^m}^\infty \log(1+q) \frac{1}{\gamma_{\phi_k^m}} e^{-\frac{q}{\gamma_{\phi_k^m}}} dq \\ &= \left(1 - \theta_{\phi_k^m}\right) \Lambda_{k+1}^m + \theta_{\phi_k^m} \Lambda_{k+1}^m \int_0^{e^{\frac{\Lambda_{k+1}^m}{c_k}} - 1} \frac{1}{\gamma_{\phi_N^m}} e^{-\frac{q}{\gamma_{\phi_N^m}}} dq \\ &\quad + c_k \theta_{\phi_k^m} \int_{e^{\frac{\Lambda_{k+1}^m}{c_k}} - 1}^\infty \log(1+q) \frac{1}{\gamma_{\phi_N^m}} e^{-\frac{q}{\gamma_{\phi_N^m}}} dq \\ &= \Lambda_{k+1}^m + c_k \theta_{\phi_k^m} e^{\frac{1}{\gamma_{\phi_k^m}}} \mathrm{Ei}\left(1, \frac{e^{\frac{\Lambda_{k+1}^m}{c_k}}}{\gamma_{\phi_k^m}}\right) \end{aligned} \tag{5}$$

According to Eqn. (3), the optimal stopping rule, i.e., optimal accessing strategy, is completely specified by the reward sequence $(\Lambda_1^m, \Lambda_2^m, \ldots, \Lambda_K^m)$: access the channel ϕ_k^m after the k^{th} sensing/probing step, if the channel is idle with achievable throughput $c_k \ln(1 + q_{\phi_k^m}) \ge \Lambda_k^m$. Otherwise, user could switch to channel ϕ_{k+1}^m for another sensing/probing step. Obviously, the accessing rule can be further simply described as a sequence of SNR thresholds, denoted as $\Xi_m = (\Gamma_1^m, \Gamma_2^m, \ldots, \Gamma_K^m)$. Hence, the access threshold $\Gamma_k^{m^*}$ is given by

$$\Gamma_k^{m^*} = \begin{cases} e^{\frac{\Lambda_{k+1}^{m^*}}{c_k}} - 1, & 1 \le k < K \\ 0, & k = K \end{cases} \tag{6}$$

Finally, Λ_1^m is the maximum expected reward user could obtain with sensing order Φ_m. The sensing order Φ_{m^*} generating the maximum $\Lambda_1^{m^*}$ is then the optimal sensing order under the given scenario with channel statistics $\{\Theta, \Upsilon\}$.

B. Complexity Analysis

An intuitive solution when channel statistics is unavailable is that, always deriving *s*-SPA strategy maximizing immediate throughput in each slot. Meanwhile, refined statistics by updating the estimations of channels have been observed.

During the slot by slot decision-making process, the estimations of channels are obtained by recording and updating the following four variables on each channel: $\hat{\theta}_i(j)$, $n_i^s(j)$, $\hat{\gamma}_i(j)$ and $n_i^p(j)$. Where $\hat{\theta}_i(j)$ is the estimated idle probability of channel i

up to slot j, and $n_i^s(j)$ is the times channel i having been sensed till slot j. They are initialized to be zero and updated as follows:

$$\hat{\theta}_i(j) = \begin{cases} \frac{\hat{\theta}_i(j-1)n_i^s(j-1)+d_i^j}{n_i^s(j-1)+1}, & \text{if channel } i \text{ is sensed} \\ \hat{\theta}_i(j-1), & \text{else} \end{cases} \quad (7)$$

$$n_i^s(j) = \begin{cases} n_i^s(j-1)+1, & \text{if channel } i \text{ is sensed} \\ n_i^s(j-1), & \text{else} \end{cases} \quad (8)$$

Similarly, $\hat{\gamma}_i(j)$ is the estimated SNR mean of channel i up to slot j, and $n_i^p(j)$ is the times channel i having been probed till slot j. They are updated as follows:

$$\hat{\gamma}_i(j) = \begin{cases} \frac{\hat{\gamma}_i(j-1)n_i^p(j-1)+q_i^j}{n_i^p(j-1)+1}, & \text{if channel } i \text{ is probed} \\ \hat{\gamma}_i(j-1), & \text{else} \end{cases} \quad (9)$$

$$n_i^p(j) = \begin{cases} n_i^p(j-1)+1, & \text{if channel } i \text{ is probed} \\ n_i^p(j-1), & \text{else} \end{cases} \quad (10)$$

Since the throughput in each slot is always maximized with the currently estimated statistics, and the channel statistics is refined slot by slot with myopic learning policy, it turns out to be a good solution for our concern.

A learning policy of non-zero-regret is equivalent to the statement that, using the learning policy, system may converge to a non-optimal solution as time goes on.

C. Challenges

However, it is really challenging to achieve optimal control because that, the reward of utilizing and learning in s-SPA process are hard to quantify. Moreover, these two rewards are both related to the sensing order and accessing rule. Specifically,

1) The closed expression of expected throughput is unavailable, which has been shown in Section IV-A. Moreover, for throughput optimal channel access scheme, the channel sensing order relies on the long-term quality, which would not show a direct relationship to the channel probing results. Temporary channel quality is not stable and would possibly contradict to the results in optimal throughput strategy.
2) Considering the exploration process, channels being learnt during a slot are unpredictable. Although intuitively one could improve channel statistics exploration by increasing the accessing thresholds, the exact relationship is complicated, and can only be described in a probabilistic way.

As a result, to achieve optimal s-SPA strategy as well as reduce the throughput loss during the learning process, one needs to consider exploring the inaccurately estimated channels while pursuing immediate reward maximization, by jointly optimizing the sensing order selection process across slots and the opportunistic accessing control process in each slot.

V. IE-OSP Algorithm

In this section, we propose the IE-OSP (i.e., Interval Estimation in OSP analytical framework) online policy, in which the statistics learning and diversity utilization processes are seamlessly integrated together for efficient spectrum access. We further analyze the convergence of the proposed policy, and prove that the IE-OSP is guaranteed to converge to the optimal s-SPA strategy with a controlled probability.

A. Algorithm Description

In our algorithm, the basic idea for guiding our system being converged to the optimal s-SPA strategy is to minimize the unreachable probability of inaccurate channels during the s-SPA process. Meanwhile, the optimal stopping analytical framework is used during the s-SPA process for obtaining diversity gain during the learning process.

For each channel, the following four variables are recorded and updated during s-SPA process for decision-making, i.e., the estimated channel idle probability $\hat{\theta}$, the times channel having been sensed n^s, the estimated channel SNR mean $\hat{\gamma}$ and the times channel having been probed n^p. They are updated according to (7)–(10), respectively.

We leverage the confidence interval bound to characterize the inaccuracy of statistical estimation. Define parameter $0 < \delta < 1$, where $1-\delta$ is the confidence coefficient of the estimations. Then, the $1-\delta$ upper confidence bound of the channel idle probability and the channel SNR mean are respectively given by

$$\hat{\theta}_i^u(j) = \min\left\{1, \hat{\theta}_i(j) + \sqrt{\frac{-\log\delta}{2n_i^s(j)}}\right\} \quad (11)$$

$$\hat{\gamma}_i^u(j) = \min\left\{q_{max}, \hat{\gamma}_i(j) + q_{max}\sqrt{\frac{-\log\delta}{2n_i^p(j)}}\right\} \quad (12)$$

where q_{max} denotes the maximum value of temporary received SNR. It is reasonable to restrict q with an upper bound q_{max}, since the probability that temporary SNR is larger than q_{max} approximates to zero if the value of q_{max} is large enough.

Then, the IE-OSP can be described as follows. Firstly, sequentially sense/probe channels until all channels are probed at least once (from line 2 to line 13). Note that, the pseudo code from line 5 to line 8 operates for the case where channel is available, and the channle is probed with property channel quality updating operations. If the channel is busy, we should move forward for next channel. Line 8 and line 10 in the pseduo are using the same operations to visit next available channels. After that, always choose the s-SPA strategy $\langle\Phi_{m^*}(j), \Xi_{m^*}^u(j)\rangle$ that achieves $\max_m \Lambda_1^{m,u}(j)$ in slot j, where $\Lambda_1^{m,u}(j)$ is a virtual throughput value defined as the maximum achievable throughput one could achieve if the real statistics is $\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\}$ (from line 14 to line 21). Obviously, $\langle\Phi_{m^*}(j), \Xi_{m^*}^u(j)\rangle$ can be derived easily with $\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\}$, using the optimal stopping analytical framework we introduced in Section IV-A.

The pseudo-code of the IE-OSP algorithm is shown as in Fig. 2.

B. Convergence Analysis

In this subsection, we analyze the convergence of IE-OSP algorithm, because the optimal convergence point is critical to online learning policy in the long run. The main result

```
IE-OSP Policy
Initialize: for all $1 \leq i \leq N$: $\hat{\theta}_i = 0$, $n_i^s = 0$, $\hat{\gamma}_i = 0$, $n_i^p = 0$, $S_0 = \Omega$, $l = 1$, $k = 1$;
while $S_0 \neq \emptyset$ do
    Sense a random channel $i \in S_0$;
    $k = k + 1$, update $\hat{\theta}_i(l)$ and $n_i^s(l)$ according to Eqn.(7) and (8), respectively;
    if $a_i(l) == 1$ then
        Probe and then access channel $i$;
        Update $\hat{\gamma}_i(l)$ and $n_i^p(l)$ according to Eqn.(9) and (10), respectively;
        $l = l + 1$, $k = 1$, $S_0 = S_0 \setminus \{i\}$;
    else if $k == K + 1$ then
        $l = l + 1$, $k = 1$, $S_0 = S_0 \setminus \{i\}$;
        Wait for next communication slot;
    end if
end while
for $j = l : L$ do
    for $m = 1 : M$ do
        Select sensing order $\Phi_m$
        for $k = K : 1$ do
            Compute $\hat{\Lambda}_k^{m,u}(j)$ with $\left\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\right\}$ according to Eqn. (4) or (5);
            Compute $\Gamma_k^{m,u}(j)$ according to Eqn. (6);
        end for
    end for
    Determine $m^*(j) = \arg\max_{1 \leq m \leq M} \left\{\hat{\Lambda}_1^{m,u}(j)\right\}$;
    Proceed $s$-SPA with strategy $\langle \Phi_{m^*}(j), \Xi_{m^*}^u(j) \rangle$;
    Update $\hat{\theta}_i(j)$, $n_i^s(j)$, $\hat{\gamma}_i(j)$ and $n_i^p(j)$, according to Eqn.(7), (8), (9) and (10), respectively;
end for
```

Fig. 2. Algorithm description on IE-OSP.

can be described by the following theorem, which provides a theoretical convergence guarantee for our proposed policy.

Theorem 1: Using *IE-OSP*, system converges to the throughput-optimal s-SPA strategy with probability at least $(1-\delta)^{2(N-1)}$. Particularly, when $\forall i : \theta_i < 1$, it converges to optimal s-SPA strategy with probability at least $(1-\delta)^{2(N-K)}$, where $1-\delta$ is used to provide bounds to the statistical channel features in channel idle probability and SNR mean, which have been formally defined in Eqn. (11), and Eqn. (12).

Before proving this theorem, it is worth noting that, the performance analysis, e.g., the regret analysis, is typically identical to previous studies [25], [33]. The difference is, since the strategy is mixed with partially known knowledge, and channel dynamics are fully used, there is no fixed optimal policy. The only concern in this work, is to know the probability that the algorithm could converge to the optimal point. To this end, the probability analysis is also challenging in our concern. Thus, an analytical bound is presented to instead of accurate p.d.f. based analysis.

Proof: To prove Theorem 1, we introduce the Chernoff-Hoeffding bound inequalities first.

Lemma 1: (Chernoff-Hoeffding bound) [39] Let $X_1, \ldots, X_n$ be random variables with range [0, 1], such that $E[X_t | X_1, \ldots, X_{t-1}] = \mu$. Moreover, let $S_n = X_1 + \ldots + X_n$. Then, for any $a > 0$,

$$\Pr[S_n \geq n\mu + a] \leq e^{-\frac{2a^2}{n}}$$

and

$$\Pr[S_n \leq n\mu - a] \leq e^{-\frac{2a^2}{n}}$$

According to Lemma 1, we can derive the following corollary directly.

Corollary 1: Let $\mathcal{D}$ be a distribution with support in [0, 1], and $E_{X \sim \mathcal{D}}[X] = \theta$. Let $X_1, \ldots, X_n$ be drawn independently from $\mathcal{D}$, and $\hat{\theta} = \frac{1}{n} \sum_t X_t$. Then

$$\Pr\left[\theta \leq \hat{\theta} + \sqrt{\frac{-\log \delta}{2n}}\right] \geq 1 - \delta$$

and

$$\Pr\left[\theta \geq \hat{\theta} - \sqrt{\frac{-\log \delta}{2n}}\right] \geq 1 - \delta$$

Moreover, let $\mathcal{D}$ denote a distribution with support in [0, q_{max}], and $E_{X \sim \mathcal{D}}[X] = \gamma$. Let $X_1, \ldots, X_n$ be drawn independently from $\mathcal{D}$, and $\hat{\gamma} = \frac{1}{n} \sum_t X_t$. Then

$$\Pr\left[\gamma \leq \hat{\gamma} + q_{max}\sqrt{\frac{-\log \delta}{2n}}\right] \geq 1 - \delta$$

and

$$\Pr\left[\gamma \geq \hat{\gamma} - q_{max}\sqrt{\frac{-\log \delta}{2n}}\right] \geq 1 - \delta$$

Proof: Corollary 1 is directly derived from Lemma 1. ■

Let θ_i' and γ_i' be the supposed channel statistics of idle probability and the averaged SNR value on channel i respectively, and let θ_i and γ_i be the real corresponding channel statistics. Denote $\langle \Phi', \Xi' \rangle$ (a pair of sensing order and accessing rule) as the throughput-optimal strategy for sequential channel sensing, probing and accessing (s-SPA) in the case that the channel statistics is $\{\Theta', \Upsilon'\}$, i.e., $\{\theta_1', \ldots, \theta_N'; \gamma_1', \ldots, \gamma_N'\}$. We have

Lemma 2: Under any given strategy $\langle \Phi', \Xi' \rangle$, if there exists an overestimated channel, it could be observed with high probability.[3]

Proof: We prove this lemma by contradiction.

Denote $V_{statistic}^{solution}$ as the expected throughput obtained by user using *solution* for sequential channel sensing and accessing, while the actual channel statistics is *statistic*. Thus:

- $V_{\{\Theta', \Upsilon'\}}^{\langle \Phi', \Xi' \rangle}$ is the maximum throughput one could obtain in the supposed scenario $\{\Theta', \Upsilon'\}$;
- $V_{\{\Theta, \Upsilon\}}^{\langle \Phi, \Xi \rangle}$ is the maximum actually achievable throughput in the scenario $\{\Theta, \Upsilon\}$;
- $V_{\{\Theta, \Upsilon\}}^{\langle \Phi', \Xi' \rangle}$ is the expected throughput one could obtain when using $\langle \Phi', \Xi' \rangle$ in the scenario $\{\Theta, \Upsilon\}$.

[3] "With high probability" means that, you can change the conditions slightly to make the probability of failure very small. The usefulness of this concept is from the power of the statement. The statement is parameterized to allow the probability to vary as necessary to prove other statements.

Suppose that for all i except i^*: $\theta_i' = \theta_1$, $\gamma_i' = \gamma_i$, while i^* is the overestimated channel, i.e., it falls into one of the following three conditions: 1) $\theta_{i^*}' > \theta_{i^*}, \gamma_{i^*}' = \gamma_{i^*}$; 2) $\theta_{i^*}' = \theta_{i^*}, \gamma_{i^*}' > \gamma_{i^*}$; and 3) or $\theta_{i^*}' > \theta_{i^*}, \gamma_{i^*}' > \gamma_{i^*}$. Then, we have

$$V_{\{\Theta',\Upsilon'\}}^{\langle\Phi',\Xi'\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi,\Xi\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi',\Xi'\rangle} \tag{13}$$

The statement that channel i^* would never be observed under the strategy $\langle\Phi', \Xi'\rangle$ is equivalent to that, the s-SPA process would stop before arriving channel i^*. If so, we have

$$V_{\{\Theta,\Upsilon\}}^{\langle\Phi',\Xi'\rangle} = V_{\{\Theta',\Upsilon'\}}^{\langle\Phi',\Xi'\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi,\Xi\rangle}$$

which contradicts the inequality (13). Hence, we can conclude that the statement is false. In other words, the overestimated channel would be observed with probability 1 as time goes on.■

We now prove Theorem 1 using Corollary 1 and Lemma 2.

Since sub-optimal convergence only happens when there exists at least one inaccurately estimated channel, where the statistics of this channel would never be updated again. Suppose that user converges to a state, i.e., a s-SPA solution, where the maximum number of achievable steps in each slot is k. Then, according to Lemma 2, the state is sub-optimal if and only if there exists some underestimated channel in remaining $N - k$ channels.

For the sake of convenient description, we denote the set of remaining channels as $S_r = \{k+1, k+2, \ldots, N\}$. For each $i \in S_r$, $p_i = \Pr[\theta_i' \le \theta_i$ or $\gamma_i' \le \gamma_i$. As in IE-OSP, we treat $\theta_i' = \theta_i^u = \hat{\theta}_i + \sqrt{-\frac{\log\delta}{2n_i^s}}$ and $\gamma_i' = \gamma_i^u = \hat{\gamma}_i + q_{max}\sqrt{-\frac{\log\delta}{2n_i^p}})$, according to Corollary 1, we have that $\Pr[\theta_i' \le \theta_i] \le \delta$, $\Pr[\gamma_i' \le \gamma_i] \le \delta$. Thus, for all i, $p_i \le p = 1 - (1-\delta)^2$. Then, the probability $P_{sub-opt}$ that system converges to a sub-optimal solution is bounded by

$$\begin{aligned} P_{sub-opt} &\le C_{N-k}^1 p\,(1-p)^{N-k-1} + C_{N-k}^2 p^2\,(1-p)^{N-k-2} \\ &\quad + \cdots + C_{N-k}^{N-k-1} p^{N-k-1}\,(1-p) + p^{N-k} \\ &= \left[p + (1-p)\right]^{N-k} - (1-p)^{N-k} \\ &= 1 - (1-\delta)^{2(N-k)} \end{aligned} \tag{14}$$

Consequently, the probability that system could converges to optimal solution is bounded by

$$P_{opt} \ge (1-\delta)^{2(N-k)} \tag{15}$$

As user needs to sense and probe at least one channel in each slot, thus $k \ge 1$, then we can derive the following probability of optimal convergence.

$$P_{opt} \ge (1-\delta)^{2(N-1)} \tag{16}$$

Particularly, when all the channel idle probabilities are less than 1, which means that when system converges to a state, all the K channels in the sensing order will be observed as time goes on (since the probability of all channel are busy is bigger than zero). In such case, we have the following statement.

$$P_{opt} \ge t(1-\delta)^{2(N-K)} \tag{17}$$

This completes the proof of Theorem 1. ■

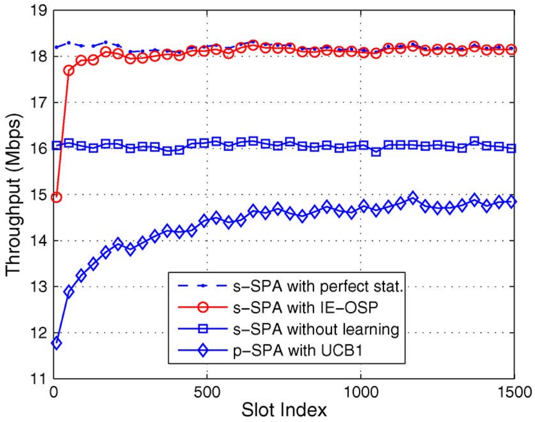


Fig. 3. Comparison on expected throughput with respect to time.

VI. Performance Evaluations

In this section, we evaluate and analyze the performance of the proposed online sequential accessing algorithm via simulations. We run our simulation code with Matlab, and an IBM X210 laptop. Our experiment settings are as follows. The idle probabilities and SNR means of independent channels are randomly generated respectively in range [0, 1] and [0, 15] dB for each round. Then, the states of channels (i.e. availability and link quality) in each slot are generated independently according to the idle probability vector as well as SNR mean vector. The channel bandwidth is set to be 6 MHz, and three channels are considered here. The normalized channel sensing/probing cost $\beta = 0.1$. The results are averaged from 1000 rounds of independent experiments, where each run lasts at least 1500 time slots.

A. Throughput Analysis

In this subsection, four policies are running under the same environment for performance comparison, briefly described as follows.

- p-*SPA with UCB1*: existing online learning solution for opportunistic channel access, in which user selects one channel to sense/access in each slot according to UCB1 [27] algorithm. Such learning policy is proved to be order-optimal in p-SPA system [26];
- s-*SPA without learning*: an intuitive method in s-SPA system without learning. User sequentially senses/probes with a random sensing order and access the first idle channel for transmission;
- s-*SPA with IE-OSP*: our proposed method, where user sequentially senses, probes and accesses according to online algorithm IE-OSP;
- s-*SPA with perfect stat.*: an ideal s-SPA strategy derived with perfect channel statistics, which leads to maximum achievable throughput.

We first study the system throughput as a function of time in Fig. 3. As depicted in Fig. 3,

1) both learning algorithms are effective in improving system throughput. This is clearly shown in the figure, where the

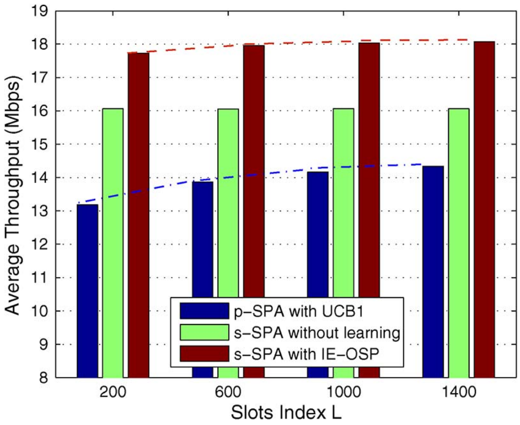


Fig. 4. Comparison on accumulated reward in the first L slots.

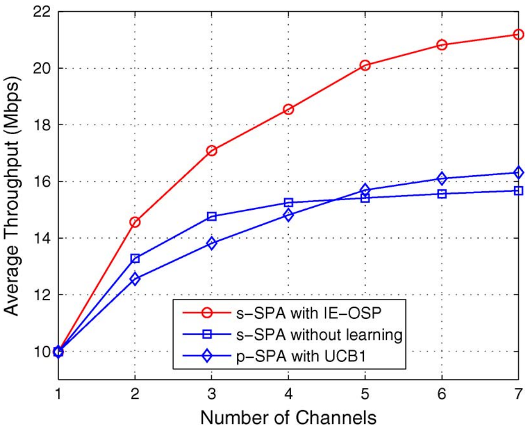


Fig. 5. Comparison on accumulated reward with respect to number of channels.

expected throughput of both p-*SPA with UCB1* and s-*SPA with IE-OSP* are increasing with time.

2) there is still a considerable gap compared with the maximum achievable throughput (i.e., the achievable throughput obtained by s-*SPA with perfect stat.*) by using existing solutions. On one hand, compare the throughput of existing learning method p-*SPA with UCB1* with that of s-*SPA with perfect stat*. It shows about 3 Mbps throughput loss even at the time $t = 1500$, where the learning algorithm converges almost to the optima status. Such a gap mainly arises from the fact that existing learning method is incompatible with temporary opportunity exploitation. On the other hand, the intuitive algorithm for exploiting diversity, i.e., s-*SPA without learning*, shows a constant gap of about 2 Mbps, comparing with the ideal strategy.
3) our proposed algorithm IE-OSP bridges the throughput gap effectively. As shown in figure, the obtained throughput of IE-OSP algorithm approaches to the ideal goal in about 500 slot.

We further investigate the accumulated reward of the three algorithms. Accumulated award in the first L slots is defied as the total transmitted bits from the beginning time, i.e., $j = 1$, to the instant $j = L$. Actually, the accumulated reward is the most concerned metric from the perspective of the user. The results are shown in Fig. 4. Here, we leverage the average throughput in the first L slots to characterize the real value of accumulated reward, which is mathematically defined as $\frac{1}{L}\sum_{j=1}^{L} r(j)$. In the figure, the average throughputs of the three practical schemes with different Ls are given. It clearly shows that, our proposed method outperforms the other two schemes in almost any time, with respect to the accumulated reward. The advantage of our proposed algorithm in time from 200 to 1400 are apparently shown in the figure. More precisely, our learning method outperforms s-*SPA* without learning as soon as $j = 50$, and outperforms p-*SPA with UCB1* in arbitrary time. In other words, applying our proposed scheme earn profits, even in where the communication session duration is relatively short. Moreover, as the gap between the average throughputs of the three schemes are tending towards stability, it is no doubt that user would gain more by applying our proposed scheme as the session duration increases.

All the above results are derived from the scenario with a constant number of channels ($N = 3$). As the number of channels is almost the most important attribute of a wireless network and relates much to the system performance, we evaluate the three schemes in scenarios with different channels in the following part of this subsection, so as to investigate the impact of channel number. We adopt the accumulated reward in the first 1500 slots as the main metric to show the impact of channel number. Similarly, we leverage ‘average throughput’ to characterize the real value of accumulated reward. With the number of channels ranging from 1 to 7, we depict the results as shown in Fig. 5. All the three curves are increasing with the number of channels; however, with different rising characteristics:

1) s-SPA without learning scheme, it shows to be a rapid growth within $N \leq 3$ (higher increasing rate compared with p-*SPA with UCB1* scheme). Such growth in throughput comes from the fact that, as the number of channels increases, it is more likely to find an available channel to use by sequentially observing channels in a slot. In other words, the increasing channels enrich diversity in temporary channel status, and thus benefit the scheme with opportunity exploitation. However, due to lack of advanced accessing control strategy, the s-*SPA without learning* scheme would fail to exploit temporary opportunity efficiently. This is why the increasing trend flattens soon when $N > 4$.
2) for the p-SPA with UCB1 scheme, the growth comes from the increasing diversity of channels' statistics. Specifically, as the expected reward of the single statistic-optimal channel is increasing with the total number of the channels, user gains more as the number of channels increases, since it could learn to converge to the optimal channel by using p-*SPA with UCB1*. Moreover, the average throughput of p-*SPA with UCB1* increases more slowly than that of s-*SPA without learning* within few channels, e.g., 14 with sustained growth.
3) our proposed s-*SPA with IE-OSP* scheme increases with the number of channels more rapidly and lasting. By using s-*SPA with IE-OSP*, user sequentially senses/probes and accesses with near-optimal strategy soon by learning.

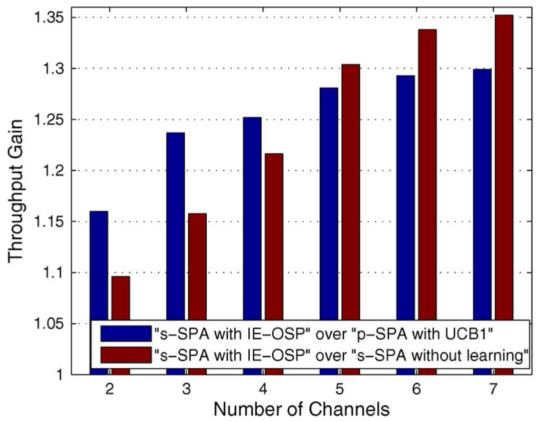


Fig. 6. Throughput gain of *s*-SPA with IE-OSP over the other two schemes.

The temporary opportunity among channels are fully and efficiently exploited. As a result, the throughput gap between our proposed policy and the existing policies is increasing with number of channels, e.g., about 5 Mbps throughput improvement is attained at $N = 7$.

To further investigate the throughput improvement of our proposed scheme over the other two schemes, we depict the throughput gain as a function of the number of channels. The throughput gain is defined as the ratio between average throughput in the first 1500 slots of s-*SPA with IE-OSP* scheme over that of p-*SPA with UCB1* or s-*SPA without learning*, respectively. As depicted in Fig. 6, with the increasing number of channels, the candidate channels are more than ever, thus the potential channel quality improvement is expected, since the probability of probing a high quality channel could be larger than ever. Specifically, we learn from this figure that:

1) the throughput gain of our opposed scheme over the other two schemes are increasing with the number of channels, which means that the proposed policy would benefit more in the scenarios with more channels.
2) at least 9.5% improvement in average throughput is achieved with our proposed scheme. This value is attained at $N = 2$ comparing with s-*SPA without learning*. When compared with p-*SPA with UCB1*, it exceeds 15%.
3) 25∼30% throughput improvement can be obtained in most scenarios, as almost all existing OSA networks are equipped with more than 5 channels.

B. Convergence Analysis

In this subsection, we evaluate the convergence property of our proposed learning algorithm by analyzing regret. Regret is an important metric for online policies, where the definition[4] of regret is presented in Eqn. (2). An online learning algorithm with higher regret means more throughput loss during learning process. Moreover, it has been proven by Lai and Robbins [40] that no policy can do better than logarithmic increasing regret in time. In other words, an online policy with logarithmic regret in time is order-optimal.

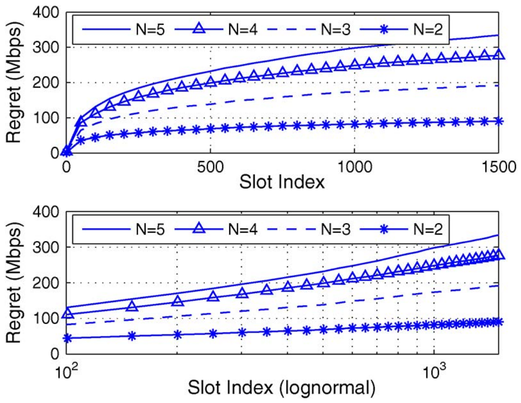


Fig. 7. Regret with respect to time.

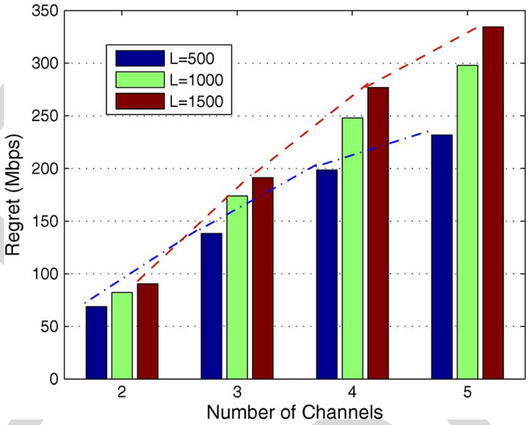
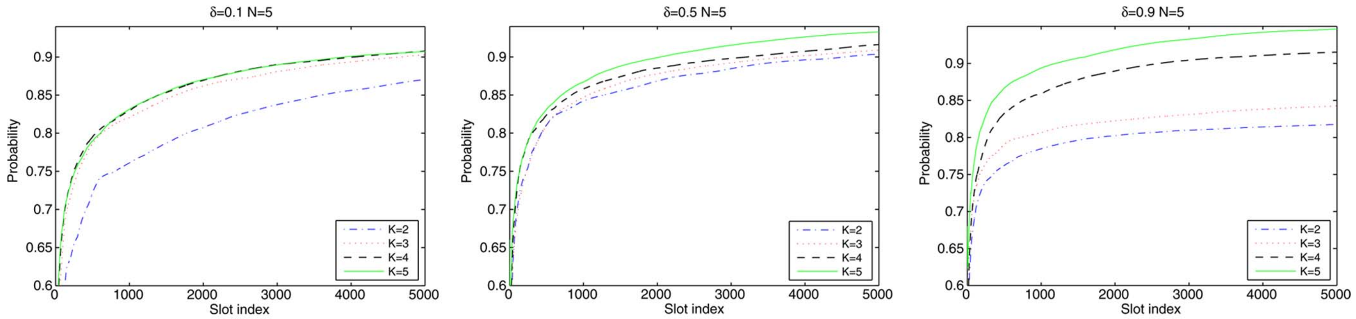


Fig. 8. Regret vs. increased number of channels.

In Fig. 7, we depict the regret of IE-OSP policy as a function of slot index, so as to study the increasing rate of regret over time. To show more widely, we present all the curves with N ranging from 2 to 5. Intuitively, we find from the upper part of this figure that, all the curves of regret show a logarithmic increasing trend over time. To further verify this logarithmic increasing property, we re-plot the regret curves in the lower part of this figure, where X-axis ranges from 100 to 1500 and is in a logarithmic form. The transformed curves show almost linear increasing trend. This verifies that, the regret is in at least asymptotically logarithmic rate, even if it is not in optimal logarithmic rate

Further, we study the increasing trend of regret with respect to the number of channels. As the regret increases infinitely with the number of slots, we take three typical value of L to determine the regret for comparison. Specifically, for each N, we depict the value of $L = 500$, $L = 1000$, and $L = 1500$. The results are presented in Fig. 8. It is intuitive that the regret values increases when adds the number of channels. This is reasonable, since the increasing number of channels extends the learning space, and thus results in higher throughput loss for learning. In spite of this, it is encouraging that the regret is sub-linearly increasing with the number of channels. As shown in the regret envelope curves, where the blue dots and red dashed line sketches the increasing trace of $\rho(500)$ and $\rho(1500)$

[4] As in our simulation, regret is the accumulated throughput loss of applying s-*SPA with IE-OSP*, comparing with always using s-*SPA with perfect stat*.



(a) (b) (c)

Fig. 9. Comparison between simulation and theoretical results. (a) $\delta = 0.1$ and $N = 5$; (b) $\delta = 0.5$ and $N = 5$; (c) $\delta = 0.9$ and $N = 5$.

respectively. Such desirable property makes the learning algorithm scalable.

C. Discussion

1) Impact of Secondary User and Reliability: The channel probing failure and primary user occupancy will lead to different results. In previous studies [41], [42], we discussed the probability of channel probing failure and effects for the statistical behavior of the primary users. Moreover, it is worth noting that, in our scheme, when the channel probing failure and primary user occupancy is stable, say, providing a probability or distribution for it, our IE-OSP policy could be adaptive to such cases. Because the threshold value could be adjustable according to this probabilistic distribution, which could be further evaluated by the rewards.

2) Validating the Theoretical Analysis: To show the matching effects of the proposed algorithm and theorem 1, we make an extended experimental study on the comparisons between the results we got from simulation study and theoretical analysis. In our simulation study, we evaluate the matching rate of the proposed algorithm and theoretical results. For each run, if the result in simulation study equals to that of theoretical analysis, the matching times could be increased by 1. And the overall matching rate is the accumulated matching times to the total number of running times.

As depicted in Fig. 9, the Y-axis denotes the matching rate with probabilistic form. We set the parameter N, K, and δ with different values, and evaluate the matching rate. To show the trends, especially when the number of probing times increases, we make observations for different values of K. This feature also validates our basic idea, i.e., providing more opportunities of probing could improve the throughput gain in temporarily high SNR channels. Large-scale evaluation needs computational intensive operations, and the theoretical results could guide us with the converging trends for the regret value. Furthermore, Fig. 10 depicts the convergenc feature of our proposed protocol, when the theoretical regret value is concerned. In that, we observe the convergence property when the parameter δ is concerned. When the confidence interval is involved, the convergence probability increases with the δ, which means, the convergence probability could be higher than the case with lower confidence interval. On the other hand, a theoretical bound value with higher confidence interval could be more difficult to achieve.

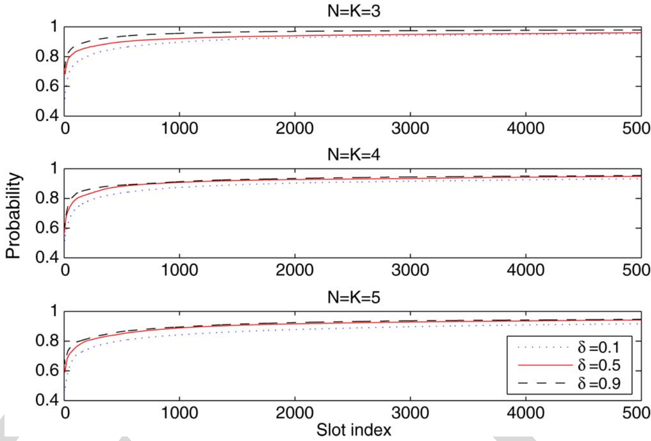


Fig. 10. Convergence property of the simulation results.

VII. Conclusion

In this work, channel learning and opportunity utilization are jointly considered for maximizing system overall throughput in an unknown environment. The sensing/probing order and accessing rule are dynamically adapted slot by slot, so as to achieve better tradeoff between maximizing diversity exploitation in current slot and exploring more channels for refining statistics. A near optimal online learning policy, so called IE-OSP, is proposed, which balances the statistics exploration and diversity exploitation by integrating confidence interval estimation into the optimal stopping analytical framework. We prove that, by using the proposed algorithm, system is guaranteed to converge to the optimal s-SPA strategy with a controllable probability. Simulation results further show that the regret of IE-OSP is asymptotically logarithmic in time and sub-linear in the number of channels, which respectively shows the optimality and scalability of our proposed learning policy. Compared with existing solutions, our proposed algorithm achieves more than 25% throughput gain in most scenarios.

In future work, we are to implement our policy to a cognitive radio platform built on USRP [43], [44], and provide a working system in real deployment [45] for validation.

References

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw. J.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006.

[2] I. F. Akyildiz, W. yeol Lee, and K. R. Chowdhury, "CRAHNs: Cognitive radio ad hoc networks," *Ad Hoc Netw.*, vol. 7, no. 5, pp. 810–836, Jul. 2009.

[3] J. Jeung, S. Jeong, and J. Lim, "Outband sensing-based dynamic frequency selection (DFS) algorithm without full DFS test in IEEE 802.11h protocol," *IEICE Trans.*, vol. 95-B, no. 4, pp. 1295–1296, Apr. 2012.

[4] "IEEE 802.22-2011(TM) standard for cognitive wireless regional area networks (RAN) for operation in tv bands." [Online]. Available: http://www.ieee802.org/22/

[5] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M. Welsh, "Whitespace networking with Wi-Fi like connectivity," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 27–38, Aug. 2009.

[6] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor, "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 3, pp. 101–116, May 2012.

[7] K. Balach, S. R. Kadaba, and S. Nanda, "Channel quality estimation and rateadaptation for cellular mobile radio," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 7, pp. 1244–1256, Jul. 1999.

[8] A. Sabharwal, A. Khoshnevis, and E. Knightly, "Opportunistic spectral usage: Bounds and a multi-band CSMA/CA protocol," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 533–545, Jun. 2007.

[9] S. Guha, K. Munagala, and S. Sarkar, "Information acquisition and exploitation in multichannel wireless systems," *arXiv preprint arXiv: 0804.1724*, 2008.

[10] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM Trans. Netw.*, vol. 17, no. 6, pp. 1805–1818, Dec. 2009.

[11] T. Shu and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," in *Proc. MobiCom*, 2009, pp. 37–48.

[12] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 297–307, Jan. 2009.

[13] Y. Zhou *et al.*, "Almost optimal channel access in multi-hop networks with unknown channel variables," in *Proc. IEEE ICDCS*, 2014, pp. 461–470.

[14] R. Fan and H. Jiang, "Channel sensing-order setting in cognitive radio networks: A two-user case," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4997–5008, Nov. 2009.

[15] J. Zhao and X. Wang, "Channel sensing order in multi-user cognitive radio networks," in *Proc. IEEE DYSPAN*, 2012, pp. 397–407.

[16] Y. Pei, Y.-C. Liang, K. C. Teh, and K. H. Li, "Energy-efficient design of sequential channel sensing in cognitive radio networks: Optimal sensing strategy, power allocation, and sensing order," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1648–1659, Sep. 2011.

[17] B. Li *et al.*, "Optimal frequency-temporal opportunity exploitation for multichannel ad hoc networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 12, pp. 2289–2302, Dec. 2012.

[18] Y. Wang, Y. He, X. Mao, Y. Liu, and X.-Y. Li, "Exploiting constructive interference for scalable flooding in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1880–1889, Dec. 2013.

[19] Y. Zhou *et al.*, "Throughput optimizing localized link scheduling for multihop wireless networks under physical interference model," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 10, pp. 2708–2720, Oct. 2014.

[20] M. Li, Z. Li, L. Shangguan, S. Tang, and X.-Y. Li, "Understanding multitask schedulability in duty-cycling sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 9, pp. 2464–2475, Sep. 2014.

[21] Z. Cao, Y. He, and Y. Liu, "L^2: Lazy forwarding in low duty cycle wireless sensor networks," in *Proc. IEEE INFOCOM*, 2012, pp. 1323–1331.

[22] P. Xu and M. Li, "Tofu: Semi-truthful online frequency allocation mechanism for wireless networks," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 433–446, Apr. 2011.

[23] P. Xu, S. Wang, and M. Li, "Salsa: Strategyproof online spectrum admissions for wireless networks," *IEEE Trans. Comput.*, vol. 59, no. 12, pp. 1691–1702, Dec. 2010.

[24] Y. Yubo *et al.*, "ZIMO: Building cross-technology mimo to harmonize zigbee smog with wifi flash without intervention," in *Proc. MobiCom*, 2013, pp. 465–476.

[25] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. New York, NY, USA: Springer-Verlag, 2008, pp. 121–151.

[26] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Trans. Mob. Comput.*, vol. 10, no. 2, pp. 239–253, Feb. 2011.

[27] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2/3, pp. 235–256, May 2002.

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

[29] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proc. IEEE INFOCOM*, 2010, pp. 1–9.

[30] 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. 731–745, Apr. 2011.

[31] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *Proc. IEEE INFOCOM*, 2011, pp. 2462–2470.

[32] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Frontiers Dyn. Spectr.*, 2010, pp. 1–9.

[33] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, Apr. 2014.

[34] W. Huang and X. Wang, "Capacity scaling of general cognitive networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1501–1513, Oct. 2012.

[35] M. Dong, G. Sun, X. Wang, and Q. Zhang, "Combinatorial auction with time-frequency flexibility in cognitive radio networks," in *Proc. IEEE INFOCOM*, 2012, pp. 2282–2290.

[36] P. Chaporkar and A. Proutiére, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1546–1555, Oct. 2008.

[37] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[38] T. S. Ferguson, *Optimal Stopping and Applications*. Los Angeles, CA, USA: Univ. of California, 2012.

[39] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

[40] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

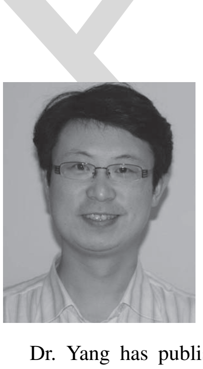
[41] B. Li *et al.*, "Almost optimal dynamically-ordered channel sensing and accessing for cognitive networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2215–2228, Oct. 2014.

[42] B. Li *et al.*, "Almost optimal accessing of nonstochastic channels in cognitive radio networks," *Proc. IEEE INFOCOM*, 2012, pp. 3081–3085.

[43] R. Dhar, G. George, and A. Malani, "Supporting integrated MAC and PHY software development for the USRP SDR," in *Proc. Netw. Technol. Softw. Defined Radio Netw.*, Mar. 2006, pp. 68–77.

[44] Y. Yan, P. Yang, L. You, and B. Li, "Demo abstract: Online optimal channel sensing, probing, accessing in usrp networks," in *Proc. IEEE/ACM ICCPS*, 2012, p. 225.

[45] Y. Liu *et al.*, "Citysee: Not only a wireless sensor network," *IEEE Netw.*, vol. 27, no. 5, pp. 42–47, Sep./Oct. 2013.



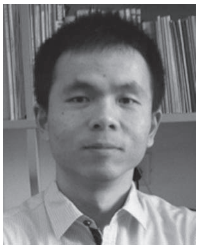
Panlong Yang (M'02) received the B.S., M.S., and Ph.D. degrees in communication and information system from Nanjing Institute of Communication Engineering, Nanjing, China, in 1999, 2002, and 2005 respectively. During September 2010 to September 2011, he was a Visiting Scholar with HKUST. He is now an Associate Professor at the Nanjing Institute of Communication Engineering, PLA University of Science and Technology. His research interests include wireless mesh networks, wireless sensor networks and cognitive radio networks.

Dr. Yang has published more than 50 papers in peer-reviewed journals and refereed conference proceedings in the areas of mobile ad hoc networks, wireless mesh networks and wireless sensor networks. He has also served as a member of program committees for several international conferences. He is a member of the IEEE Computer Society and ACM SIGMOBILE Society.

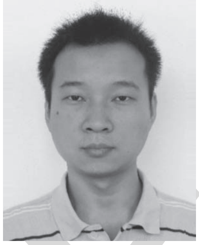
Bowen Li (S'11) received the B.S. degree in wireless communication from the Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, China, in 2007. He is currently working toward the Ph.D. degree from PLA University of Science and Technology. His current research interests include stochastic optimization in cognitive radio networks, and energy efficient algorithm design for wireless sensor networks. He is a student member of the IEEE.

Jinlong Wang received the B.S. degree in mobile communications and the M.S. and Ph.D. degrees in communications engineering and information systems from Institute of Communications Engineering, Nanjing, China, in 1983, 1986, and 1992, respectively. He is a Full Professor of the Institute of Communications Engineering, PLA University of Science and Technology. His current research interests are the broad area of digital communications systems with emphasis on cooperative communication, adaptive modulation, multiple-input-multiple-output systems, soft defined radio, cognitive radio, green wireless communications, and game theory.

Xiang-Yang Li (M'99–SM'08–F'15) received the bachelor's degrees from the Department of Computer Science and the Department of Business Management, Tsinghua University, P.R. China, both in 1995, and the M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign in 2000 and 2001, respectively. He is a Professor at the Illinois Institute of Technology. He is an IEEE Fellow and an ACM Distinguished Scientist. He holds EMC-Endowed Visiting Chair Professorship at Tsinghua University. He is a recipient of China NSF Outstanding Overseas Young Researcher (B). His research interests include wireless networking, mobile computing, security and privacy, cyber physical systems, smart grid, social networking, and algorithms. He and his students won four best paper awards, one best demo award and was nominated for best paper awards twice (ACM MobiCom 2008 and ACM MobiCom 2005). He published a monograph "Wireless Ad Hoc and Sensor Networks: Theory and Applications."



Zhiyong Du (S'12) received the B.S. degree in electronic information engineering from Wuhan University of Technology, Wuhan, China, in 2009. He is currently working toward the Ph.D. degree in communications and information system at the College of Communications Engineering, PLA University of Science and Technology. His research interests include heterogeneous wireless networks, 5G, quality of experience (QoE), learning theory and game theory.



Yubo Yan (S'10) received the B.S. and M.S. degrees in communication and information system from the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2006 and 2011, respectively. He is currently working towards the Ph.D. degree at the PLA University of Science and Technology. His current research interests include software radio systems and wireless sensor networks. He is a student member of the IEEE and the IEEE Computer Society.

Yan Xiong was born in Anhui Province, in 1960. He is a Professor with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include distributed processing, mobile computation, and information security.

Online Sequential Channel Accessing Control: A Double Exploration vs. Exploitation Problem

Panlong Yang, *Member, IEEE*, Bowen Li, *Student Member, IEEE*, Jinlong Wang, Xiang-Yang Li, *Fellow, IEEE*, Zhiyong Du, *Student Member, IEEE*, Yubo Yan, *Student Member, IEEE*, and Yan Xiong

***Abstract*—In opportunistic channel access, the user needs to make real time decisions on when and which channel to access with uncertainty. Assuming perfect channel statistics, several studies have applied optimal stopping theory to derive control strategy for sequential sensing/probing based opportunistically accessing (*s*-SPA), exploiting temporary opportunities among multiple channels. Meanwhile, numerous multi-arm bandit (MAB)-based approaches have been proposed for online learning of channel selection in periodical sensing/accessing system, however, these schemes fail to exploit the opportunistic diversity in short term. In this paper, we investigate online learning of optimal control in *s*-SPA systems, where both statistics learning and temporary opportunity utilization are jointly considered. An effective and efficient online policy, so called IE-OSP, is proposed, which theoretically guarantees system converges to the optimal *s*-SPA strategy with bounded probability. Experimental results further show that, the regret of IE-OSP is almost in optimal logarithmic increasing rate over time, and is sub-linear with the increasing number of channels. Compared with existing solutions, our proposed algorithm achieves 25 ∼ 30% throughput gain in typical scenarios.**

***Index Terms*—Opportunistic spectrum access, sequential sensing and accessing, online learning, diversity exploitation.**

I. Introduction

OPPORTUNISTIC channel access (OSA), due to its flexibility and efficiency in spectrum utilization, has become a well established concept in designing wireless systems [1], [2]. With the success of OSA-based standards such as IEEE 802.11h [3], 802.22 [4], and 802.11af [5], more and more organizations are considering to adopt OSA in future communication standards. In achieving perfect opportunistic channel utilization, the key challenge comes from the unpredictable channel status. Specifically, to acquire the exact channel state, user needs to detect whether the channel is available with spectrum sensing [6], and evaluate the link quality with probing [7]. Online accessing control, i.e., making real time decisions on when and which channel to access, plays a critical role in improving system performance as well as avoiding interference to primary users.

Based on sequential channel sensing and probing, user could opportunistically access a good channel for communication, so as to exploit diversity of temporary channel status among channels. The sequential accessing control problem is firstly studied in multiple i.i.d Rayleigh channels scenario [8], where a multichannel opportunistic auto rate protocol is proposed. Further, more generalized scenarios allowing users to recall pre-probed channels [9], [10] or considering the activities of primary users [11], [12] are further studied. The major concern in these studies is to balance exploration and exploitation on temporary channel status. Corresponding control strategies are constructed on the ideal assumption that the user has perfect knowledge of channel statistics. Since channel statistics are usually unavailable in advance, obtaining complete channel statistics before a communication session will be costly, and would also result in unacceptable delay and overhead.

Our work aims to achieve more throughput gain under the rule of MAB. The reason is, the short-term statistical results could be leveraged for such improvement. We find that, even when no recall action is allowed, the optimal stopping rule could still be applied, where users could opportunistically select the temporary 'good' channel to access, if the user could sense more channels. This motivation relies on two basic facts. First, most of the channels are slow fading, especially for indoor WiFi transmissions. Second, with the advances of wireless communication technology, the channel probing efficiency could be improved in relatively smaller time. Motivated by the aforementioned two conditions, we believe that, the statistical channel knowledge accumulated in the probing process could be leveraged for performance improvements.

To this end, this paper attempts to combine the following two models that have each been quite extensively studied in recent literature: (1) using online learning methods to make sequential channel access decisions when the average channel qualities are unknown a priori (which involves exploration and exploitation); and (2) optimal stopping time methods to determine whether to

Manuscript received June 26, 2014; revised December 4, 2014; accepted April 3, 2015. This research is partially supported by NSF China under Grants No 61232018, 61272487, 61170216, 61172062, 61228202, 61273210, NSF CNS-0832120, NSF CNS-1035894, NSF ECCS-1247944, and NSF CMMI 1436786. The associate editor coordinating the review of this paper and approving it for publication was C. Ghosh.

P. Yang is with the Institute of Communication Engineering, People's Liberation Army University of Science and Technology (PLAUST), Nanjing 210007, China, and also with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China (e-mail: panlongyang@gmail.com).

B. Li is with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China.

J. Wang, Z. Du, and Y. Yan are with the Institute of Communication Engineering, People's Liberation Army University of Science and Technology (PLAUST), Nanjing 210007, China.

X.-Y. Li is with the Tsinghua National Laboratory for Information Science and Technology (TNLIST), Tsinghua University, Beijing 100084, China, and also with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616-3793 USA.

Y. Xiong is with the Department of Computer Science and Technology, University of Science and Technology, Hefei 230026, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2424413

continue sensing the qualities of a given sequence of channels or stop and use the channel for data transmission.

We first analyze the property of optimal sequential sensing, probing and accessing strategy with perfect channel statistics, and then propose an intuitive solution, i.e., myopic learning policy, to help understanding the online accessing control problem. After analyzing the convergence of the myopic learning policy, we find that properly exploring the inaccurately estimated channels is critical for guaranteeing the convergence property. Inspired by this observation, we develop an online policy referred to as IE-OSP, which achieves nearly optimal balance between exploration and exploitation. The main contribution of this paper is two-folds:

First, the brand new double exploration vs. exploitation problem is well studied under the myopic learning policy. We show that, such learning policy with greedy exploitation is non-zero-regret, which indicates that, optimizing opportunity exploitation during a slot is incompatible with that of statistics exploration. Thus, a tradeoff between them is needed for maximizing overall system throughput. Moreover, both the sensing order and accessing rule play critical roles in designing effective and efficient online learning policy.

Secondly, we present a statistical learning based online policy referred to as IE-OSP, which integrates confidence interval estimation into the optimal stopping analytical framework. We've proved that, using the IE-OSP policy, system is guaranteed to converge to the optimal *s*-SPA strategy with bounded probability. Extensive simulation results show that, the expected regret of the IE-OSP policy achieves near optimal logarithmic increasing rate over time, and is sub-linear increasing with the number of channels. Comparing with existing solutions, our proposed scheme achieves 25$\sim$30% throughput gain in most scenarios.

The rest of the paper is organized as follows. The related work is introduced in Section II and in Section III, we briefly present the system model and problem formulation. Further, we analyze the online sequential channel accessing control problem with an intuitive learning policy in Section IV. In Section V, the proposed IE-OSP algorithm and corresponding analysis are presented. Our evaluation results are presented in Section VI. Finally, we conclude our paper in Section VII.

II. Related Work

Opportunistic spectrum accessing control have received much attention recently. Online decisions are made under channel uncertainty, maximizing the system throughput by flexibly exploiting communication opportunities. The most relevant studies to our work can be classified to the following two broad categories:

A. Optimal Control for Sequential Sensing, Probing, and Accessing

To efficiently explore and exploit *diversity on temporary channel status* among multiple channels, optimal control algorithms for sequential channel sensing, probing and accessing scheme have been widely studied. The real time decisions, i.e., whether to access channel or continue to observe another channel immediately, are made on the observed temporary channel status.

Considering i.i.d. Rayleigh fading channels, Sabharwal *et al.* [8] firstly analyze the gains from opportunistic band selection. To obtain such gain, sequential probing based opportunistic channel accessing scheme is proposed, and optimal skipping rule is derived by finite-horizon optimal stopping formulation. More generalized scenarios, e.g., with arbitrary number of channels, statistically non-identical channels, and possibly different probing costs, are studied in seminar work [9], [10], [13]. Moreover, recalling a pre-probed channel as well as accessing an unobserved channel are allowed in their considered communication model.[1] The corresponding optimal strategies are derived by comprehensive theoretic proofs. In [11], Shu and Krunz consider an OSA network with primary users, and thus channel quality as well as availability are considered when making accessing decisions. States of different channels are considered to be i.i.d. to each other, and an infinite-horizon optimal stopping model is leveraged to formulate the online control problem during the *s*-SPA process. For scenarios with non-identical channels, sensing order plays a critical role in achieving maximum throughput. Jiang *et al.* firstly considered the problem of acquiring the optimal sensing/probing order for a single user case in [12]. A computational efficient algorithm is constructed by appealing to dynamic program. Later, Fan *et al.* [14] extends sensing order selection to a two-user case, where a coordinator in the network to determine the sensing orders for each of the two users is required. Recently, Zhao *et al.* [15] propose a novel sensing metric that integrate the channel availability, link quality and access collisions, to guide the sensing order selection. A dynamic programming algorithm is presented, which allows each node to efficiently determine its sensing order in coordination with neighboring nodes. More recently, Pei *et al.* [16] extend the sequential channel sensing and accessing control to a new area, where energy-efficiency is mainly concerned. In their work, sensing order, accessing strategy and transmit power are jointly optimized with dynamic programming. Unlike assuming time-independent channels, i.e., channel states are considered to be independent across slots, Li *et al.* [17] consider Markovian channels and investigate the sequential probing based opportunistic channel accessing and releasing scheme, where a two-dimension optimal stopping framework is proposed for achieving optimal action point under Rayleigh fading. Wang *et al.* [18] exploit constructive interference for scalable flooding. Reference [19]–[21] propose schedule schemes to optimize throughput. Other works [22]–[24] are proposed to exploit the frequency diversity.

The major difference between our work and the above-mentioned studies can be explained as follows. In all the above-mentioned studies, the optimal control strategies are constructed on the assumption of perfect channel statistics. In contrast, we consider more practical scenarios that channel

[1] "Recalling a channel" means revisit the previous probed channel. Such that, the reward could be increased if the user found the previously probed channel is better. Comparing with scheme without recalling, such scheme could achieve lower regret value.

statistics are unknown in the beginning, and focus on investigating online learning method to achieve optimal control of sequential sensing, probing and accessing.

B. Online Learning of Dynamic Channel Selection

Online learning framework for opportunistic spectrum access when channel statistics is unknown a priori, especially formulated as multi-armed bandit (MAB) problems [25], has been fully investigated for periodical sensing/accessing system. The main concern in these studies is to explore and exploit *diversity on channel statistics* among multiple channels efficiently. Specifically, the dynamic selection process is expected to converge to choosing the statistically optimal channel, i.e., the channel with maximum expected reward, thus to achieve diversity gain over channel statistics.

Lai *et al.* [26] firstly apply multi-arm bandit formulations to user-channel selection problems in OSA networks. Especially for the single user case, the UCB1 [27] algorithm is proposed, which is order-optimal with respect to regret. And for decentralized multiple users, a randomized access policy is presented for learning the unknown parameters efficiently. Liu and Zhao [28] formulate the secondary user channel selection to a decentralized multi-armed bandit problem, where contentions among multiple users are considered. A policy achieving asymptotically logarithmic regret is proposed in their work. Anandkumar in [29] and [30] proposed two policies for distributed learning and accessing rule, lead to order-optimal throughput. In addition to learning the channel availability, the secondary users also learn others' strategies, even the total number of users, through channel level feedback. Tekin and Liu [31] modeled each channel as a restless Markov chain rather than time-independent channels as studied before, and multiple channel states rather than binary states are considered. They present a sample-mean based index policy, showing that, under mild conditions, it could achieve logarithmic regret uniformly over time. For the multiuser-multichannel matching problem, Gai *et al.* [32] develop a combinatorial multi-armed bandits (MAB) formulation to address the channel allocation problem under centralized setting. An online learning algorithm that achieves $O(\log T)$ regret uniformly over time is derived. Later, Kalathil *et al.* [33] consider a decentralized setting where there is no dedicated communication channel for coordination among the users. An online index-based distributed learning policy called the dUCB4 algorithm is developed, which achieves the expected regret growing at most as $near - O(\log^2 T)$. Huang *et al.* [34] study the scaling problem of general cognitive radio networks, Dong *et al.* [35] propose a auction scheme.

The main difference between our work and existing online learning frameworks can be explained as follows. All existing studies are focused on periodical sensing/accessing system, where the user only needs to select one channel at a slot. While we consider online learning of optimal control in sequential sensing, probing and accessing systems, where a series of decisions are needed to be made in each slot.

Remark: To the best of our knowledge, it is the first work on integrating OSP and MAB in one unified theoretic framework, making a good balance between statistical exploration across slots and opportunity exploitation during a slot.

III. System Model and Problem Formulation

Considering an OSA network with potential channel set $\Omega = \{1, 2, \ldots, N\}$, each cognitive user could sense/probe/access only one channel at a time, and is operated in *constant access time* (CAT) mode [8], i.e., users could have a constant duration T for channel observation and data transmission, once they would win a communication chance. The communication chances of users come from wining competition with the control channel in distributed wireless system [8], or assigned by a center node as in one hop access system [36]. We denote the duration of each access time as a slot.

The channel state consists of two elements: channel availability and link quality. Denote $a_i(j)$ as the availability of channel i in the j^{th} slot, and availability state $a_i(j) \in \{0, 1\}$, where $a_i(j) = 0$ indicates that the primary user is transmitting over channel i in the j^{th} slot, and $a_i(j) = 1$, otherwise. The channel quality is characterized by the temporary 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). Denote $q_i(j)$ as the quality of channel i in the j^{th} slot. We consider slow-varying Rayleigh fading channels, which is typical for multipath propagation environment [11], [17]. Thus the received temporary SNR is distributed exponentially [12], [37], 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. Both the channel idle probability vector $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ and the SNR mean vector $\Upsilon = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ are unknown to user at the beginning, but can be available through learning. Channel state is considered to be stable during T, as slot duration in OSA system is set to be much shorter than channel coherence time, as well as the sojourn time of primary user activities. Moreover, as the interval time between consecutive communication chances is relatively long in multi-user networks (as discussed in [8]), the channel states in different slots are commonly treated to be independent of each other. This assumption is consistent with previous studies [8]–[12], [26], [28]–[30], [32]. Also, there is another concern that, since the channel states are assumed i.i.d over time, there is no need to assume constant channel quality during T, and allowing the recall process could improve the results. The main reason is to protect primary users' communication. Since there is contention among users, and the primary users could use the licensed channel anytime, we need to set the duration T short enough for this concern. Thus, there is no chance to recall back the previous probed channels.

We depict the online accessing control process in Fig. 1. The s-SPA proceeds slot by slot. For a given slot, says slot j, s-SPA process can be described as follows. Firstly, user senses a channel $\phi_1(j)$ to acquire the channel availability $a_{\phi_1(j)}(j)$. If $a_{\phi_1(j)}(j) = 1$ (i.e., the sensed channel is idle), user further probes the channel via physical layer measurement mechanism (which also has been applied in [17]), acquiring temporary link

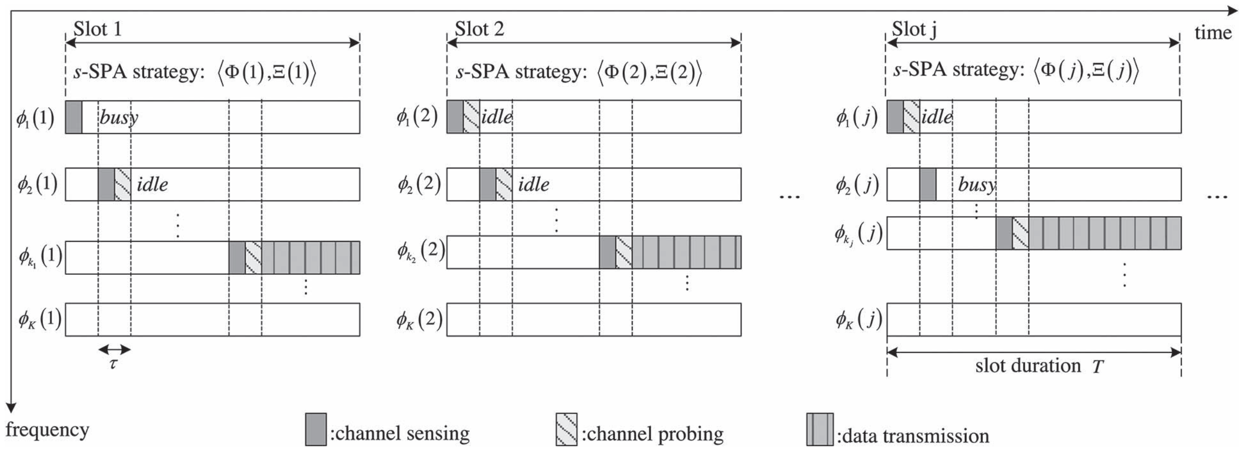


Fig. 1. Online sequential sensing, probing and accessing (*s*-SPA) control.

quality $q_{\phi_1(j)}(j)$. With the observed result, user needs to make a real time decision on whether to access the channel $\phi_1(j)$, or go on *s*-SPA process by switching to another channel, says $\phi_2(j)$. During the *s*-SPA process, if a channel is sensed to be busy, the user is forbidden to send measurement packet for primary user protection. However, the user still needs to wait for a constant channel probing time before switching to next channel. Such scheme is introduced for transceiver synchronization under the case that the channel availability of transmitter and receiver is different [11]. As a result, each sensing/probing step costs a constant time τ. Correspondingly, the maximum number of steps one could take in one slot is $K = \min\left(N, \left\lfloor \frac{T}{\tau} \right\rfloor\right)$, where $\lfloor\cdot\rfloor$ represents round-down function.

When user decides to access channel for data transmission after the k^{th} channel sensing/probing step, the immediate normalized throughput is given by

$$\begin{aligned} r(j) &= c_k \ln\left(1 + q_{\phi_k(j)}(j)\right) \\ &= (1 - k\beta)\ln\left(1 + q_{\phi_k(j)}(j)\right) \end{aligned} \tag{1}$$

where $\beta = \frac{\tau}{T}$ is a normalized observation cost, which is a factor to show the fraction of time a probing duration occupies the whole time slot. As we know, in evaluating the probing time overhead, the normalized β factor is used to evaluate this overhead. In our work, we use $c_k = 1 - k\beta$ to evaluate the pure data transmission time in each slot. The actual throughput can be easily obtained by scaling our reward[2] with a constant $\frac{T}{\ln 2}$.

We define the deterministic learning policy χ, mapping from the observation history $\mathcal{F}_{j-1}$ to a *s*-SPA strategy $\langle\Phi(j), \Xi(j)\rangle$ at each slot j, where $\Phi(j) = (\phi_1(j), \phi_2(j), \ldots, \phi_K(j))$ is a permutation of channels that determines the channel sensing/probing order in a slot, and $\Xi(j)$ is the corresponding accessing rule determining when to access which channel. For notation convenience, we define Ψ as the set of all possible sensing orders, and denote the m^{th} element in it as $\Phi_m = (\phi_1^m, \phi_2^m, \ldots, \phi_K^m)$. Correspondingly, the number of all possible sensing orders $|\Psi| = M = \binom{N}{K}K!$. Then, deriving a *s*-SPA strategy $\langle\Phi, \Xi\rangle$ in a slot includes:

1) selecting K channels from channel set Ω;
2) arranging the order of the selected K channels for sequential channel sensing/probing;
3) deriving an accessing rule for opportunistic channel accessing.

Our main goal is to devise a learning policy guiding the system converging to the throughput-optimal *s*-SPA strategy. Meanwhile, the accumulated throughput loss during the learning process should be as small as possible. We use *regret* value to characterize the accumulated throughput loss, which is defined as the gap between the accumulated reward gained by always using the perfect *s*-SPA strategy, and using the *s*-SPA strategy proposed by learning policy in each slot. Mathematically, the regret of learning policy χ up to slot L is

$$\rho_\chi(L) = LV^*_{\{\Theta,\Upsilon\}} - \sum_{j=1}^{L} \chi\, V^{\langle\Phi(j),\Xi(j)\rangle}_{\{\Theta,\Upsilon\}} \tag{2}$$

Here, $V^*_{\{\Theta,\Upsilon\}}$ is the maximum expected throughput one could obtain in one slot under the environment $\{\Theta, \Upsilon\}$, which is achieved by user applying the ideal *s*-SPA strategy $\langle\Phi^*, \Xi^*\rangle$ derived with perfect statistical knowledge. $V^{\langle\Phi(j),\Xi(j)\rangle}_{\{\Theta,\Upsilon\}}$ is the corresponding reward user obtains with the strategy $\langle\Phi(j), \Xi(j)\rangle$ derived by learning policy χ.

The main notations and definitions of this paper are summarized in Table I.

IV. Understanding Sequential Accessing Control in *s*-SPA

In this section, we are aiming to demonstrate the fundamental tradeoff problem behind the sequential accessing control in *s*-SPA. We first propose a preliminary on the throughput-optimal sequential sensing, probing and accessing strategy with perfect statistics. After that, an intuitive strategy referred to as myopic learning policy is studied, and several observations are derived from the convergence analysis of this learning policy.

[2] The reward is directly related with the throughput. The difference is, when we use the reward for denotation, it mainly focuses on the regret analysis, where the reward value is evaluated with expectation value in the long run. On the other hand, when the term 'throughput' is used, it mainly focuses on the achievable data transmission rate, which is an instant value for evaluation.

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Description
N:	total number of channels
K:	number of maximum observing steps in one slot, where $K = \min\left(N, \frac{T}{\tau}\right)$
M:	total number of possible sensing orders
i:	channel index, $1 \leq i \leq N$
j:	slot index, $1 \leq j \leq L$
k:	step index during a slot, $1 \leq k \leq K$
δ:	a tunable parameter in IE-OSP, where $1 - \delta$ is the confidence coefficient of the estimations
c_k:	normalized remaining time for data transmission after k^{th}: $c_k = 1 - k\beta$
Φ_m:	the m^{th} sensing order in sensing order set Ψ: $\Phi_m = \left(\phi_1^m, \ldots, \phi_K^m\right)$
ϕ_k^m:	ID of the k^{th} channel in sensing order Φ_m
Ξ:	accessing rule, described by a sequence of SNR thresholds, i.e.,$\Xi = (\Gamma_1, \Gamma_2, \ldots, \Gamma_K)$
Λ_1^m:	maximum expected reward user could obtain with sensing order Φ_m
$a_i(j)$:	availability of channel i in epoch j
$q_i(j)$:	quality of channel i in epoch j
$\rho_\chi(L)$:	regret of learning policy χ up to slot L
$\{\Theta, \Upsilon\}$:	channel statistics, where $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, and $\Upsilon = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$
$\left\{\hat{\Theta}, \hat{\Upsilon}\right\}$:	estimated channel statistics, where $\hat{\Theta} = \left\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N\right\}$ and $\hat{\Upsilon} = \{\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_N\}$
$\left\{\hat{\Theta}^u, \hat{\Upsilon}^u\right\}$:	upper confidence bound of estimated channel statistics, where $\hat{\Theta}^u = \left\{\hat{\theta}_1^u, \hat{\theta}_2^u, \ldots, \hat{\theta}_N^u\right\}$ and $\hat{\Upsilon}^u = \left\{\hat{\gamma}_1^u, \hat{\gamma}_2^u, \ldots, \hat{\gamma}_N^u\right\}$
$V_{\{\Theta', \Upsilon'\}}^{\langle\Phi', \Xi'\rangle}$:	expected throughput obtained by strategy $\langle\Phi', \Xi'\rangle$ when statistics is $\{\Theta', \Upsilon'\}$
$\langle\Phi(j), \Xi(j)\rangle$:	s-SPA strategy in the j^{th} slot

A. Optimal s-SPA Strategy Under Perfect Statistics

Given a channel sensing order Φ_m and the channel statistics $\{\Theta, \Upsilon\}$, obtaining the optimal s-SPA strategy can be formulated as an optimal stopping problem (OSP) [38]: during the sequential sensing/probing process, user makes a real time decision on when to stop channel sensing by accessing an observed channel. We formulate the problem as follows.

After sensing/probing channel ϕ_k^m, if the observed channel is idle with channel quality $q_{\phi_k^m}$, the achievable reward in step k is given by:

$$r_k^m = \begin{cases} c_k \ln\left(1 + q_{\phi_k^m}\right), c_k \ln\left(1 + q_{\phi_k^m}\right) > \Lambda_{k+1}^m \\ \Lambda_{k+1}^m, \qquad\qquad\quad \text{else} \end{cases} \tag{3}$$

where $\Lambda_{k+1}^m = E[r_{k+1}^m]$ is the expected reward when user decides to skip the current channel under sensing order Φ_m.

Since in the last step K, the optimal choice is always to access the channel if it is available. Therefore,

$$\Lambda_K^m = E\left[r_K^m\right] = c_K E\left[\theta_{\phi_K^m} \ln\left(1 + q_{\phi_K^m}\right)\right]$$

Then, the expected reward in each step $\Lambda_{K-1}^m, \Lambda_{K-2}^m, \ldots, \Lambda_1^m$ can be obtained using backward deduction according to Eqn. (3).

Specifically, with the channel statistics $\{\Theta, \Upsilon\}$, the expected reward Λ_K^m is given by

$$\begin{aligned} \Lambda_K^m &= c_K \theta_{\phi_K^m} \int_0^\infty \log(1+q) \frac{1}{\gamma_{\phi_K^m}} e^{-\frac{q}{\gamma_{\phi_K^m}}} dq \\ &= c_K \theta_{\phi_N^m} e^{\frac{1}{\gamma_{\phi_K^m}}} \operatorname{Ei}\left(1, \frac{1}{\gamma_{\phi_K^m}}\right) \end{aligned} \tag{4}$$

where function Ei is the exponential integral function defined as $\operatorname{Ei}(1, x) = \int_x^\infty \frac{e^{-t}}{t} dt$ for $x > 0$.

For $1 \leq k < K$, the Λ_k^m can be computed using the following recursion [8], [12], [38].

$$\begin{aligned} \Lambda_k^m = &\left(1 - \theta_{\phi_k^m}\right) \Lambda_{k+1}^m \\ &+ \theta_{\phi_k^m} \Lambda_{k+1}^m \int_0^{c_k \log(1+q) \leq \Lambda_{k+1}^m} \frac{1}{\gamma_{\phi_k^m}} e^{-\frac{q}{\gamma_{\phi_k^m}}} dq \\ &+ c_k \theta_{\phi_k^m} \int_{c_k \log(1+q) > \Lambda_{k+1}^m}^\infty \log(1+q) \frac{1}{\gamma_{\phi_k^m}} e^{-\frac{q}{\gamma_{\phi_k^m}}} dq \\ = &\left(1 - \theta_{\phi_k^m}\right) \Lambda_{k+1}^m + \theta_{\phi_k^m} \Lambda_{k+1}^m \int_0^{e^{\frac{\Lambda_{k+1}^m}{c_k}} - 1} \frac{1}{\gamma_{\phi_N^m}} e^{-\frac{q}{\gamma_{\phi_N^m}}} dq \\ &+ c_k \theta_{\phi_k^m} \int_{e^{\frac{\Lambda_{k+1}^m}{c_k}} - 1}^\infty \log(1+q) \frac{1}{\gamma_{\phi_N^m}} e^{-\frac{q}{\gamma_{\phi_N^m}}} dq \\ = &\Lambda_{k+1}^m + c_k \theta_{\phi_k^m} e^{\frac{1}{\gamma_{\phi_k^m}}} \operatorname{Ei}\left(1, \frac{e^{\frac{\Lambda_{k+1}^m}{c_k}}}{\gamma_{\phi_k^m}}\right) \end{aligned} \tag{5}$$

According to Eqn. (3), the optimal stopping rule, i.e., optimal accessing strategy, is completely specified by the reward sequence $(\Lambda_1^m, \Lambda_2^m, \ldots, \Lambda_K^m)$: access the channel ϕ_k^m after the k^{th} sensing/probing step, if the channel is idle with achievable throughput $c_k \ln(1 + q_{\phi_k^m}) \geq \Lambda_k^m$. Otherwise, user could switch to channel ϕ_{k+1}^m for another sensing/probing step. Obviously, the accessing rule can be further simply described as a sequence of SNR thresholds, denoted as $\Xi_m = (\Gamma_1^m, \Gamma_2^m, \ldots, \Gamma_K^m)$. Hence, the access threshold $\Gamma_k^{m^*}$ is given by

$$\Gamma_k^{m^*} = \begin{cases} e^{\frac{\Lambda_{k+1}^{m^*}}{c_k}} - 1, & 1 \leq k < K \\ 0, & k = K \end{cases} \tag{6}$$

Finally, Λ_1^m is the maximum expected reward user could obtain with sensing order Φ_m. The sensing order Φ_{m^*} generating the maximum $\Lambda_1^{m^*}$ is then the optimal sensing order under the given scenario with channel statistics $\{\Theta, \Upsilon\}$.

B. Complexity Analysis

An intuitive solution when channel statistics is unavailable is that, always deriving s-SPA strategy maximizing immediate throughput in each slot. Meanwhile, refined statistics by updating the estimations of channels have been observed.

During the slot by slot decision-making process, the estimations of channels are obtained by recording and updating the following four variables on each channel: $\hat{\theta}_i(j)$, $n_i^s(j)$, $\hat{\gamma}_i(j)$ and $n_i^p(j)$. Where $\hat{\theta}_i(j)$ is the estimated idle probability of channel i

up to slot j, and $n_i^s(j)$ is the times channel i having been sensed till slot j. They are initialized to be zero and updated as follows:

$$\hat{\theta}_i(j) = \begin{cases} \frac{\hat{\theta}_i(j-1)n_i^s(j-1)+d_i^j}{n_i^s(j-1)+1}, & \text{if channel } i \text{ is sensed} \\ \hat{\theta}_i(j-1), & \text{else} \end{cases} \tag{7}$$

$$n_i^s(j) = \begin{cases} n_i^s(j-1)+1, & \text{if channel } i \text{ is sensed} \\ n_i^s(j-1), & \text{else} \end{cases} \tag{8}$$

Similarly, $\hat{\gamma}_i(j)$ is the estimated SNR mean of channel i up to slot j, and $n_i^p(j)$ is the times channel i having been probed till slot j. They are updated as follows:

$$\hat{\gamma}_i(j) = \begin{cases} \frac{\hat{\gamma}_i(j-1)n_i^p(j-1)+q_i^j}{n_i^p(j-1)+1}, & \text{if channel } i \text{ is probed} \\ \hat{\gamma}_i(j-1), & \text{else} \end{cases} \tag{9}$$

$$n_i^p(j) = \begin{cases} n_i^p(j-1)+1, & \text{if channel } i \text{ is probed} \\ n_i^p(j-1), & \text{else} \end{cases} \tag{10}$$

Since the throughput in each slot is always maximized with the currently estimated statistics, and the channel statistics is refined slot by slot with myopic learning policy, it turns out to be a good solution for our concern.

A learning policy of non-zero-regret is equivalent to the statement that, using the learning policy, system may converge to a non-optimal solution as time goes on.

C. Challenges

However, it is really challenging to achieve optimal control because that, the reward of utilizing and learning in s-SPA process are hard to quantify. Moreover, these two rewards are both related to the sensing order and accessing rule. Specifically,

1) The closed expression of expected throughput is unavailable, which has been shown in Section IV-A. Moreover, for throughput optimal channel access scheme, the channel sensing order relies on the long-term quality, which would not show a direct relationship to the channel probing results. Temporary channel quality is not stable and would possibly contradict to the results in optimal throughput strategy.
2) Considering the exploration process, channels being learnt during a slot are unpredictable. Although intuitively one could improve channel statistics exploration by increasing the accessing thresholds, the exact relationship is complicated, and can only be described in a probabilistic way.

As a result, to achieve optimal s-SPA strategy as well as reduce the throughput loss during the learning process, one needs to consider exploring the inaccurately estimated channels while pursuing immediate reward maximization, by jointly optimizing the sensing order selection process across slots and the opportunistic accessing control process in each slot.

V. IE-OSP Algorithm

In this section, we propose the IE-OSP (i.e., Interval Estimation in OSP analytical framework) online policy, in which the statistics learning and diversity utilization processes are seamlessly integrated together for efficient spectrum access. We further analyze the convergence of the proposed policy, and prove that the IE-OSP is guaranteed to converge to the optimal s-SPA strategy with a controlled probability.

A. Algorithm Description

In our algorithm, the basic idea for guiding our system being converged to the optimal s-SPA strategy is to minimize the unreachable probability of inaccurate channels during the s-SPA process. Meanwhile, the optimal stopping analytical framework is used during the s-SPA process for obtaining diversity gain during the learning process.

For each channel, the following four variables are recorded and updated during s-SPA process for decision-making, i.e., the estimated channel idle probability $\hat{\theta}$, the times channel having been sensed n^s, the estimated channel SNR mean $\hat{\gamma}$ and the times channel having been probed n^p. They are updated according to (7)–(10), respectively.

We leverage the confidence interval bound to characterize the inaccuracy of statistical estimation. Define parameter $0 < \delta < 1$, where $1 - \delta$ is the confidence coefficient of the estimations. Then, the $1 - \delta$ upper confidence bound of the channel idle probability and the channel SNR mean are respectively given by

$$\hat{\theta}_i^u(j) = \min\left\{1, \hat{\theta}_i(j) + \sqrt{\frac{-\log\delta}{2n_i^s(j)}}\right\} \tag{11}$$

$$\hat{\gamma}_i^u(j) = \min\left\{q_{max}, \hat{\gamma}_i(j) + q_{max}\sqrt{\frac{-\log\delta}{2n_i^p(j)}}\right\} \tag{12}$$

where q_{max} denotes the maximum value of temporary received SNR. It is reasonable to restrict q with an upper bound q_{max}, since the probability that temporary SNR is larger than q_{max} approximates to zero if the value of q_{max} is large enough.

Then, the IE-OSP can be described as follows. Firstly, sequentially sense/probe channels until all channels are probed at least once (from line 2 to line 13). Note that, the pseudo code from line 5 to line 8 operates for the case where channel is available, and the channle is probed with property channel quality updating operations. If the channel is busy, we should move forward for next channel. Line 8 and line 10 in the pseduo are using the same operations to visit next available channels. After that, always choose the s-SPA strategy $\langle\Phi_{m^*}(j), \Xi_{m^*}^u(j)\rangle$ that achieves $\max_m \Lambda_1^{m,u}(j)$ in slot j, where $\Lambda_1^{m,u}(j)$ is a virtual throughput value defined as the maximum achievable throughput one could achieve if the real statistics is $\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\}$ (from line 14 to line 21). Obviously, $\langle\Phi_{m^*}(j), \Xi_{m^*}^u(j)\rangle$ can be derived easily with $\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\}$, using the optimal stopping analytical framework we introduced in Section IV-A.

The pseudo-code of the IE-OSP algorithm is shown as in Fig. 2.

B. Convergence Analysis

In this subsection, we analyze the convergence of IE-OSP algorithm, because the optimal convergence point is critical to online learning policy in the long run. The main result

```
IE-OSP Policy
Initialize: for all $1 \leq i \leq N$: $\hat{\theta}_i = 0$, $n_i^s = 0$, $\hat{\gamma}_i = 0$, $n_i^p = 0$, $S_0 = \Omega$, $l = 1$, $k = 1$;
while $S_0 \neq \emptyset$ do
    Sense a random channel $i \in S_0$;
    $k = k + 1$, update $\hat{\theta}_i(l)$ and $n_i^s(l)$ according to Eqn.(7) and (8), respectively;
    if $a_i(l) == 1$ then
        Probe and then access channel $i$;
        Update $\hat{\gamma}_i(l)$ and $n_i^p(l)$ according to Eqn.(9) and (10), respectively;
        $l = l + 1$, $k = 1$, $S_0 = S_0 \setminus \{i\}$;
    else if $k == K + 1$ then
        $l = l + 1$, $k = 1$, $S_0 = S_0 \setminus \{i\}$;
        Wait for next communication slot;
    end if
end while
for $j = l : L$ do
    for $m = 1 : M$ do
        Select sensing order $\Phi_m$
        for $k = K : 1$ do
            Compute $\hat{\Lambda}_k^{m,u}(j)$ with $\left\{\hat{\Theta}^u(j), \hat{\Upsilon}^u(j)\right\}$ according to Eqn. (4) or (5);
            Compute $\Gamma_k^{m,u}(j)$ according to Eqn. (6);
        end for
    end for
    Determine $m^*(j) = \arg\max_{1 \leq m \leq M} \left\{\hat{\Lambda}_1^{m,u}(j)\right\}$;
    Proceed $s$-SPA with strategy $\langle \Phi_{m^*}(j), \Xi_{m^*}^u(j) \rangle$;
    Update $\hat{\theta}_i(j)$, $n_i^s(j)$, $\hat{\gamma}_i(j)$ and $n_i^p(j)$, according to Eqn.(7), (8), (9) and (10), respectively;
end for
```

Fig. 2. Algorithm description on IE-OSP.

can be described by the following theorem, which provides a theoretical convergence guarantee for our proposed policy.

Theorem 1: Using *IE-OSP*, system converges to the throughput-optimal s-SPA strategy with probability at least $(1-\delta)^{2(N-1)}$. Particularly, when $\forall i : \theta_i < 1$, it converges to optimal s-SPA strategy with probability at least $(1-\delta)^{2(N-K)}$, where $1-\delta$ is used to provide bounds to the statistical channel features in channel idle probability and SNR mean, which have been formally defined in Eqn. (11), and Eqn. (12).

Before proving this theorem, it is worth noting that, the performance analysis, e.g., the regret analysis, is typically identical to previous studies [25], [33]. The difference is, since the strategy is mixed with partially known knowledge, and channel dynamics are fully used, there is no fixed optimal policy. The only concern in this work, is to know the probability that the algorithm could converge to the optimal point. To this end, the probability analysis is also challenging in our concern. Thus, an analytical bound is presented to instead of accurate p.d.f. based analysis.

Proof: To prove Theorem 1, we introduce the Chernoff-Hoeffding bound inequalities first.

Lemma 1: (Chernoff-Hoeffding bound) [39] Let $X_1, \ldots, X_n$ be random variables with range [0, 1], such that $E[X_t | X_1, \ldots, X_{t-1}] = \mu$. Moreover, let $S_n = X_1 + \ldots + X_n$. Then, for any $a > 0$,

$$\Pr[S_n \geq n\mu + a] \leq e^{-\frac{2a^2}{n}}$$

and

$$\Pr[S_n \leq n\mu - a] \leq e^{-\frac{2a^2}{n}}$$

According to Lemma 1, we can derive the following corollary directly.

Corollary 1: Let $\mathcal{D}$ be a distribution with support in [0, 1], and $E_{X\sim\mathcal{D}}[X] = \theta$. Let $X_1, \ldots, X_n$ be drawn independently from $\mathcal{D}$, and $\hat{\theta} = \frac{1}{n}\sum_t X_t$. Then

$$\Pr\left[\theta \leq \hat{\theta} + \sqrt{\frac{-\log\delta}{2n}}\right] \geq 1 - \delta$$

and

$$\Pr\left[\theta \geq \hat{\theta} - \sqrt{\frac{-\log\delta}{2n}}\right] \geq 1 - \delta$$

Moreover, let $\mathcal{D}$ denote a distribution with support in [0, q_{max}], and $E_{X\sim\mathcal{D}}[X] = \gamma$. Let $X_1, \ldots, X_n$ be drawn independently from $\mathcal{D}$, and $\hat{\gamma} = \frac{1}{n}\sum_t X_t$. Then

$$\Pr\left[\gamma \leq \hat{\gamma} + q_{max}\sqrt{\frac{-\log\delta}{2n}}\right] \geq 1 - \delta$$

and

$$\Pr\left[\gamma \geq \hat{\gamma} - q_{max}\sqrt{\frac{-\log\delta}{2n}}\right] \geq 1 - \delta$$

Proof: Corollary 1 is directly derived from Lemma 1. ∎

Let θ_i' and γ_i' be the supposed channel statistics of idle probability and the averaged SNR value on channel i respectively, and let θ_i and γ_i be the real corresponding channel statistics. Denote $\langle \Phi', \Xi' \rangle$ (a pair of sensing order and accessing rule) as the throughput-optimal strategy for sequential channel sensing, probing and accessing (s-SPA) in the case that the channel statistics is $\{\Theta', \Upsilon'\}$, i.e., $\{\theta_1', \ldots, \theta_N'; \gamma_1', \ldots, \gamma_N'\}$. We have

Lemma 2: Under any given strategy $\langle \Phi', \Xi' \rangle$, if there exists an overestimated channel, it could be observed with high probability.[3]

Proof: We prove this lemma by contradiction.

Denote $V_{statistic}^{solution}$ as the expected throughput obtained by user using *solution* for sequential channel sensing and accessing, while the actual channel statistics is *statistic*. Thus:

- $V_{\{\Theta', \Upsilon'\}}^{\langle \Phi', \Xi' \rangle}$ is the maximum throughput one could obtain in the supposed scenario $\{\Theta', \Upsilon'\}$;
- $V_{\{\Theta, \Upsilon\}}^{\langle \Phi, \Xi \rangle}$ is the maximum actually achievable throughput in the scenario $\{\Theta, \Upsilon\}$;
- $V_{\{\Theta, \Upsilon\}}^{\langle \Phi', \Xi' \rangle}$ is the expected throughput one could obtain when using $\langle \Phi', \Xi' \rangle$ in the scenario $\{\Theta, \Upsilon\}$.

[3] "With high probability" means that, you can change the conditions slightly to make the probability of failure very small. The usefulness of this concept is from the power of the statement. The statement is parameterized to allow the probability to vary as necessary to prove other statements.

Suppose that for all i except i^*: $\theta_i' = \theta_1$, $\gamma_i' = \gamma_i$, while i^* is the overestimated channel, i.e., it falls into one of the following three conditions: 1) $\theta_{i^*}' > \theta_{i^*}, \gamma_{i^*}' = \gamma_{i^*}$; 2) $\theta_{i^*}' = \theta_{i^*}, \gamma_{i^*}' > \gamma_{i^*}$; and 3) or $\theta_{i^*}' > \theta_{i^*}, \gamma_{i^*}' > \gamma_{i^*}$. Then, we have

$$V_{\{\Theta',\Upsilon'\}}^{\langle\Phi',\Xi'\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi,\Xi\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi',\Xi'\rangle} \tag{13}$$

The statement that channel i^* would never be observed under the strategy $\langle\Phi', \Xi'\rangle$ is equivalent to that, the s-SPA process would stop before arriving channel i^*. If so, we have

$$V_{\{\Theta,\Upsilon\}}^{\langle\Phi',\Xi'\rangle} = V_{\{\Theta',\Upsilon'\}}^{\langle\Phi',\Xi'\rangle} > V_{\{\Theta,\Upsilon\}}^{\langle\Phi,\Xi\rangle}$$

which contradicts the inequality (13). Hence, we can conclude that the statement is false. In other words, the overestimated channel would be observed with probability 1 as time goes on.■

We now prove Theorem 1 using Corollary 1 and Lemma 2.

Since sub-optimal convergence only happens when there exists at least one inaccurately estimated channel, where the statistics of this channel would never be updated again. Suppose that user converges to a state, i.e., a s-SPA solution, where the maximum number of achievable steps in each slot is k. Then, according to Lemma 2, the state is sub-optimal if and only if there exists some underestimated channel in remaining $N - k$ channels.

For the sake of convenient description, we denote the set of remaining channels as $S_r = \{k+1, k+2, \dots, N\}$. For each $i \in S_r$, $p_i = \Pr[\theta_i' \le \theta_i$ or $\gamma_i' \le \gamma_i$. As in IE-OSP, we treat $\theta_i' = \theta_i^u = \hat{\theta}_i + \sqrt{-\frac{\log\delta}{2n_i^s}}$ and $\gamma_i' = \gamma_i^u = \hat{\gamma}_i + q_{max}\sqrt{-\frac{\log\delta}{2n_i^p}}$), according to Corollary 1, we have that $\Pr[\theta_i' \le \theta_i] \le \delta$, $\Pr[\gamma_i' \le \gamma_i] \le \delta$. Thus, for all i, $p_i \le p = 1 - (1-\delta)^2$. Then, the probability $P_{sub-opt}$ that system converges to a sub-optimal solution is bounded by

$$\begin{aligned} P_{sub-opt} &\le C_{N-k}^1 p\,(1-p)^{N-k-1} + C_{N-k}^2 p^2\,(1-p)^{N-k-2} \\ &\quad + \cdots + C_{N-k}^{N-k-1} p^{N-k-1}\,(1-p) + p^{N-k} \\ &= \left[p + (1-p)\right]^{N-k} - (1-p)^{N-k} \\ &= 1 - (1-\delta)^{2(N-k)} \end{aligned} \tag{14}$$

Consequently, the probability that system could converges to optimal solution is bounded by

$$P_{opt} \ge (1-\delta)^{2(N-k)} \tag{15}$$

As user needs to sense and probe at least one channel in each slot, thus $k \ge 1$, then we can derive the following probability of optimal convergence.

$$P_{opt} \ge (1-\delta)^{2(N-1)} \tag{16}$$

Particularly, when all the channel idle probabilities are less than 1, which means that when system converges to a state, all the K channels in the sensing order will be observed as time goes on (since the probability of all channel are busy is bigger than zero). In such case, we have the following statement.

$$P_{opt} \ge t(1-\delta)^{2(N-K)} \tag{17}$$

This completes the proof of Theorem 1. ■

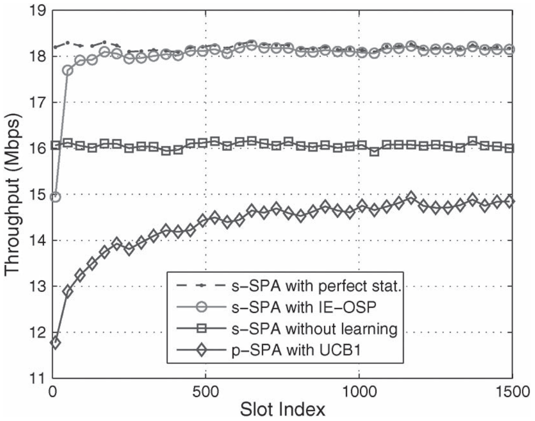


Fig. 3. Comparison on expected throughput with respect to time.

VI. Performance Evaluations

In this section, we evaluate and analyze the performance of the proposed online sequential accessing algorithm via simulations. We run our simulation code with Matlab, and an IBM X210 laptop. Our experiment settings are as follows. The idle probabilities and SNR means of independent channels are randomly generated respectively in range [0, 1] and [0, 15] dB for each round. Then, the states of channels (i.e. availability and link quality) in each slot are generated independently according to the idle probability vector as well as SNR mean vector. The channel bandwidth is set to be 6 MHz, and three channels are considered here. The normalized channel sensing/probing cost $\beta = 0.1$. The results are averaged from 1000 rounds of independent experiments, where each run lasts at least 1500 time slots.

A. Throughput Analysis

In this subsection, four policies are running under the same environment for performance comparison, briefly described as follows.

- p-*SPA with UCB1*: existing online learning solution for opportunistic channel access, in which user selects one channel to sense/access in each slot according to UCB1 [27] algorithm. Such learning policy is proved to be order-optimal in p-SPA system [26];
- s-*SPA without learning*: an intuitive method in s-SPA system without learning. User sequentially senses/probes with a random sensing order and access the first idle channel for transmission;
- s-*SPA with IE-OSP*: our proposed method, where user sequentially senses, probes and accesses according to online algorithm IE-OSP;
- s-*SPA with perfect stat.*: an ideal s-SPA strategy derived with perfect channel statistics, which leads to maximum achievable throughput.

We first study the system throughput as a function of time in Fig. 3. As depicted in Fig. 3,

1) both learning algorithms are effective in improving system throughput. This is clearly shown in the figure, where the

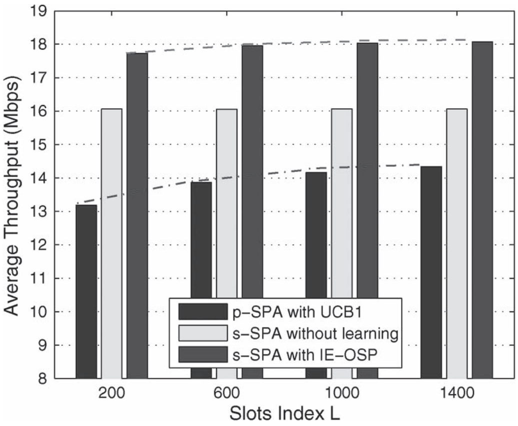


Fig. 4. Comparison on accumulated reward in the first L slots.

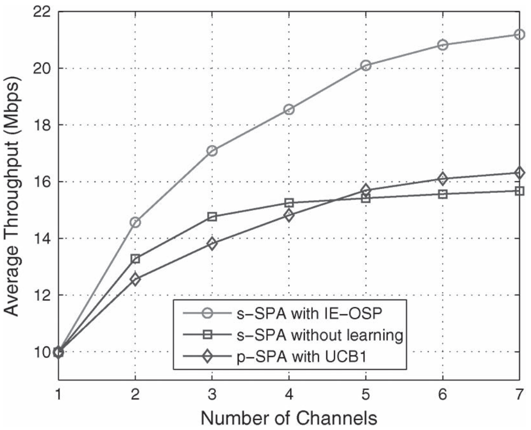


Fig. 5. Comparison on accumulated reward with respect to number of channels.

expected throughput of both p-*SPA with UCB1* and s-*SPA with IE-OSP* are increasing with time.

2) there is still a considerable gap compared with the maximum achievable throughput (i.e., the achievable throughput obtained by s-*SPA with perfect stat.*) by using existing solutions. On one hand, compare the throughput of existing learning method p-*SPA with UCB1* with that of s-*SPA with perfect stat*. It shows about 3 Mbps throughput loss even at the time $t = 1500$, where the learning algorithm converges almost to the optima status. Such a gap mainly arises from the fact that existing learning method is incompatible with temporary opportunity exploitation. On the other hand, the intuitive algorithm for exploiting diversity, i.e., s-*SPA without learning*, shows a constant gap of about 2 Mbps, comparing with the ideal strategy.
3) our proposed algorithm IE-OSP bridges the throughput gap effectively. As shown in figure, the obtained throughput of IE-OSP algorithm approaches to the ideal goal in about 500 slot.

We further investigate the accumulated reward of the three algorithms. Accumulated award in the first L slots is defied as the total transmitted bits from the beginning time, i.e., $j = 1$, to the instant $j = L$. Actually, the accumulated reward is the most concerned metric from the perspective of the user. The results are shown in Fig. 4. Here, we leverage the average throughput in the first L slots to characterize the real value of accumulated reward, which is mathematically defined as $\frac{1}{L}\sum_{j=1}^{L} r(j)$. In the figure, the average throughputs of the three practical schemes with different Ls are given. It clearly shows that, our proposed method outperforms the other two schemes in almost any time, with respect to the accumulated reward. The advantage of our proposed algorithm in time from 200 to 1400 are apparently shown in the figure. More precisely, our learning method outperforms s-*SPA* without learning as soon as $j = 50$, and outperforms p-*SPA with UCB1* in arbitrary time. In other words, applying our proposed scheme earn profits, even in where the communication session duration is relatively short. Moreover, as the gap between the average throughputs of the three schemes are tending towards stability, it is no doubt that user would gain more by applying our proposed scheme as the session duration increases.

All the above results are derived from the scenario with a constant number of channels ($N = 3$). As the number of channels is almost the most important attribute of a wireless network and relates much to the system performance, we evaluate the three schemes in scenarios with different channels in the following part of this subsection, so as to investigate the impact of channel number. We adopt the accumulated reward in the first 1500 slots as the main metric to show the impact of channel number. Similarly, we leverage ‘average throughput’ to characterize the real value of accumulated reward. With the number of channels ranging from 1 to 7, we depict the results as shown in Fig. 5. All the three curves are increasing with the number of channels; however, with different rising characteristics:

1) s-SPA without learning scheme, it shows to be a rapid growth within $N \leq 3$ (higher increasing rate compared with p-*SPA with UCB1* scheme). Such growth in throughput comes from the fact that, as the number of channels increases, it is more likely to find an available channel to use by sequentially observing channels in a slot. In other words, the increasing channels enrich diversity in temporary channel status, and thus benefit the scheme with opportunity exploitation. However, due to lack of advanced accessing control strategy, the s-*SPA without learning* scheme would fail to exploit temporary opportunity efficiently. This is why the increasing trend flattens soon when $N > 4$.
2) for the p-SPA with UCB1 scheme, the growth comes from the increasing diversity of channels’ statistics. Specifically, as the expected reward of the single statistic-optimal channel is increasing with the total number of the channels, user gains more as the number of channels increases, since it could learn to converge to the optimal channel by using p-*SPA with UCB1*. Moreover, the average throughput of p-*SPA with UCB1* increases more slowly than that of s-*SPA without learning* within few channels, e.g., 14 with sustained growth.
3) our proposed s-*SPA with IE-OSP* scheme increases with the number of channels more rapidly and lasting. By using s-*SPA with IE-OSP*, user sequentially senses/probes and accesses with near-optimal strategy soon by learning.

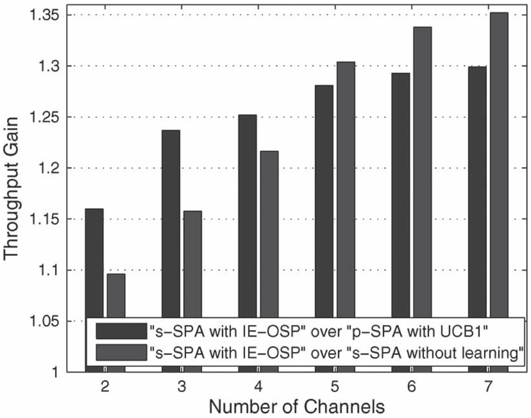


Fig. 6. Throughput gain of *s*-SPA with IE-OSP over the other two schemes.

The temporary opportunity among channels are fully and efficiently exploited. As a result, the throughput gap between our proposed policy and the existing policies is increasing with number of channels, e.g., about 5 Mbps throughput improvement is attained at $N = 7$.

To further investigate the throughput improvement of our proposed scheme over the other two schemes, we depict the throughput gain as a function of the number of channels. The throughput gain is defined as the ratio between average throughput in the first 1500 slots of s-*SPA with IE-OSP* scheme over that of p-*SPA with UCB1* or s-*SPA without learning*, respectively. As depicted in Fig. 6, with the increasing number of channels, the candidate channels are more than ever, thus the potential channel quality improvement is expected, since the probability of probing a high quality channel could be larger than ever. Specifically, we learn from this figure that:

1) the throughput gain of our opposed scheme over the other two schemes are increasing with the number of channels, which means that the proposed policy would benefit more in the scenarios with more channels.
2) at least 9.5% improvement in average throughput is achieved with our proposed scheme. This value is attained at $N = 2$ comparing with s-*SPA without learning*. When compared with p-*SPA with UCB1*, it exceeds 15%.
3) 25∼30% throughput improvement can be obtained in most scenarios, as almost all existing OSA networks are equipped with more than 5 channels.

B. Convergence Analysis

In this subsection, we evaluate the convergence property of our proposed learning algorithm by analyzing regret. Regret is an important metric for online policies, where the definition[4] of regret is presented in Eqn. (2). An online learning algorithm with higher regret means more throughput loss during learning process. Moreover, it has been proven by Lai and Robbins [40] that no policy can do better than logarithmic increasing regret in time. In other words, an online policy with logarithmic regret in time is order-optimal.

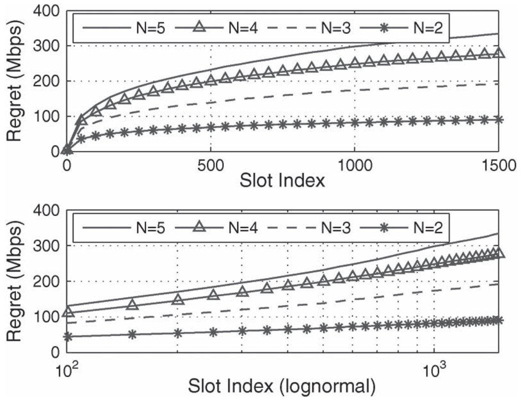


Fig. 7. Regret with respect to time.

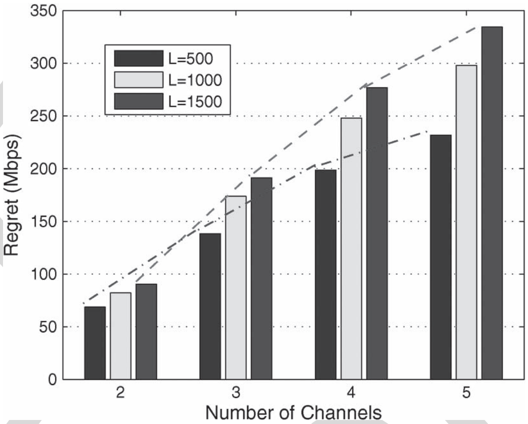
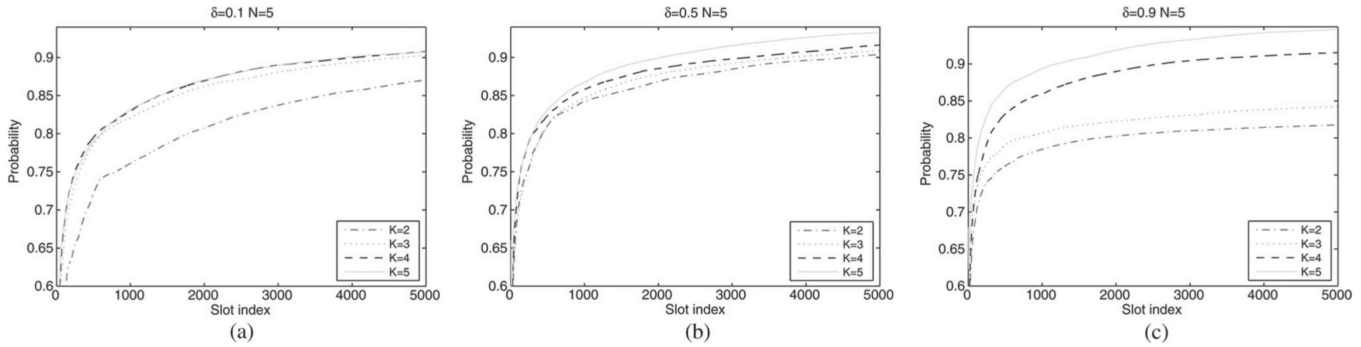


Fig. 8. Regret vs. increased number of channels.

In Fig. 7, we depict the regret of IE-OSP policy as a function of slot index, so as to study the increasing rate of regret over time. To show more widely, we present all the curves with N ranging from 2 to 5. Intuitively, we find from the upper part of this figure that, all the curves of regret show a logarithmic increasing trend over time. To further verify this logarithmic increasing property, we re-plot the regret curves in the lower part of this figure, where X-axis ranges from 100 to 1500 and is in a logarithmic form. The transformed curves show almost linear increasing trend. This verifies that, the regret is in at least asymptotically logarithmic rate, even if it is not in optimal logarithmic rate

Further, we study the increasing trend of regret with respect to the number of channels. As the regret increases infinitely with the number of slots, we take three typical value of L to determine the regret for comparison. Specifically, for each N, we depict the value of $L = 500$, $L = 1000$, and $L = 1500$. The results are presented in Fig. 8. It is intuitive that the regret values increases when adds the number of channels. This is reasonable, since the increasing number of channels extends the learning space, and thus results in higher throughput loss for learning. In spite of this, it is encouraging that the regret is sub-linearly increasing with the number of channels. As shown in the regret envelope curves, where the blue dots and red dashed line sketches the increasing trace of $\rho(500)$ and $\rho(1500)$

[4] As in our simulation, regret is the accumulated throughput loss of applying s-*SPA with IE-OSP*, comparing with always using s-*SPA with perfect stat*.



(a) (b) (c)

Fig. 9. Comparison between simulation and theoretical results. (a) $\delta = 0.1$ and $N = 5$; (b) $\delta = 0.5$ and $N = 5$; (c) $\delta = 0.9$ and $N = 5$.

respectively. Such desirable property makes the learning algorithm scalable.

C. Discussion

1) Impact of Secondary User and Reliability: The channel probing failure and primary user occupancy will lead to different results. In previous studies [41], [42], we discussed the probability of channel probing failure and effects for the statistical behavior of the primary users. Moreover, it is worth noting that, in our scheme, when the channel probing failure and primary user occupancy is stable, say, providing a probability or distribution for it, our IE-OSP policy could be adaptive to such cases. Because the threshold value could be adjustable according to this probabilistic distribution, which could be further evaluated by the rewards.

2) Validating the Theoretical Analysis: To show the matching effects of the proposed algorithm and theorem 1, we make an extended experimental study on the comparisons between the results we got from simulation study and theoretical analysis. In our simulation study, we evaluate the matching rate of the proposed algorithm and theoretical results. For each run, if the result in simulation study equals to that of theoretical analysis, the matching times could be increased by 1. And the overall matching rate is the accumulated matching times to the total number of running times.

As depicted in Fig. 9, the Y-axis denotes the matching rate with probabilistic form. We set the parameter N, K, and δ with different values, and evaluate the matching rate. To show the trends, especially when the number of probing times increases, we make observations for different values of K. This feature also validates our basic idea, i.e., providing more opportunities of probing could improve the throughput gain in temporarily high SNR channels. Large-scale evaluation needs computational intensive operations, and the theoretical results could guide us with the converging trends for the regret value. Furthermore, Fig. 10 depicts the convergenc feature of our proposed protocol, when the theoretical regret value is concerned. In that, we observe the convergence property when the parameter δ is concerned. When the confidence interval is involved, the convergence probability increases with the δ, which means, the convergence probability could be higher than the case with lower confidence interval. On the other hand, a theoretical bound value with higher confidence interval could be more difficult to achieve.

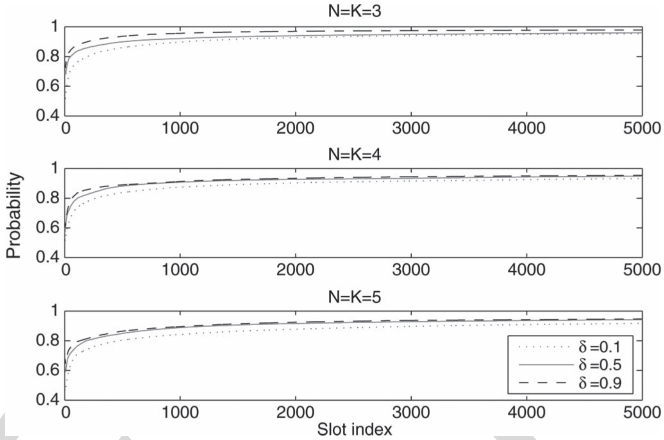


Fig. 10. Convergence property of the simulation results.

VII. Conclusion

In this work, channel learning and opportunity utilization are jointly considered for maximizing system overall throughput in an unknown environment. The sensing/probing order and accessing rule are dynamically adapted slot by slot, so as to achieve better tradeoff between maximizing diversity exploitation in current slot and exploring more channels for refining statistics. A near optimal online learning policy, so called IE-OSP, is proposed, which balances the statistics exploration and diversity exploitation by integrating confidence interval estimation into the optimal stopping analytical framework. We prove that, by using the proposed algorithm, system is guaranteed to converge to the optimal s-SPA strategy with a controllable probability. Simulation results further show that the regret of IE-OSP is asymptotically logarithmic in time and sub-linear in the number of channels, which respectively shows the optimality and scalability of our proposed learning policy. Compared with existing solutions, our proposed algorithm achieves more than 25% throughput gain in most scenarios.

In future work, we are to implement our policy to a cognitive radio platform built on USRP [43], [44], and provide a working system in real deployment [45] for validation.

References

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw. J.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006.

[2] I. F. Akyildiz, W. yeol Lee, and K. R. Chowdhury, "CRAHNs: Cognitive radio ad hoc networks," *Ad Hoc Netw.*, vol. 7, no. 5, pp. 810–836, Jul. 2009.

[3] J. Jeung, S. Jeong, and J. Lim, "Outband sensing-based dynamic frequency selection (DFS) algorithm without full DFS test in IEEE 802.11h protocol," *IEICE Trans.*, vol. 95-B, no. 4, pp. 1295–1296, Apr. 2012.

[4] "IEEE 802.22-2011(TM) standard for cognitive wireless regional area networks (RAN) for operation in tv bands." [Online]. Available: http://www.ieee802.org/22/

[5] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M. Welsh, "Whitespace networking with Wi-Fi like connectivity," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 27–38, Aug. 2009.

[6] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor, "Spectrum sensing for cognitive radio: State-of-the-art and recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 3, pp. 101–116, May 2012.

[7] K. Balach, S. R. Kadaba, and S. Nanda, "Channel quality estimation and rateadaptation for cellular mobile radio," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 7, pp. 1244–1256, Jul. 1999.

[8] A. Sabharwal, A. Khoshnevis, and E. Knightly, "Opportunistic spectral usage: Bounds and a multi-band CSMA/CA protocol," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 533–545, Jun. 2007.

[9] S. Guha, K. Munagala, and S. Sarkar, "Information acquisition and exploitation in multichannel wireless systems," *arXiv preprint arXiv: 0804.1724*, 2008.

[10] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM Trans. Netw.*, vol. 17, no. 6, pp. 1805–1818, Dec. 2009.

[11] T. Shu and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," in *Proc. MobiCom*, 2009, pp. 37–48.

[12] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 297–307, Jan. 2009.

[13] Y. Zhou *et al.*, "Almost optimal channel access in multi-hop networks with unknown channel variables," in *Proc. IEEE ICDCS*, 2014, pp. 461–470.

[14] R. Fan and H. Jiang, "Channel sensing-order setting in cognitive radio networks: A two-user case," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 4997–5008, Nov. 2009.

[15] J. Zhao and X. Wang, "Channel sensing order in multi-user cognitive radio networks," in *Proc. IEEE DYSPAN*, 2012, pp. 397–407.

[16] Y. Pei, Y.-C. Liang, K. C. Teh, and K. H. Li, "Energy-efficient design of sequential channel sensing in cognitive radio networks: Optimal sensing strategy, power allocation, and sensing order," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1648–1659, Sep. 2011.

[17] B. Li *et al.*, "Optimal frequency-temporal opportunity exploitation for multichannel ad hoc networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 12, pp. 2289–2302, Dec. 2012.

[18] Y. Wang, Y. He, X. Mao, Y. Liu, and X.-Y. Li, "Exploiting constructive interference for scalable flooding in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1880–1889, Dec. 2013.

[19] Y. Zhou *et al.*, "Throughput optimizing localized link scheduling for multihop wireless networks under physical interference model," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 10, pp. 2708–2720, Oct. 2014.

[20] M. Li, Z. Li, L. Shangguan, S. Tang, and X.-Y. Li, "Understanding multitask schedulability in duty-cycling sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 9, pp. 2464–2475, Sep. 2014.

[21] Z. Cao, Y. He, and Y. Liu, "L^2: Lazy forwarding in low duty cycle wireless sensor networks," in *Proc. IEEE INFOCOM*, 2012, pp. 1323–1331.

[22] P. Xu and M. Li, "Tofu: Semi-truthful online frequency allocation mechanism for wireless networks," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 433–446, Apr. 2011.

[23] P. Xu, S. Wang, and M. Li, "Salsa: Strategyproof online spectrum admissions for wireless networks," *IEEE Trans. Comput.*, vol. 59, no. 12, pp. 1691–1702, Dec. 2010.

[24] Y. Yubo *et al.*, "ZIMO: Building cross-technology mimo to harmonize zigbee smog with wifi flash without intervention," in *Proc. MobiCom*, 2013, pp. 465–476.

[25] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. New York, NY, USA: Springer-Verlag, 2008, pp. 121–151.

[26] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Trans. Mob. Comput.*, vol. 10, no. 2, pp. 239–253, Feb. 2011.

[27] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2/3, pp. 235–256, May 2002.

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

[29] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proc. IEEE INFOCOM*, 2010, pp. 1–9.

[30] 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. 731–745, Apr. 2011.

[31] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *Proc. IEEE INFOCOM*, 2011, pp. 2462–2470.

[32] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Frontiers Dyn. Spectr.*, 2010, pp. 1–9.

[33] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, Apr. 2014.

[34] W. Huang and X. Wang, "Capacity scaling of general cognitive networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1501–1513, Oct. 2012.

[35] M. Dong, G. Sun, X. Wang, and Q. Zhang, "Combinatorial auction with time-frequency flexibility in cognitive radio networks," in *Proc. IEEE INFOCOM*, 2012, pp. 2282–2290.

[36] P. Chaporkar and A. Proutiére, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1546–1555, Oct. 2008.

[37] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[38] T. S. Ferguson, *Optimal Stopping and Applications*. Los Angeles, CA, USA: Univ. of California, 2012.

[39] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

[40] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

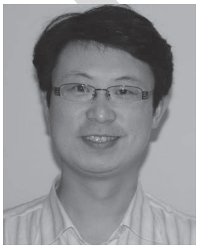
[41] B. Li *et al.*, "Almost optimal dynamically-ordered channel sensing and accessing for cognitive networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2215–2228, Oct. 2014.

[42] B. Li *et al.*, "Almost optimal accessing of nonstochastic channels in cognitive radio networks," *Proc. IEEE INFOCOM*, 2012, pp. 3081–3085.

[43] R. Dhar, G. George, and A. Malani, "Supporting integrated MAC and PHY software development for the USRP SDR," in *Proc. Netw. Technol. Softw. Defined Radio Netw.*, Mar. 2006, pp. 68–77.

[44] Y. Yan, P. Yang, L. You, and B. Li, "Demo abstract: Online optimal channel sensing, probing, accessing in usrp networks," in *Proc. IEEE/ACM ICCPS*, 2012, p. 225.

[45] Y. Liu *et al.*, "Citysee: Not only a wireless sensor network," *IEEE Netw.*, vol. 27, no. 5, pp. 42–47, Sep./Oct. 2013.



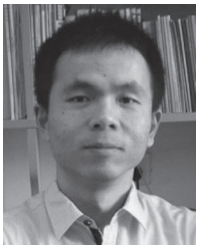
Panlong Yang (M'02) received the B.S., M.S., and Ph.D. degrees in communication and information system from Nanjing Institute of Communication Engineering, Nanjing, China, in 1999, 2002, and 2005 respectively. During September 2010 to September 2011, he was a Visiting Scholar with HKUST. He is now an Associate Professor at the Nanjing Institute of Communication Engineering, PLA University of Science and Technology. His research interests include wireless mesh networks, wireless sensor networks and cognitive radio networks.

Dr. Yang has published more than 50 papers in peer-reviewed journals and refereed conference proceedings in the areas of mobile ad hoc networks, wireless mesh networks and wireless sensor networks. He has also served as a member of program committees for several international conferences. He is a member of the IEEE Computer Society and ACM SIGMOBILE Society.

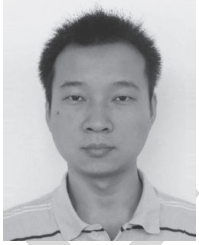
Bowen Li (S'11) received the B.S. degree in wireless communication from the Institute of Communication Engineering, PLA University of Science and Technology, Nanjing, China, in 2007. He is currently working toward the Ph.D. degree from PLA University of Science and Technology. His current research interests include stochastic optimization in cognitive radio networks, and energy efficient algorithm design for wireless sensor networks. He is a student member of the IEEE.

Jinlong Wang received the B.S. degree in mobile communications and the M.S. and Ph.D. degrees in communications engineering and information systems from Institute of Communications Engineering, Nanjing, China, in 1983, 1986, and 1992, respectively. He is a Full Professor of the Institute of Communications Engineering, PLA University of Science and Technology. His current research interests are the broad area of digital communications systems with emphasis on cooperative communication, adaptive modulation, multiple-input-multiple-output systems, soft defined radio, cognitive radio, green wireless communications, and game theory.

Xiang-Yang Li (M'99–SM'08–F'15) received the bachelor's degrees from the Department of Computer Science and the Department of Business Management, Tsinghua University, P.R. China, both in 1995, and the M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign in 2000 and 2001, respectively. He is a Professor at the Illinois Institute of Technology. He is an IEEE Fellow and an ACM Distinguished Scientist. He holds EMC-Endowed Visiting Chair Professorship at Tsinghua University. He is a recipient of China NSF Outstanding Overseas Young Researcher (B). His research interests include wireless networking, mobile computing, security and privacy, cyber physical systems, smart grid, social networking, and algorithms. He and his students won four best paper awards, one best demo award and was nominated for best paper awards twice (ACM MobiCom 2008 and ACM MobiCom 2005). He published a monograph "Wireless Ad Hoc and Sensor Networks: Theory and Applications."



Zhiyong Du (S'12) received the B.S. degree in electronic information engineering from Wuhan University of Technology, Wuhan, China, in 2009. He is currently working toward the Ph.D. degree in communications and information system at the College of Communications Engineering, PLA University of Science and Technology. His research interests include heterogeneous wireless networks, 5G, quality of experience (QoE), learning theory and game theory.



Yubo Yan (S'10) received the B.S. and M.S. degrees in communication and information system from the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2006 and 2011, respectively. He is currently working towards the Ph.D. degree at the PLA University of Science and Technology. His current research interests include software radio systems and wireless sensor networks. He is a student member of the IEEE and the IEEE Computer Society.

Yan Xiong was born in Anhui Province, in 1960. He is a Professor with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include distributed processing, mobile computation, and information security.