# Almost Optimal Dynamically-Ordered Channel Sensing and Accessing for Cognitive Networks

Bowen Li, Panlong Yang, Jinlong Wang, Qihui Wu, Shaojie Tang, Xiang-Yang Li, Yunhao Liu

**Abstract**—For cognitive wireless networks, one challenge is that the status and statistics of the channels' availability are difficult to predict. Numerous learning based online channel sensing and accessing strategies have been proposed to address such challenge. In this work, we propose a novel channel sensing and accessing strategy that carefully balances the channel statistics exploration and multichannel diversity exploitation. Unlike traditional MAB-based approaches, in our scheme, a secondary cognitive radio user will sequentially sense the status of multiple channels in a carefully designed order. We formulate the online sequential channel sensing and accessing problem as a *sequencing multi-armed bandit problem*, and propose a novel policy whose regret is in optimal logarithmic rate in time and polynomial in the number of channels. We conduct extensive simulations to compare the performance of our method with traditional MAB-based approach. Simulation results show that the proposed scheme improves the throughput by more than $30\%$ and speeds up the learning process by more than $100\%$.

**Index Terms**—cognitive radio networks, online sequential sensing and accessing, sequencing multi-armed bandit problem, multichannel diversity.

## 1 Introduction

Driven by regulatory initiatives and radio technology advances, dynamic spectrum access (DSA), especially enabled by cognitive radios (CR), has been well recognized as an effective method to improve spectrum utilization. In DSA system, the cognitive user is regulated to perform spectrum sensing before transmitting over a candidate channel, so as to protect primary user's communication. Due to hardware limitations, cognitive user can only sense a small portion of the spectrum band at a time[1], and thus properly arranging sensing and accessing is critical for improving system throughput as well as reducing access delay. In achieving optimal channel sensing and accessing scheme, a major challenge is predicting the channel status and quality. Online learning scheme, due to the adaptivity and efficiency inherently for dynamic wireless network, has received much attention.

Assuming cognitive user would only sense/access one channel in each time slot, existing online channel sensing and accessing solutions often model the learning process as a multi-armed bandit (MAB) problem [1]. Although the *one channel per slot* scheme is widely used in periodical and synchronized spectrum sensing system, it fails to exploit instantaneous opportunities among channels, i.e. multichannel diversity. Such diversity is widespread in distributed dynamic spectrum access system, since the available channels are usually much more than users could use, e.g., with half of the US population having more than 20 TV channels available for white-space communication at a time [2]. Meanwhile, the channel sensing time is much shorter than the duration of an access time slot, e.g., the sensing time is typically about 10ms, while the access duration is 2s in TV band [3]. In such system, user could gain more by sensing multiple channels sequentially and opportunistically access the idle channel in each time slot.

Motivated by these facts, we investigate the online Sequential channel Sensing and Accessing (SSA) schemes, where the channels' statistics is unknown initially. Our objective is to optimize the total throughput achieved during system lifetime by carefully selecting the sequence of channels to be sensed in each time slot. We formulate the problem on learning the optimal channel sensing order in a stochastic setting as a new bandit problem, which we referred as a *sequencing multi-armed bandit problem* (SMAB). In this formulation, we map each sensing order (i.e. a sequence of channels) to an arm. Same as all MAB-based approaches, we will study the regret of our scheme, where the regret is defined as the difference between the expected reward gained by a genie-based optimal choice (i.e. optimal sensing order), and the reward obtained by our learning policy.

Clearly, both long-term statistics and short-term diversity among different licensed channels would be jointly explored and exploited in our model, which makes our problem distinctive and challenging. Observe that the number of arms in SMAB is exponential, *i.e.*, it is $O(N^K)$ where $N$ is the total number of channels and $K$ is the maximum number of channels user could sense in one time slot. This complexity brings the first

- *B. Li, Y. Liu and X. Li are with the School of Software and TNLIST, Tsinghua University, and X. Li is also with CS Department, Illinois Institute of Technology.*
- *Panlong Yang, Jinlong Wang, and Qihui Wu are with Institute of Communication Engineering, PLAUST, Nanjing, China.*
- *S. Tang is with the Department of Computer and Information Science, Temple University.*

1. Without loss of generosity, in this work, we consider that user can only sense (or transmit over) one channel at a time

Digital Object Indentifier 10.1109/TMC.2013.98 

challenge in devising an efficient online learning policy, as traditional MAB solutions [4]–[6] would result in exponential throughput loss with the increasing number of channels. Moreover, the rewards from different arms are no longer independent to each other, because the elements of different arms are derived from the same channel set. Consequently, previous results under the assumption of independent arms [4]–[6] are no longer applicable to our model. We notice that recent studies [7]–[9] paid attention to the case where the rewards of different arms are correlated, however, all their work focused on linear rewards model, where the expected reward of an arm is defined as a linear function of random variables. While in our model, the correlation between arms is nonlinear, which is the second challenge in designing and analyzing our scheme. Finally, unlike previous work applying MAB in dynamic spectrum access [10]–[16] that channels being sensed during a time slot is fixed, the channels being sensed in a time slot is unpredictable in our model, even the sensing order is given. This distinctive feature makes the learning process in SMAB much more difficult to quantify and analyze.

The main contributions of our paper are as follows.

Firstly, we analyze the performance of classic UCB1 algorithm [6] in handling the online SSA problem, and show that both regret and storage are exponential with the increasing number of channels. We then develop an improved policy referred to as *UCB1 with virtual sampling* (UCB1-VS), exploring dependencies among arms. We show by analytical analysis and extensive simulations that the UCB1-VS scheme significantly improves the convergence speed of the learning process.

Secondly, we propose a novel *sequencing confidence bound* (SCB) algorithm for the SMAB problem. Our analysis shows that, the expected regret of SCB is bounded by a function grows in order $O(NK \log L)$, where $L$ is the number of time slots. That is, the regret is increasing in the optimal logarithmic rate in time and polynomial in the number of unknown parameters. Meanwhile, the storage overhead is significantly reduced from $O(N^K)$ to $O(N)$.

The rest of the paper is organized as follows. We present our system model with problem formulation in Section 2. The proposed online sequential channel sensing and accessing policies are presented in Section 3. We present the extensive simulations and results in Section 4 and review the related work in Section 5. Section 6 concludes our work.

## 2 System Model and Problem Formulation

Consider a cognitive radio network with potential channel set $S = \{1, 2, \ldots, N\}$. Each cognitive user is operated in *constant access time* (CAT) mode [17], i.e., user would have a constant duration $T$ for either channel observation or data transmission once it obtains a communication chance. The communication chance may come from wining competition in control channel as in ad hoc system [17] or assigned by a center node in one hop system [18].

Denote $a_i(j) \in \{0, 1\}$ as the availability of channel $i$ in the $j^{th}$ slot, where $a_i(j) = 0$ indicates the primary user is transmitting over channel $i$ in the $j^{th}$ slot, and $a_i(j) = 1$, vice versa. In order to protect the primary users' communication process, the duration $T$ is commonly set to be much shorter than the sojourn time of primary user activities. It is reasonable to consider that the channel state is relatively stable during $T$. On the other hand, as the interval time between consecutive communication chances is relatively long in multi-user networks (as discussed in [17]), the channel status is considered to be independent among slots. This basic assumption is consistent with previous studies such as [11], [15], [19], [20]. We consider that the channel idle probability $\theta_i \in [0, 1]$ ($i \in S$) is not known to user at the beginning, but can be available through learning. For denotation convenience, we sort the channel according to idle probability, where $\theta_{1'} \geq \theta_{2'} \geq \ldots \geq \theta_{N'}$.

It is worth noting that, in our cognitive radio network scenario, our scheme can work in both cases where primary users presented or not. Note that, the channel idle probability is dominated by the access frequency of the primary users. What left to secondary user is to learn and select the dynamic available channels wisely. In case there is no primary users, e.g. the dynamic spectrum utilization scenario, the channel availability is dominatedby the channel contention process as well as the channel quality. In summary, we consider the typical cognitive radio network scenario as well as dynamic spectrum utilization problem, and solve them in one theoretical framework.

The sequential channel sensing and transmission procedure can be described as follows. At each slot, user senses the channels sequentially according to a given sensing order, until it arrives at an idle channel, and transmits over this channel during the remainder of the time slot with data rate $R$. Each channel sensing is denoted as a step in a slot, which costs a constant time $\tau_s$. We denote $\Psi$ as the set of all possible sensing orders. Each element in $\Psi$, $\Phi_m = (\phi_1^m, \phi_2^m, \ldots, \phi_K^m)$, is a permutation of the $K$ channels, where $K$ is the maximum number of steps in each decision slot, and $\phi_k^m$ denotes the ID of $k^{th}$ channel in $\Phi_m$. Correspondingly, $K = \min\left(N, \lfloor \frac{T}{\tau_s} \rfloor\right)$ ($\lfloor \cdot \rfloor$ is round-down function), and $|\Psi| = M = \begin{pmatrix} N \\ K \end{pmatrix} K!$. When a user stops at step $k$ (i.e, $a_{s_k} = 1$ in current slot), it obtains an immediate data transmission reward $R(T - k\tau_s)$.

We define the deterministic policy $\pi(j)$ at each time $j$, mapping from the observation history $\mathcal{F}_{j-1}$ to a sequence of channels $\Phi(j)$. The problem is how to make sequential decision on sensing order selection, offering stochastic rewards with unknown distribution. Our main goal is to devise a learning policy maximizing the accu-

TABLE 1
Summary of Notations

| Notation | Description |
|---|---|
| $S$ | channel set, $S = \{1, 2, \ldots, N\}$ |
| $\Psi$ | sensing order set, $\Psi = \{\Phi_1, \Phi_2, \ldots, \Phi_M\}$ |
| $M$ | number of sensing orders |
| $N$ | number of channels |
| $K$ | number of maximum sensing steps in one slot |
| $i$ | channel index, $1 \leq i \leq N$ |
| $j$ | slot index, $1 \leq j \leq L$ |
| $k$ | sensing/probing step index in a slot, $1 \leq k \leq K$ |
| $m$ | sensing order index, $1 \leq m \leq M$ |
| $\alpha$ | normalized time cost for sensing one channel |
| $\phi_k^m$ | ID of the $k^{th}$ channel in $\Phi_m$ |
| $r_{\Phi_m}(j)$ | immediate reward obtained by user using order $\Phi_m$ in the $j^{th}$ slot |
| $\mu_m$ | expected reward per slot of user selecting $\Phi_m$ |
| $\hat{\mu}_m$ | estimated average reward of selecting $\Phi_m$ |
| $n_m$ | number of times that sensing order $\Phi_m$ has been selected up to the current slot |
| $a_i(j)$: | availability of channel $i$ in the $j^{th}$ slot |
| $\Theta$: | channel idle probability, $\Theta = (\theta_1, \theta_2, \ldots, \theta_N)$ |
| $\hat{\Theta}$: | estimated idle probability, $\hat{\Theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N\right)$ |
| $\Theta^u$: | upper confidence of idle probability, $\Theta^u = \left(\theta_1^u, \theta_2^u, \ldots, \theta_N^u\right)$ |
| $n_i^s$ | number of times that channel $i$ has been sensed up to the current slot |
| $L_k$ | number of slots user arrives the $k^{th}$ sensing step up to $L$ slots |
| $T_i(j)$ | number of times $i$ has been sensed up to the $j^{th}$ slot |
| $T_i^k(j)$ | number of times $i$ has been sensed in the $k^{th}$ step up to the $j^{th}$ slot |

mulated throughput, i.e.,

$$\max \lim_{L\to\infty} \sum_{j=1}^{L} \mu_{\pi(j)}$$

where $\mu$ is the expected reward in one slot time according to order $\Phi$. Let $\alpha = \frac{\tau_s}{T}$. When order $\Phi_m$ is chosen, the expected per-slot reward is given by

$$\mu_m = E\left[r_{\Phi_m}\right] = \sum_{k=1}^{K} \left\{ (1 - k\alpha)\theta_{\phi_k^m} \prod_{\kappa=1}^{k-1} \left(1 - \theta_{\phi_\kappa^m}\right) \right\} \quad (1)$$

Here, $r_{\Phi_m}$ is the normalized immediate reward obtained using order $\Phi_m$. Without special notification, the rewards we talked about are normalized. To obtain the actual throughput, the reward should be scaled by constant factor $RT$.

Since maximizing accumulated throughput is equivalent to minimizing the *regret*, we can rewrite the objective function as

$$\min \lim_{L\to\infty} \rho_\pi(L) = L\mu^* - \sum_{j=1}^{L} \mu_{\pi(j)} \quad (2)$$

A scheme is *zero-regret* if the average regret per time slot tends to 0 when the duration $L$ goes to infinity, i.e. $\lim_{L\to\infty} \frac{\rho_\pi(L)}{L} = 0$.

We summarize the main notations in Table 1.

## 3 Almost Optimal Online Sequential Sensing and Accessing

In this section, we first propose two intuitive methods to construct sensing order selection strategy. The first one directly applies UCB1 [6], and the second one is referred to as *UCB1 with virtual sampling* (UCB1-VS), which is an improved version of UCB1 by exploring the dependency among arms. We then develop *sequencing confidence bound* (SCB) for such SMAB problem and analyze the regret of this novel algorithm.

### 3.1 Intuitive Solutions Based on UCB1

#### 3.1.1 UCB1 Algorithm

An intuitive approach to solve the sequencing multi-armed bandit problem is to use the UCB1 policy given by Auer et al. [6]. In supporting sensing order selection, two variables are used for each candidate order $\Phi_m$ ($1 \leq m \leq M$): $\hat{\mu}_m(j)$ is the averaged value of all the obtained rewards in order $\Phi_m$ up to slot $j$, and $n_m(j)$ is the number of times that $\Phi_m$ has been chosen up to slot $j$. They are both initialized to zero and updated according to the following rules:

$$\hat{\mu}_m(j) = \begin{cases} \frac{\hat{\mu}_m(j-1)n_m(j-1) + r_m(j)}{n_m(j-1)+1}, & \Phi_m \text{ is selected} \\ \hat{\mu}_m(j-1), & \text{else} \end{cases} \quad (3)$$

$$n_m(j) = \begin{cases} n_m(j-1) + 1, & \Phi_m \text{ is selected} \\ n_m(j-1), & \text{else} \end{cases} \quad (4)$$

Then, the intuitive policy can be described as: at the very beginning, the user should choose each sensing order only once. After that, the user should select the order $\Phi_m$ that maximizes $\hat{\mu}_m + \sqrt{\frac{2\log j}{n_m}}$ in the $j^{th}$ slot. The description of such policy is presented in Fig.1.

The regret bound of the UCB1 policy is given by the following theorem.

*Theorem 1:* The expected regret of sequential sensing/accessing under policy UCB1 is at most

$$\left[8 \sum_{m:\mu_m<\mu^*} \left(\frac{\log L}{\xi_m}\right)\right] + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{m:\mu_m<\mu^*} \xi_m\right) \quad (5)$$

where $\xi_m = \mu^* - \mu_m$.

*Proof:* See ( [6], Theorem 1). □

As the number of candidate orders $M = \begin{pmatrix} N \\ K \end{pmatrix} K!$, according to Equ. (5), we conclude that the regret under UCB1 policy is $O(N^K \log L)$, which is in optimal logarithmic increasing rate in time. Intuitively, the UCB1 policy is zero regret. However, it is worth to note that the regret of UCB1 is exponential increasing in the number

UCB1

1: Initialize: $j = 0$; for all $1 \leq m \leq M$: $\hat{\mu}_m = 0$, $n_m = 0$
2: **for** $j = 1$ to $M$ **do**
3: Sequentially sensing/accessing with $\Phi(j) = \Phi_j$ in the $j^{th}$ slot
4: Update $\hat{\mu}_j(j)$, $n_j(j)$ using Equ. (3)-(4) respectively
5: **end for**
6: **for** $j = M+1$ to $L$ **do**
7: Sequentially sensing/accessing with $\Phi(j) = \Phi_m$ that maximizes $\hat{\mu}_m + \sqrt{\frac{2\log j}{n_m}}$ in the $j^{th}$ slot
8: Update $\hat{\mu}_m(j)$, $n_m(j)$ using Equ. (3)-(4) respectively
9: **end for**

Fig. 1. UCB1 algorithm description

UCB1 with Virtual Sampling

1: Initialize: $j = 0$; for all $1 \leq m \leq M$: $\hat{\mu}_m = 0$, $n_m = 0$
2: **for** $j = 1$ to $M$ **do**
3: Sequentially sense and access with $\Phi(j) = \Phi_j$ in the $j^{th}$ slot
4: Update $\hat{\mu}_j(j)$, $n_j(j)$ according to sub-algorithm (as in Fig.3) with Equ. (3)-(4) respectively
5: **end for**
6: **for** $j = M+1$ to $L$ **do**
7: Sequentially sense and access with $\Phi(j) = \Phi_m$ that maximizes $\hat{\mu}_m + \sqrt{\frac{2\log j}{n_m}}$ in the $j^{th}$ slot
8: Update $\hat{\mu}_j(j)$, $n_j(j)$ according to sub-algorithm (as in Fig.3) with Equ. (3)-(4) respectively
9: **end for**

Fig. 2. UCB1-VS algorithm description

Virtual Sampling sub-algorithm

1: **if** $r(j) == 0$ **then**
2: Update statistics of all sensing orders consisting of $\{\psi_1(j), \psi_2(j), \ldots, \psi_K(j)\}$ with $r(j) = 0$
3: **else if** $r(j) == 1-\alpha$ **then**
4: Update statistics of all sensing orders starting with $\psi_1(j)$ using $r(j) = 1-\alpha$
5: **else if** $r(j) == 1-k\alpha$, $(k > 1)$ **then**
6: Update statistics of all sensing orders starting with $\psi_k(j)$ using $r(j) = 1-\alpha$
7: Update statistics of all sensing orders starting with a permutation of $\{\psi_1(j), \psi_2(j), \ldots, \psi_{k-1}(j)\}$ and the $k^{th}$ channel is $\psi_k(j)$, using $r(j) = 1-k\alpha$
8: **end if**

Fig. 3. Virtual Sampling sub-algorithm description

of channels, which results in high throughput loss in actual sequential sensing/accessing scenario[2].

Note that since all the orders are composed of the channels from the same set, the rewards of different orders can not be independent to each other. Hence, it is possible to improve learning efficiency by further exploring dependency among arms. Motivated by this fact, we develop an improved UCB1-based policy as follows.

#### 3.1.2 Improved UCB1-VS Algorithm

Analyze the reward function in SSA process (as in Equ.1). The reward in our sequencing multi-armed bandit problem depends on the specific channels as well as their positions in the order, which indicates that the orders with identical sub-sequence would result in similar rewards. This basic finding provides us an important hint that we could obtain reward information about multiple arms by playing a single arm. As a result, it is possible to update statistics of multiple sensing orders without really playing with it in each slot, which we referred as *virtual sampling*.

We illustrate the principle of our proposed UCB1-VS policy as follows. Suppose that user selects an order $\Phi_m = (\psi_1^m, \psi_2^m, \ldots, \psi_K^m)$ in a slot and access $\psi_k^m$ ($\psi_k^m$ is idle in current slot), then user obtains reward $1-k\alpha$, and we can conclude that, in current slot:

- If user selects an arbitrary sensing order starting with a subsequence consisting of $\psi_1^m, \psi_2^m, \ldots, \psi_{k-1}^m$ and the $k^th$ channel is $\psi_k^m$, the reward must be $1-k\alpha$;
- If user selects an arbitrary sensing order starting with $\psi_k^m$, the reward must be $1-\alpha$.

Moreover, if user observes that all channels are busy in current slot, then we can conclude that all sensing orders consisting of $\{\psi_1^m, \psi_2^m, \ldots, \psi_K^m\}$ would result in zero reward in current slot. Motivated by these facts, we develop the UCB1-VS algorithm as presented in Fig.2.

Clearly, with virtual sampling, the learning process can be greatly accelerated, while the zero-regret property still holds. We show the benefit of virtual sampling by analyzing the number of arms being updated in one slot. Suppose that user chooses order $\Phi_m = (\psi_1^m, \psi_2^m, \ldots, \psi_K^m)$ to run SSA in a slot, the expected number of sensing orders that would be updated in this slot is then given by

$$\begin{aligned} M_{vs} =& \sum_{k=1}^{K} \left\{ \left[(K-1)! + (k-1)!\right] \theta_{\psi_k^m} \prod_{\kappa=1}^{k-1} \left(1-\theta_{\psi_\kappa^m}\right) \right\} \\ &+ K! \prod_{\kappa=1}^{K} \left(1-\theta_{\psi_\kappa^m}\right) \\ >& (K-1)! \left\{ \sum_{k=1}^{K} \left[ \theta_{\psi_k^m} \prod_{\kappa=1}^{k-1} \left(1-\theta_{\psi_\kappa^m}\right) \right] + \prod_{\kappa=1}^{K} \left(1-\theta_{\psi_\kappa^m}\right) \right\} \\ =& (K-1)! \end{aligned}$$

2. As the number of channels is commonly large, e.g., the number of whitespace channels for DSA in TV band is over 50 [3].

It is clearly shown that at least $(K-1)!$ arms are updated in each slot, rather than only one as in UCB1. Moreover, as the number $(K-1)!$ is obtained when user accesses after only one step channel sensing, it would be further increased when the number of channels sensed in a slot increases. This indicates that the virtual sampling would benefit more in a spectrum scarcity environment, where the learning gain is most needed.

Although the expected regret of UCB1 and UCB1-VS are in optimal logarithmic rate over time, they are exponentially increasing with the number of channels. Moreover, as the choices are made on arm-specific statistics, two variables are recorded for each sensing order. Thus, the storage overhead of running these two algorithms is $O(N^K)$. On the other hand, in each decision-making, e.g., in the $j^{th}$ slot, the user needs to choose the order $\Phi_m$ with maximum $\hat{\mu}_m + \sqrt{\frac{2\log j}{n_m}}$ from $M = \begin{pmatrix} N \\ K \end{pmatrix} K!$ candidate sensing orders. Hence, $O(N^K)$ compare operations are needed for each decision-making. In a word, both storage overhead and computation complexity of these two algorithms are exponentially growing with the number of channels. This property makes them infeasible in most practical scenarios, where the number of channels is commonly up to ten or even more.

### 3.2 A Novel Algorithm for Sequencing Bandit Problem

In this subsection, we propose a novel learning policy for sequential channel sensing and accessing, in which decisions are made on channel-specific statistics. Later, theoretical analysis on the regret bound is proposed, which shows that the regret of our proposed algorithm is in polynomial order of the number of channels.

#### 3.2.1 Algorithm Description

We describe our proposed learning algorithm by introducing the concept of *optimistic throughput*, which is defined as the supposed achievable throughput under the optimistic channel estimation. We use the confidence interval estimation to characterize the optimism of uncertainty in channel statistics, and *sequencing confidence bound* (SCB) is proposed for quantifying the optimistic throughput. Then, the core idea of our proposed SCB learning policy can be briefly described as: choosing the sensing order that maximizes the sequencing confidence bound in each slot.

In decision-making process, the channel statistics is learnt by recording and updating the following two variables: $\hat{\theta}_i(j)$ and $n_i^s(j)$, where $\hat{\theta}_i(j)$ and $n_i^s(j)$ are the statistic value of idle probability and the times having been sensed for channel $i$ till slot $j$ respectively. They

Sequencing Confidence Bound

1: Initialize: for all $1 \le i \le N$: $\hat{\theta}_i = 0$, $n_i^s = 0$; $S_0 = S$; $l = 1$, $k = 1$;
2: **while** $S_0 \neq \emptyset$ **do**
3: Sense a random channel $i \in S_0$
4: $k = k + 1$, $S_0 = S_0 \setminus \{i\}$
5: Update $\hat{\theta}_i(l)$, $n_i^s(l)$ according to Equ.(6) (7) respectively
6: **if** $a_i^l == 1$ **then**
7: $l = l + 1$, $k = 1$
8: Access the idle channel
9: **else if** $k == K + 1$ **then**
10: $l = l + 1$, $k = 1$
11: Wait for the next slot
12: **end if**
13: **end while**
14: **for** $j = l$ to $L$ **do**
15: Sequentially sense and access according to $\Phi$ where
$$\Phi = \arg \max_{\Phi_m \in \Psi} SCB_m(j)$$
16: Update $\hat{\theta}_i(j)$, $n_i^s(j)$ accordingly
17: **end for**

Fig. 4. SCB algorithm description

are initialized to zero and updated as follows:

$$\hat{\theta}_i(j) = \begin{cases} \frac{\hat{\theta}_i(j-1)n_i(j-1)+a_i^j}{n_i(j-1)+1}, & \text{if channel } i \text{ is sensed} \\ \hat{\theta}_i(j-1), & \text{else} \end{cases} \tag{6}$$

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

Then, the SCB learning policy can be described as follows. Firstly, user will sequentially sense channels until all channels are visited at least once. After that, in time slot $j$, the user will choose the sensing order $\Phi_m$ with maximum $SCB_m(j)$, which is defined by

$$SCB_m(j) = \sum_{k=1}^{K} \left\{ (1-k\alpha)\theta^u_{\psi_k^m}(j) \prod_{\kappa=1}^{k-1} \left(1 - \theta^u_{\psi_\kappa^m}(j)\right) \right\} \tag{8}$$

Here $\theta_i^u(j) = \hat{\theta}_i(j) + \sqrt{\frac{2\log j}{n_i^s(j)}}$ is the upper confidence bound of the idle probability on channel $i$ up to slot $j$. The detailed SCB algorithm is presented in Fig.4.

We now analyze the complexity of running SCB algorithm. Compare with the arm-specific algorithms (i.e. UCB1 and UCB1-VS) that need $O\left(N^K\right)$ storage space, the storage overhead of SCB is greatly reduced to be $O\left(N\right)$, since the decision making in SCB needs only two variables for each channel. Moreover, for the computational complexity, we have the following lemma.

*Lemma 1:* The channel sequence with first $K$ elements in the descending order of $\theta^u$ is optimal for maximizing $SCB_m$.

*Proof:* Given the real channel statistics is $\Theta^u$, $SCB_m$ is the expected reward of sensing order $\Phi_m$. Hence, the lemma is equivalent to the statement that *the order* $\Phi = (1', 2', \ldots, K')$, *where* $\theta_{1'} \geq \theta_{2'} \geq \ldots \geq \theta_{K'} \geq \ldots \geq \theta_{N'}$, *is optimal for maximizing expected reward.*

Define $U_k$ as the expected reward one could obtain in the $k^{th}$ sensing step. $U_1$is then the expected reward one could obtain by using the sense $U_{K+1} = 0$ (i.e., all the channels sensed in this slot are busy). We prove it by contradiction in following two phases.

*Phase 1:* Suppose an optimal sensing order is $\Phi = (1', 2', \ldots, K')$, and there exists $k_1 < K$, $k_2 > K$ such that $\theta_{k_1'} < \theta_{k_2'}$. Then, we have

$$U_{k_1'} = \theta_{k_1'}(1 - k_1\alpha) + U_{k_1+1'}$$

We get a new sensing order by switching the order of channels $k_1{}'$ and $k_2{}'$, and keeping all the other channels the same. We have

$$U_{k_1'}^{new} = \theta_{k_2'}(1 - k_1\alpha) + U_{k_1+1'}$$

Since $\theta_{k_1'} < \theta_{k_2'}$, it is easily to acquire that $U_{k_1'} < U_{k_1'}^{new}$.

Further, we have

$$\begin{aligned} U_{k_1-1'} &= \theta_{k_1-1'}(1 - k_1\alpha + \alpha) + U_{k_1'} \\ &< \theta_{k_1-1'}(1 - k_1\alpha + \alpha) + U_{k_1'}^{new} = U_{k_1-1'}^{new} \end{aligned}$$

Similarly, we have $U_1 < U_1^{new}$. This contradicts the assumption that the sensing order $\Phi = (1', 2', \ldots, K')$ is optimal. In other words, if $\Phi = (1', 2', \ldots, K')$ is optimal, then $\theta_{k_1'} \geq \max_{k_2>K}\{\theta_{k_2'}\}, \forall k_1 \leq K$.

*Phase 2:* Suppose an optimal sensing order is $\Phi = (1', 2', \ldots, K')$, and there exists $k < K$ such that $\theta_{k'} < \theta_{k+1'}$. Then, we have $U_k = \theta_{k'}(1 - k\alpha) + (1 - \theta_{k'})[\theta_{k+1'}(1 - k\alpha - \alpha) + (1 - \theta_{k+1'})U_{k+2'}]$. We get a new sensing order by switching the order of channels $k'$ and $k + 1'$, and keeping all the other channels the same. We have $U_k^{new} = \theta_{k+1'}(1 - k\alpha) + (1 - \theta_{k+1'})[\theta_{k'}(1 - k\alpha - \alpha) + (1 - \theta_{k'})U_{k+2'}]$.

Then, $U_k^{new} - U_k = \alpha(\theta_{k+1'} - \theta_{k'}) > 0$, since $\theta_{k'} < \theta_{k+1'}$. Further, we have

$$\begin{aligned} U_{k-1} &= \theta_{k-1'}(1 - k\alpha + \alpha) + (1 - \theta_{k-1'})U_k \\ &< \theta_{k-1'}(1 - k\alpha + \alpha) + (1 - \theta_{k-1'})U_k^{new} = U_{k-1}^{new} \end{aligned}$$

Similarly, we have $U_1 < U_1^{new}$. This contradicts the assumption that the sensing order $\Phi = (1', 2', \ldots, K')$ is optimal. In other words, if $\Phi = (1', 2', \ldots, K')$ is optimal, then $\theta_{k_1'} \geq \theta_{k_2'}, \forall k_1 < k_2 \leq K$.

This completes the proof. □

.

**Remark:** As the decisions are made on channel-specific statistics in SCB, the storage overhead for running SCB algorithm is only $O(N)$, i.e., linear to the number of channels. In the computation complexity aspect, the Lemma1 indicates that one could acquire the order with maximum SCB by arranging $K$ channels from all the $N$ channels with the descending order of $\theta^u$. In this case, the user needs only $NK$ compare operations to derive the exact sensing order. Moreover, the decision on determining a sensing order in a slot can be accomplished by a sequence of decisions on selecting channel step by step during the slot. Specifically, in each slot, the user could maximize $SCB$ in practice by always choosing the channel with highest $\theta^u$ from the candidate channel set in each step (and then sweeping the chosen channel from candidate channel set for the channel selection in the next step). By such way, although the user needs to make a series of decisions on selecting one channel from the candidate channel set in each slot; the computation complexity in each decision-making is at most $O(N)$ compare operations.

### 3.2.2 Analysis of Regret

In this subsection, we study the regret of the proposed SCB policy. Traditionally, the regret of a policy for a multi-armed bandit problem is upper-bounded by the number of times each sub-optimal arm being played. Summing over all sub-optimal arms can get the upper bound. However, since our proposed approach focuses on the basic elements of each arm (i.e. the sub-sequences in each sensing order), it requires more finely analysis. We analyze the number of times each sub-optimal channel being sensed in each step, and sum up this expectation over all channels with all steps. Our analysis provides an upper bound which is polynomial to $N$ and logarithmic to time. We present our analytical result in the following theorem.

*Theorem 2:* The expected regret of sequential sensing/accessing under the SCB policy is at most

$$\Pi(L)K\left[N - \frac{K+1}{2} - \frac{\alpha(K+1)(3N-2K-1)}{6}\right]$$

where $\Pi(L) = \frac{8\log L}{\Delta_{min}} + \left(1 + \frac{\pi^2}{3}\right)\Delta_{max}$, and $\Delta_{min} = \min_{i,j}|\theta_i - \theta_j|$ $(i \neq j)$, $\Delta_{max} = \max_{i,j}|\theta_i - \theta_j|$.

To prove Theorem 2, we introduce the Chernoff-Hoeffding bound inequalities first.

*Lemma 2:* (Chernoff-Hoeffding bound) 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

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

We now prove Theorem 2 as follows.

*Proof:* Denote $T_i(L)$ as the number of times that channel $i$ having been sensed in the first $L$ slots. $T_i^k(L)$ is the number of times that $i$ has been sensed in the $k^{th}$ step in the first $L$ slots, and $T_i(L) = \sum_{k=1}^{N} T_i^k(L)$. Let $L_0$ be the number of slots SCB consumes till all channels are observed at least once, and $\ell$ be an arbitrary positive

integer, then we have

$$\begin{aligned}T_i^1(L) &\le 1+\sum_{j=L_0+1}^{L} I\left\{ch_j^1=i\right\}\\ &\le \ell+\sum_{j=L_0+1}^{L} I\left\{ch_j^1=i, T_i^1(j-1)\ge \ell\right\}\\ &\le \ell+\sum_{j=L_0+1}^{L} I\Big\{\hat{\theta}_{1'}(j-1)+c_{j-1,T_{1'}(j-1)}\\ &\qquad \le \hat{\theta}_i(j-1)+c_{j-1,T_i(j-1)}, T_i^1(j-1)\ge \ell\Big\}\end{aligned}$$

where $I\{\}$ is an indicator function, $ch_j^1=i$ represents the event that user senses channel $i$ in the $1^{st}$ sensing step of slot $j$, and $c_{x,y}\doteq\sqrt{\frac{2\log x}{y}}$.

Since $\Pr\left\{T_i^1(j-1)\ge\ell\right\} \le \Pr\left\{T_i(j-1)\ge\ell\right\}$, we then obtain

$$\begin{aligned}T_i^1(L) &\le \ell+\sum_{j=L_0+1}^{L} I\Big\{\min_{0<t<j}\hat{\theta}_{1'}(t)+c_{j-1,t}\\ &\qquad \le \max_{\ell\le t_i<j}\hat{\theta}_i(t_i)+c_{j-1,t_i}\Big\}\\ &\le \ell+\sum_{j=1}^{\infty}\sum_{t=1}^{j-1}\sum_{t_i=\ell}^{j-1} I\left\{\hat{\theta}_{1'}(t)+c_{j,t}\le\hat{\theta}_i(t_i)+c_{j,t_i}\right\}\end{aligned}$$

Note that $\hat{\theta}_{1'}(t)+c_{j,t}\le\hat{\theta}_i(t_i)+c_{j,t_i}$, which implies at least one of the following conditions must be held

$$\hat{\theta}_{1'}(t)\le\theta_{1'}-c_{j,t} \tag{9}$$

$$\hat{\theta}_i(t_i)\ge\theta_i+c_{j,t_i} \tag{10}$$

$$\theta_{1'}<\theta_i+2c_{j,t_i} \tag{11}$$

We bound the probability of events happening in Equ.(9) and (10) using Lemma 2

$$\Pr\left\{\hat{\theta}_{1'}(t)\le\theta_{1'}-c_{j,t}\right\}\le e^{-4\log j}=e^{-4}$$

$$\Pr\left\{\hat{\theta}_i(t_i)\ge\theta_i+c_{j,t_i}\right\}\le e^{-4\log j}=e^{-4}$$

For $\ell=\lceil\frac{(8\log L)}{(\theta_1-\theta_i)^2}\rceil$, we have $\theta_{1'}-\theta_i-2c_{j,t_i}=\theta_{1'}-\theta_i-2\sqrt{\frac{2\log j}{t_i}}\ge 0$, which indicates that Equ.(11) is false. Denote $\Delta_i^j\doteq\theta_i-\theta_j$. Then, we get

$$\begin{aligned}E\left[T_i^1(L)\right] &\le \lceil\frac{(8\log L)}{{\Delta_{1'}^i}^2}\rceil+\sum_{j=1}^{\infty}\sum_{t=1}^{j-1}\sum_{t_i=\lceil\frac{(8\log L)}{{\Delta_{1'}^i}^2}\rceil}^{j-1}2j^{-4}\\ &\le \frac{8\log L}{{\Delta_{1'}^i}^2}+1+\frac{\pi^2}{3}\end{aligned}$$

Consequently, the regret of the $1^{st}$ step up to $L$ slots is

$$\begin{aligned}\rho_1(L) &= (1-\alpha)\sum_{i:\theta_i<\theta_{1'}}\Delta_{1'}^i E\left[T_i^1(L)\right]\\ &\le (1-\alpha)\left[8\sum_{i:\theta_i<\theta_{1'}}\frac{\log L}{\Delta_{1'}^i}+\left(1+\frac{\pi^2}{3}\right)\sum_{i:\theta_i<\theta_{1'}}\Delta_{1'}^i\right]\end{aligned}$$

Denote $L_k$ as the expected number of times that a system reaches the $k^{th}$ sensing step up to $L$ slots. Obviously, $L_1=L$. And $L_2$ is given by

$$\begin{aligned}L_2 &= \sum_{i:\theta_i<\theta_{1'}}E\left[T_i^1(L_1)\right](1-\theta_i)\\ &\qquad +(1-\theta_{1'})\left\{L_1-\sum_{i:\theta_i<\theta_{1'}}E\left[T_i^1(L_1)\right]\right\}\\ &= L_1(1-\theta_{1'})+\sum_{i:\theta_i<\theta_{1'}}E\left[T_i^1(L_1)\right]\Delta_{1'}^i\end{aligned}$$

Since $E\left[T_i^1(L)\right]\le\frac{8\log L}{(\theta_{1'}-\theta_i)^2}+1+\frac{\pi^2}{3}$ and $\forall i\ne 1'$, $(1-\theta_i)\ge(1-\theta_{1'})$, we have

$$\begin{aligned}L_2 &\le L_1(1-\theta_{1'})\\ &\qquad +\left[8\sum_{i:\theta_i<\theta_{1'}}\frac{\log L_1}{\Delta_{1'}^i}+\left(1+\frac{\pi^2}{3}\right)\sum_{i:\theta_i<\theta_{1'}}\Delta_{1'}^i\right]\end{aligned}$$

We now consider the regret in the $2^{nd}$ step up to time slot $L$, $\rho_2(L_2)$. In order to derive the bound of the $\rho_2(L_2)$, we analyze the bound of $\rho_1^2(L_2)$ first. Here, $\rho_1^2(L_2)$ denotes the $1^{st}$ step regret under the case that there is totally $L_2$ slots. Denote the candidate channel set $S_2=S_1\setminus\{1'\}$. Under such case, the optimal choice is channel $2'$ with probability $\theta_{2'}$. Similar to the analysis process of $\rho_1(L)$, we have

$$\rho_1^2(L_2)\le(1-\alpha)\left[8\sum_{i:\theta_i<\theta_{2'}}\frac{\log L_2}{\Delta_{2'}^i}+\left(1+\frac{\pi^2}{3}\right)\sum_{i:\theta_i<\theta_{2'}}\Delta_{2'}^i\right]$$

Note that during the sequential channel sensing and accessing process under SCB, user can not always choose channel $1'$ in the first step (the throughput loss due to error selections are counted in regret $\rho_1(L)$). When channel $1'$ is not sensed in the $1^{st}$ step, user will choose channel $1'$ with high probability in the $2^{nd}$ sensing step, rather than channel $i$ with lower $\theta_i$. As a result, we have

$$\rho_2(L_2)\le\frac{(1-2\alpha)}{(1-\alpha)}\rho_1^2(L_2)$$

Let

$$X_k=\left[8\sum_{i:\theta_i<\theta_{k'}}\frac{\log L_k}{\Delta_{k'}^i}+\left(1+\frac{\pi^2}{3}\right)\sum_{i>k}^{N}\Delta_{2'}^i\right] \tag{12}$$

Then, we have

$$\rho_1(L_1)\le(1-\alpha)X_1$$

$$\rho_2(L_2)\le(1-2\alpha)X_2$$

and

$$L_2\le L_1(1-\theta_{1'})+X_1$$

$$L_3\le L_2(1-\theta_{2'})+X_2$$

It is easily to be generalized to arbitrary $1 \leq k \leq K$, i.e.,

$$L_k \leq L_{k-1}\left(1-\theta_{k-1'}\right)+X_{k-1} \tag{13}$$

$$\rho_k(L_k) \leq (1-k\alpha)\, X_k \tag{14}$$

where the initial $L_1 = L$.

Then, the the total regret

$$\rho(L) = \sum_{k=1}^{K} \rho_k(L_k) \leq \sum_{k=1}^{K} (1-k\alpha)\, X_k \tag{15}$$

Leveraging the intermediate results in Equ. (12)(13) and (14), we have

$$\begin{aligned}\rho(L) \leq (1-\alpha) & \sum_{i:\theta_i<\theta_{1'}} \left[\frac{8\log L_1}{\Delta_{1'}^i} + \left(1+\frac{\pi^2}{3}\right)\Delta_{1'}^i\right] \\ & + (1-2\alpha) \sum_{i:\theta_i<\theta_{2'}} \left[\frac{8\log L_2}{\Delta_{2'}^i} + \left(1+\frac{\pi^2}{3}\right)\Delta_{2'}^i\right] \\ & + \cdots \\ & + (1-K\alpha) \sum_{i:\theta_i<\theta_{K'}} \left[\frac{8\log L_K}{\Delta_{K'}^i} + \left(1+\frac{\pi^2}{3}\right)\Delta_{K'}^i\right]\end{aligned}$$

As for arbitrary $x' < y' \leq N'$: $\Delta_{x'}^{x+1'} \leq \Delta_{x'}^{y'} \leq \Delta_{x'}^{N'}$, the above inequality can be rewritten as

$$\begin{aligned}\rho(L) \leq (1-\alpha)\,(N-1) & \left[\frac{8\log L_1}{\Delta_{1'}^{2'}} + \left(1+\frac{\pi^2}{3}\right)\Delta_{1'}^{N'}\right] \\ & + (1-2\alpha)\,(N-2) \left[\frac{8\log L_2}{\Delta_{2'}^{3'}} + \left(1+\frac{\pi^2}{3}\right)\Delta_{2'}^{N'}\right] \\ & + \cdots \\ & + (1-K\alpha)\,(N-K) \left[\frac{8\log L_K}{\Delta_{K'}^{K+1'}} + \left(1+\frac{\pi^2}{3}\right)\Delta_{K'}^{N'}\right]\end{aligned}$$

Define function $\Pi(L) = \frac{8\log L}{\Delta_{min}} + \left(1+\frac{\pi^2}{3}\right)\Delta_{max}$, where $\Delta_{min} = \min_{(i,j)} |\Delta_j^i|$ and $\Delta_{max} = \max_{(i,j)} |\Delta_j^i| = \Delta_{1'}^{N'}$. Note that $L_K < L_{K-1} < \ldots < L_1 = L$. We then have

$$\begin{aligned}\rho(L) &< \Pi(L) \sum_{k=1}^{K} [(1-k\alpha)\,(N-k)] \\ &= \Pi(L) \left[\frac{\alpha}{6}K(K+1)(2K+1) + KN \right. \\ &\qquad \left. -\frac{1}{2}(1+N\alpha)K(K+1)\right] \\ &= \Pi(L) K \left[N - \frac{K+1}{2} - \frac{\alpha(K+1)(3N-2K-1)}{6}\right]\end{aligned}$$

which concludes the proof. □

As $N \geq K$ and $K \geq 1$, we conclude that $3N-2K-1 \geq 0$. Thus, $N - \frac{K+1}{2} - \frac{\alpha(K+1)(3N-2K-1)}{6} < N$. As a result, our policy achieves an expected regret upper bounded in the order of $O(NK\log L)$, which is in polynomial order to the number of channels and strictly in logarithmic order to time.

## 4 Simulations and Performance Analysis

In this section, we evaluate and analyze the performance of the proposed online sequential channel sensing and accessing algorithms via simulations.

Eight policies are running under the same environment for performance comparison, where UCB1, UCB1-VS and SCB are presented in Section3. The other five policies are described as follows.

- *Randomized Single Channel* (Ran. Sin.): chooses a random channel for sensing/accessing at each slot;
- *Single Index* (Sin. Index): an optimal online learning policy for *one channel per slot* scheme, which is first presented by Lai et al in [10];
- *Optimal Single Channel* (Opt. Sin.): a genie-based policy that user always senses/accesses the channel with highest idle probability at each slot;
- *Randomized Sequence* (Ran. Seq.): chooses a random channel sequence for sequential sensing/accessing at each slot;
- *Optimal Sequence* (Opt. Seq.): a genie-based policy that user always chooses the optimal sensing order for SSA at each slot;

### 4.1 Performance Comparison on Policies

Three performance metrics are considered in this subsection: throughput, regret and learning progress. Throughput presents a direct impression on the variation of system reward over time. Regret shows the accumulated system performance, where a higher regret indicates higher throughput loss over time and thus is worse in terms of overall reward. Finally, learning progress describes the convergence speed of the learning algorithms, which is critical for burst communication traffics.

We derive the throughput as a function of slot index in Fig.5. The results are averaged from 1500 rounds of independent experiments, where each lasts 6000 time slots. Our experiment settings are as follows. The idle probabilities of independent channels are randomly generated in range $[0, 1]$ for each round. Then, the states of channels (i.e. idle or busy) in each slot are generated independently according to the idle probability vector. Here, $N = 3$ and the normalized per channel sensing cost $\alpha = 0.2$. According to this figure, we conclude that: 1) all the policies that exploit diversity (i.e., sequential sensing/accessing) outperform the policy in the scheme of "one channel per slot", e.g., even *Randomized Sequence* outperforms *Optimal Single Channel* about 15% in our simulation; 2) all the learning policies converge to the optimal solution under either sequential sensing scheme or one channel per slot scheme, which means the learning policies are zero-regret; 3) our proposed SCB policy outperforms all other three online policies in both expected throughput and learning speed: our SCB policy converges to optimal sensing order within almost 400 slots, while all other three policies are still climbing slowly even in 5500 slot; and 4) UCB1-VS outperforms

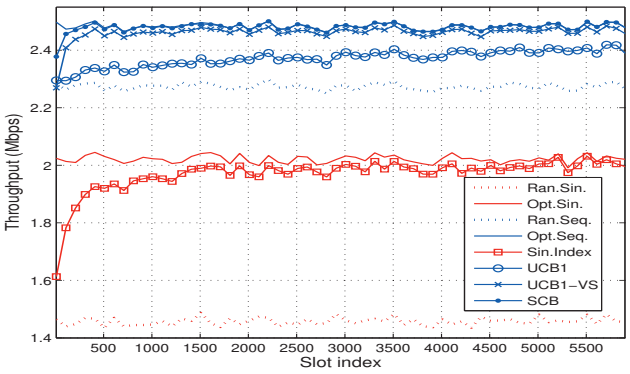



Fig. 5. Throughput comparison

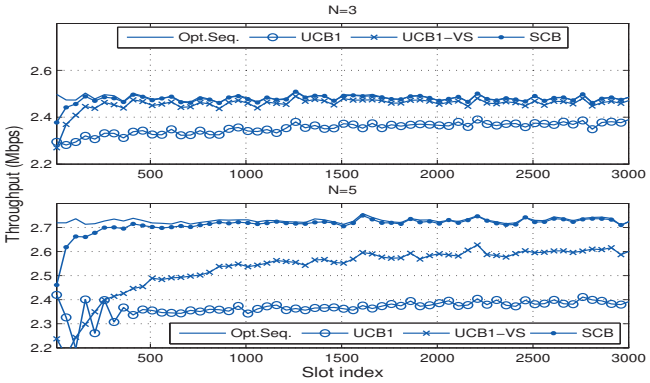



Fig. 6. Comparison on throughput with different $N$

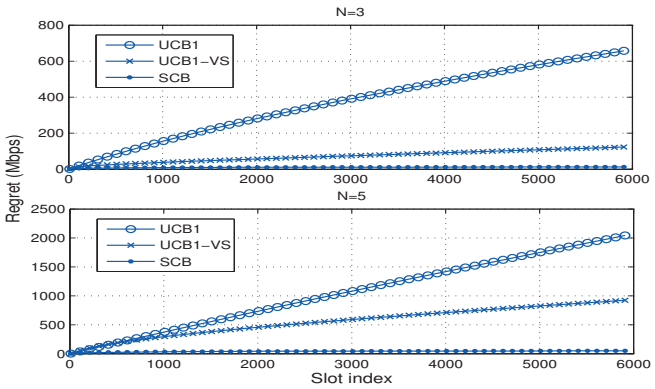



Fig. 7. Comparison on regret with different $N$

UCB1 greatly by virtual sampling, and we will further discuss the benefit of virtual sampling via regret and learning progress analysis.

In Fig.6, we further compare the performance of the sequential learning policies (i.e., UCB1, UCB1-VS and SCB) with different $N$, where the upper sub-figure characterizes the the case $N = 3$ and the lower sub-figure denotes the case $N = 5$. We obtain following observations by comparing these two sub-figures. Firstly, as the number of channels increases, user could obtain more throughput gain through learning. e.g., the achievable throughput increases from about $2.48Mbps$ to $2.72Mbps$ as number of channels increases from 3 to 5. This is reasonable since the potential opportunity increases with the number of channels. Moreover, the curves clearly show that, the learning speed of order-specific algorithms (UCB1 and UCB1-VS) would sharply decreased as $N$ increases, meanwhile, the UCB1-VS greatly accelerates the learning progress over traditional UCB1. These conform to the analysis we stated in Section.3.

Regret is an important metric in evaluating overall system performance of online policies. It has been stated by Lai and Robbins [4] 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. We have proved that the proposed algorithms are order-optimal theoretically in Section.3. In this part, we validate this property by simulations. In Fig.7, we depict the regret of all the three policies, where the upper sub-figure describes the case $N = 3$ and the lower one presents the case $N = 5$. It is intuitive that the regret of all the online policies are increased when increasing the number of channels. This is reasonable, since more channels need more cost for learning. Moreover, it is clearly that SCB outperforms the other two policies significantly in terms of regret. The results also clearly show the benefit of applying virtual sampling: compared with UCB1 policy, the regret (i.e., accumulated throughput loss) of UCB1-VS is apparently reduced. Note that in this figure, the logarithmic increasing trend of UCB1 and UCB1-VS's regret is not evident perhaps due to simulation time limitation, however, the logarithmic increasing property of SCB is fully revealed in 6000 simulation slot (as we would discussed in Fig.8).

To learn more about the increasing regret of SCB policy, we further depict the regret of SCB policy with different value of $N$ particularly in Fig.8. The upper sub-figure clearly shows the logarithmic increasing rate of SCB's regret in time for both $N = 3$ and $N = 5$. To further verify this logarithmic increasing property, we plot out the regret with logarithmic X-axis in the lower sub-figure. It validates the accuracy of our analysis.

We now study the convergence speed of the learning policies in detail. To quantify the convergence speed of an online policy, we propose the concept of *learning progress*. Specifically, "$\sigma$-LP", i.e., $\sigma$-learning-progress ($0 < \sigma < 1$), is defined as the event that "system obtains $\sigma$ times the maximum achievable throughput [3] under the online policies". Mathematically, it can be described as the event that

$$\left\{ \frac{E\left[r_{\pi}(j) - r_{\pi_{rand}}(j)\right]}{E\left[r_{\pi^*}(j) - r_{\pi_{rand}}(j)\right]} = \sigma \right\}$$

where $r_{\pi}$, $r_{rand}$ and $r_{\pi^*}$ are the reward derived by using the online policy $\pi$, randomized policy and the optimal policy respectively. Denote $t_{\sigma}^{\pi}$ as the number of slots user

3. The maximum achievable throughput is achieved by the optimal genie-based solution.

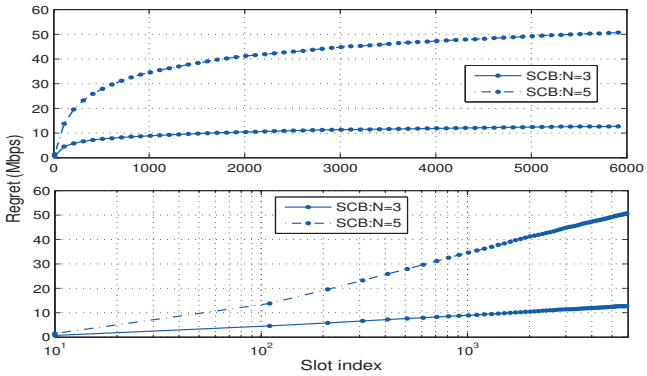



Fig. 8. Regret analysis of SCB

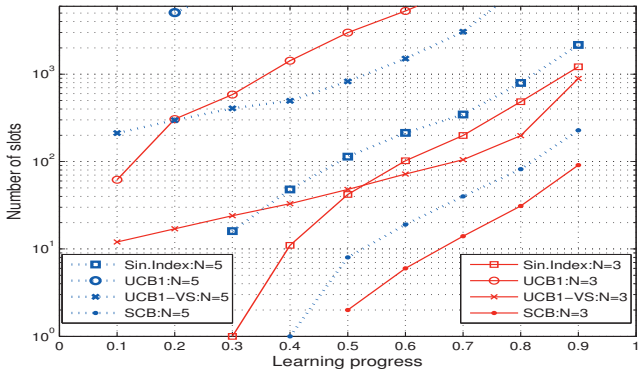



Fig. 9. Analysis of learning progress

experienced before achieving $\sigma$-LP via policy $\pi$. Then, $t_\sigma^\pi$ can describe system learning process well.

In Fig.9, we study $t_\sigma^\pi$ as a function of $\sigma$ in different policies. We analyze all the four online policies. Two points in regarding to the date processing remain to be explained: 1) for fairness consideration, $\pi^*$ is set to be *Optimal Sequence* and $\pi_{rand}$ is *Randomized Sequence* for sequential online policies (i.e., UCB1, UCB1-VS and SCB), while $\pi^*$ is set to be *Optimal Single Channel* and $\pi_{rand}$ is *Randomized Single Channel* for *Single Index*; and 2) to eliminate the singular point caused by randomness, we consider the online policy achieves $\sigma$-LP only when there are 10 continuous slots whose throughput achieve $\sigma$ times the corresponding maximum achievable throughput. The results are shown in Fig.9, where the red-real lines represent the case $N = 3$ and blue-dashed lines represent the case $N = 5$. It tells that:

- All the policies show exponentially increasing $t_\sigma$ with $\sigma$ (note that the $t_\sigma$ is logarithmic value in the figure), i.e., the learning progress increases fast at the beginning, and slower down as learning goes on. This conforms to the laws of learning.
- Virtual sampling greatly accelerates the learning process. As shown in the figure, UCB1-VS outperforms UCB1 19dB in terms of time slots for achieving $0.6-LP$. Note that the learning progress curves on $\sigma > 0.6$ for the case $N = 3$ and $\sigma > 0.1$ for the case $N = 5$ are not depicted here due to space limitation.
- The learning speed of SCB is much faster than all the other policies. As in the case $N = 3$, SCB algorithm outperforms respectively UCB1 and UCB1-VS about 29dB and 10dB in terms of learning speed (quantified by the time slots for achieving 0.9-LP). It even outperforms *Single Index* 12dB, whose total number of learning objects is only $N$.
- Compare the two groups of curves with different number of channels. It shows that, when the number of channels $N$ increases from 3 the 5, the increment of $t_\sigma$ is about 12dB for the order-specific policies (i.e., UCB1 and UCB1-VS), while that for SCB is only 5dB. This indicates that the SCB policy is more tolerant to the increase of channel number.

### 4.2 Impact of Environmental Factors

In this subsection, we evaluate the performance of SCB, *Randomized Sequence* and *Single Index*, with the varying environmental parameters, so as to study the impact of environmental factors. Three metrics are considered for performance comparison: maximum achievable throughput, normalized sensing time and the number of slots to achieve 0.9-LP. The maximum achievable throughput is the expected throughput user could obtain when system converges. Normalized sensing cost is the ratio between the total sensing in a slot and the slot duration. It is adopted to study the energy cost for spectrum sensing in these three policies, as the energy cost for sensing is approximately proportional to the sensing time. Note that for the learning speed, we measure $t_{0.9}^\pi$ here rather than $t_1^\pi$, because the time for achieving 1-LP is usually infinity for online learning policy, and meanwhile, 90% of the maximum throughput is sufficient for most applications.

In Fig.10, we study the first two metrics as functions of channel idle probability, by fixing the number of channels $N = 5$ and normalized per channel sensing cost $\alpha = 0.1$. Two parameters, i.e., average idle probability $\bar{\theta}$ and deviation of channel idle probability $\delta$, are used to control the generation of channel idle probability. At the beginning of each experiment round, channel idle probabilities are randomly generated in the range $\left[\bar{\theta}-\delta, \bar{\theta}+\delta, \right]$. Obviously, a higher $\bar{\theta}$ indicates that there are much more available spectrum resource, while a bigger $\delta$ means more diversity among multiple channels.

The results relating to achievable throughput are shown in the upper part of Fig.10. It is no doubt that: 1) SCB outperforms *Sin.Index* by exploiting instantaneous opportunities among channels and achieves throughput gain over *Rnd.Seq.* with better sensing order; 2) all the three polices would obtain more throughput as the idle probability increases, since higher idle probability indicates better transmission environment. More interesting observations is that, with the decrease of

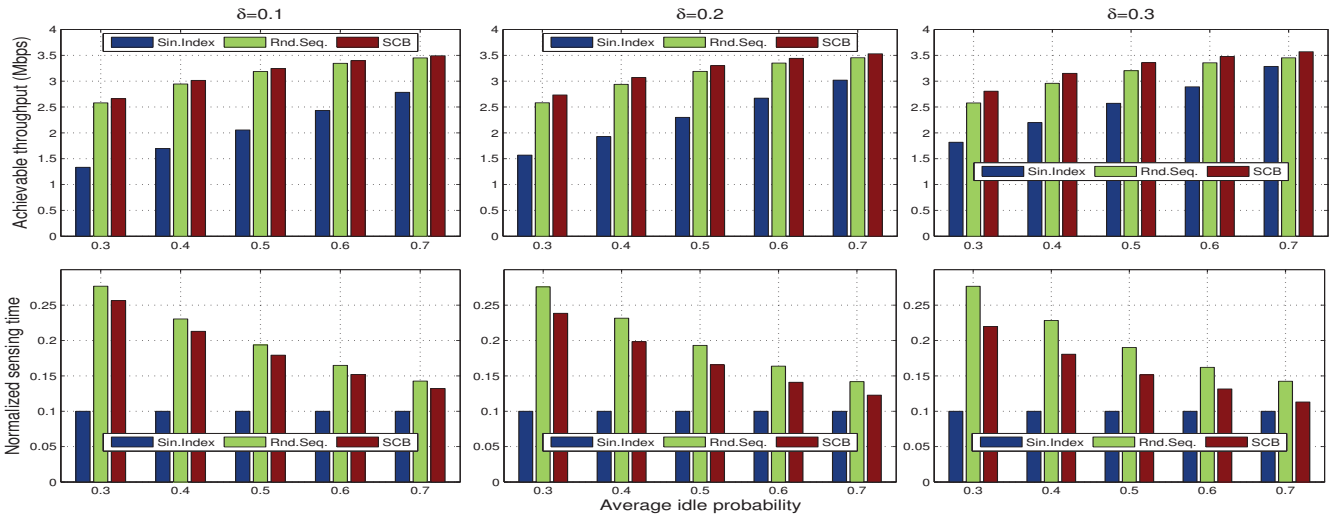



Fig. 10. Achievable throughput & normalized sensing cost vs. idle probability distribution

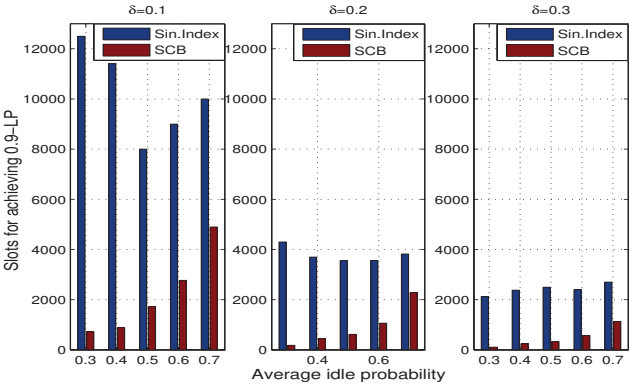



Fig. 11. Convergence vs. idle probability distribution

$\bar{\theta}$, the throughput gains of SCB over the other two policies are increasing. This indicates that SCB would benefit more in the spectrum scarcity scenario, where throughput improvement is most needed. Specifically, it shows nearly two times throughput over *Sin.Index* when $\bar{\theta} = 0.3$. Consider the variation of $\delta$. The throughput gain of SCB over *Sin.Index* deceases as $\delta$ increases, while that of SCB over *Rnd.Seq.* is increasing with $\delta$. As in practical scenarios, $\delta$ is commonly bigger than 0.3. In such case, the SCB achieves more than $10\%$ throughput gain over *Rnd.Seq.* when $\bar{\theta} = 0.3$.

The results relating to normalized sensing cost are shown in the lower part of Fig.10. It is no doubt that the sensing time of *Sin.Index* is constant, as the user always senses one channel per slot in *Sin.Index*. Comparing with *Sin.Index*, SCB and *Rnd.Seq.* lead to more sensing time in each slot. In other words, it tells that the throughput gain of instantaneous opportunity exploitation is achieved at the cost of more energy consumption for spectrum sensing. However, compare the sensing cost of SCB with *Rnd.Seq.*. It shows that SCB achieves more throughput gain with less energy consumption. This clearly shows the benefit of learning.

Further, we study the learning speed (quantified by $t^{\pi}_{0.9}$) of the two learning-based policies under the same experimental settings in Fig.11. It is clearly shown that SCB scheme greatly reduced the time cost in achieving 0.9 learning progress, e.g., less than half even when $\bar{\theta} = 0.7$. In other words, SCB accelerates the learning speed by more than $100\%$ in all cases. Compare with the three sub-figures. We find that the learning speeds of the two policies are strictly increasing with $\delta$. We can explain this phenomenon in such way: as $\delta$ characterizes the deviation of channel statistics, higher $\delta$ indicates bigger difference between channels, and thus is easier to classify them through learning. Meanwhile, it is shown that the learning speed of SCB is increasing with $\bar{\theta}$, due to the fact that less channels would be observed in a slot when $\bar{\theta}$ increases.

In Fig.12 and Fig.13, we explore the impact of $N$ and $\alpha$ on system performance. The results relating to achievable throughput are shown in the upper part of Fig.12. It is no doubt that: 1) the achievable throughput of all the three policies are increasing with $N$; and 2) the achievable throughput decreases as per channel sensing cost $\alpha$ increases. More interesting observations is that, with the increase of $N$, the achievable throughput of all the three policies are increasing sub-linearly. However, the throughput gaps between SCB and the other two policies are increasing with the number of channels. As the number of channels are commonly more than 10 in practical dynamic spectrum access systems, our proposed SCB would outperform *Sin.Index* and *Rnd.Seq.* by more than 30% and 10%, respectively. Consider the variation of $\alpha$. The throughput gain of SCB over *Sin.Index* deceases as $\alpha$ increases, while that of SCB over *Rnd.Seq.* is increasing with $\alpha$, which means online SSA is more sensitive to sensing cost.

The results relating to normalized sensing cost are shown in the lower part of Fig.12. Similar to the results

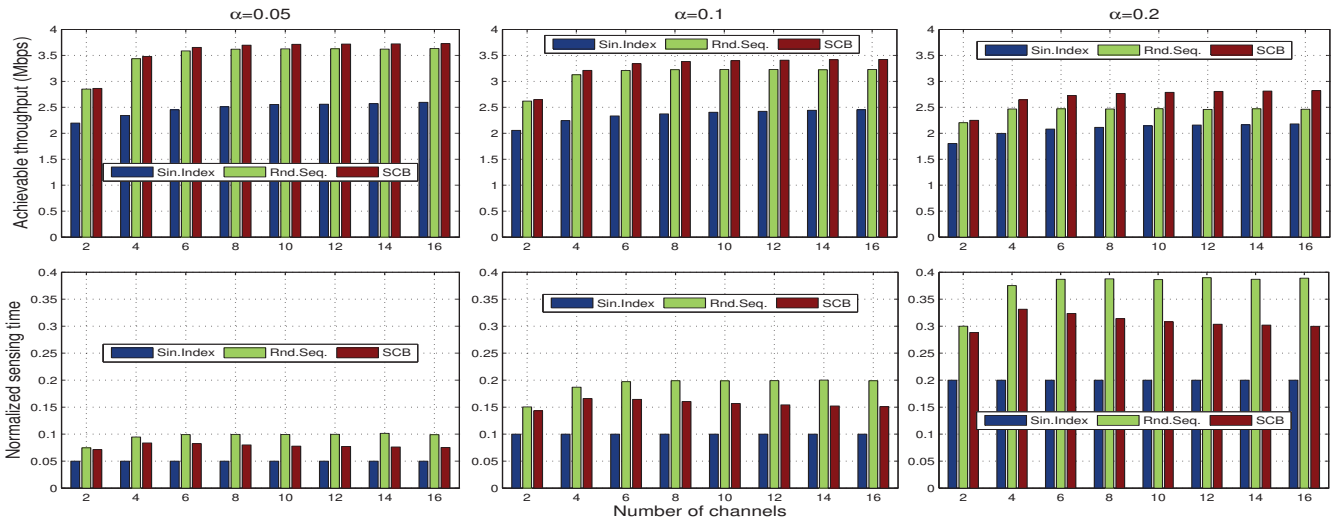



Fig. 12. Achievable throughput & normalized sensing cost vs. number of channels

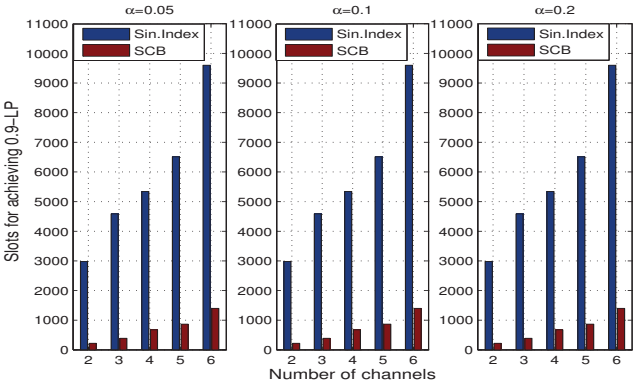



Fig. 13. Convergence vs. number of channels

in Fig.10, the normalized sensing cost of SCB is larger than that of *Sin.Index*, but less than that of *Rnd.Seq.*. Consider the variation of environmental parameters. The total sensing time of all the three policies are linearly increasing with per channel sensing cost $\alpha$. While the relationship between total sensing time and $N$ is much more diverse: 1) the normalized sensing time of *Sin.Index* is constant; 2) with the increase of $N$, the normalized sensing time of *Rnd.Seq.* is increasing sub-linearly when $N < 6$, but goes to be a constant when $N > 6$ (it is approximately double sensing time comparing with *Sin.Index*, as the average idle probability in this settings is 0.5); 3) the normalized sensing time of SCB is increasing with $N$ when $N <\leq 4$ but decreases when $N > 4$, which shows a good scalability with respect to $N$.

In Fig.13, it is clearly show that the learning speed of SCB is much faster than *Sin.Index* (at least 9 times) and the value of normalized per channel sensing cost $\alpha$ has little impact on learning speed of the both policies. Moreover, the learning time of SCB for achieving 0.9-LP is almost linearly increasing with $N$, which indicates SCB is scalable with the increase of $N$ in respect to learning speed. Note that we depict only the cases when $N \leq 6$. This is because the time used by *Sin.Index* for achieving 0.9-LP is too long in $N > 6$. In that case, the corresponding height of SCB bar would be too small (especially at N=2) as we need plot the results of both SCB and *Sin.Index* in a figure for comparison.

## 5 Related Work

In classic multi-armed bandit problem, there are $M$ independent arms, each generating rewards are i.i.d. over time from a given family of distributions with unknown parameters. Lai and Robbins [4] write one of the earliest paper, presenting a general policy that provides expected regret that is $O(M \log L)$. They also show that this policy is order optimal, in which, no policy can do better than $\Omega(M \log L)$. Agrawal et al. [5] present policies easier to compute, leveraging the sample mean with asymptotically logarithmic regret. Further, Auer et al. [6] consider an arbitrary un-parameterized i.i.d arms with non-negative rewards. They provide a simple policy (referred to as UCB1), which achieves logarithmic regret uniformly over time, differs from asymptotical result. However, as we show in this paper, a direct application of the algorithms proposed for classic MAB problem performs poorly, due to the exponential regret value with the increasing number of channels.

Several recent work pay attention to the case where the rewards of different arms are correlated. Mersereau et al. [7] consider a bandit problem where the expected reward is defined as a linear function with random variables. They show the upper bound of the regret is $O(\sqrt{L})$ and the lower bound is $\Omega(\sqrt{L})$. Dani et al. [8] consider linear reward models and presented a randomized policy for the case of a compact set of arms, showing the regret is upper bounded by $O\left(N\sqrt{L}\log^{3/2} L\right)$ for sufficiently large $L$ with high probability, and lower bounded by $\Omega(N\sqrt{L})$. Seminar work of [7] and [8] assume that the

reward is always observable at each time. While in [9], authors present a deterministic policy with a finite time bound of regret, in the order of $\mathcal{O}\left(N^4 \log L\right)$. All those work consider the linear reward model, while we consider non-linear correlated arms in our model.

It should be noted that the parallels between cognitive medium access and the multi-armed bandit problem has been explored in various works. Lai *et al.* [10] firstly apply multi-arm bandit formulations to user-channel selection problems in OSA networks and the UCB1 [6] algorithm is applied. Liu and Zhao [11] formulate the secondary user channel selection to a decentralized multi-armed bandit problem, where contentions among multiple users are considered. Anandkumar in [14] and [15] 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 through channel level feedback. Tekin and Liu [12] model each channel as a restless Markov chain with multiple channel states, and 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.* [13] develop a combinatorial multi-armed bandits (MAB) formulation. An centralized online learning algorithm that achieves $O\left(\log T\right)$ regret uniformly over time is derived. Later, Kalathil *et al.* [21] consider a decentralized setting where there is no dedicated communication channel for coordination among the users. An online index-based distributed learning policy is developed, which achieves the expected regret growing at most as $near-O\left(\log^2 T\right)$. However, all these studies focus on *one channel per slot* scheme and fail to exploit multichannel diversity during the online learning process. This is the key insight where our sequencing multi-armed bandit formulation is novel.

There are also some other papers in the area of optimal control in sequential sensing and accessing systems. Considering i.i.d. Rayleigh fading channels, Sabharwal *et al.* [17] firstly analyze the gains from opportunistic band selection. More generalized scenarios, e.g., with arbitrary number of channels and statistically non-identical channels, are studied in seminar work [19], [22]. In [23], Shu and Krunz consider cognitive radio networks with i.i.d channels, and an infinite-horizon optimal stopping model is leveraged to formulate the online control problem. Jiang *et al.* firstly considered the problem of acquiring the optimal sensing/probing order for a single user case in [20]. Later, Fan *et al.* [24] extends sensing order selection to a two-user case. Zhao *et al.* [25] propose a novel sensing metric to guide the sensing order selection in multiuser case with dynamic programming. Pei *et al.* [26] extend the sequential channel sensing and accessing control to a new scenario, where energy-efficiency is mainly concerned. However, all these work are constructed on perfect knowledge of channel statistics, which differs from the online learning paradigm.

## 6 Conclusion

Sequential channel sensing and accessing (SSA) is an efficient channel utilization scheme for acquiring instantaneous opportunities among multiple channels. In this work, we investigated online learning policy for achieving optimal SSA strategy in unknown environment. A sequencing multi-armed bandit (SMAB) model is proposed for formulating our problem. We first applied the classic UCB1 algorithm and developed an improved version, i.e. UCB1-VS, in solving the SMAB problem. Analysis results show that such UCB1 based policies lead to exponentially increasing regret and complexity. Then, a novel SCB algorithm was proposed, whose storage and computation complexity are linear to the number of channels. We further proved that the expected regret of SCB is upper bounded by $O(NK \log L)$, which is in optimal logarithmic rate in time and polynomial in the number of channels.

## Acknowledgment

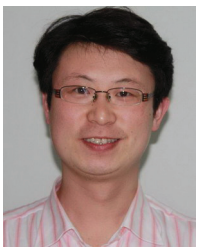
This research is partially supported by NSF China under Grants No. 61003277, 61232018, 61272487, 61170216, 61172062, 61228202, 61273210, NSF CNS-0832120, NSF CNS-1035894, NSF EECS-1247944, China 973 Program under Grants No. 2009CB320400, 2010CB328100, 2010CB334707, 2011CB302705.

## References

[1] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," 2009.
[2] M. Mishra and A. Sahai, "How much white space is there?" University of California, Berkeley, Tech. Rep. Jan. 2009.
[3] "IEEE 802.22-2011(TM) standard for cognitive wireless regional area networks (RAN) for operation in TV bands."
[4] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
[5] R. Agrawal, "Sample mean based index policies with $O\left(\log n\right)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 1, pp. 1054–1078, 1995.
[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, May 2002.
[7] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis, "A structured multiarmed bandit problem and the greedy policy," *Automatic Control, IEEE Transactions on*, vol. 54, no. 12, pp. 2787–2802, Nov. 2009.
[8] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *COLT*, 2008, pp. 355–366.
[9] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. on Networking*, vol. to appear, 2012.
[10] 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, 2011.
[11] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, pp. 5667–5681, 2010.
[12] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *INFOCOM*, 2011.
[13] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *DYSPAN*, 2010.

[14] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *INFOCOM*, 2010, pp. 803–811.
[15] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.
[16] P. Arora, C. Szepesvari, and R. Zheng, "Sequential learning for optimal monitoring of multi-channel wireless networks," in INFOCOM 2011, IEEE. Shanghai, China: IEEE, April 2011.
[17] A. Sabharwal, A. Khoshnevis, and E. Knightly, "Opportunistic spectral usage: bounds and a multi-band CSMA/CA protocol," *IEEE/ACM Trans. Netw.*, vol. 15, pp. 533–545, June 2007.
[18] P. Chaporkar and A. Proutiére, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1546–1555, 2008.
[19] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 17, pp. 1805–1818, 2009.
[20] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 297–307, Jan. 2009.
[21] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized online learning for optimal matching of users to network resources," in *INFOCOM 2012*.
[22] S. Guha, K. Munagala, and S. Sarkar, "Information acquisition and exploitation in multichannel wireless systems," *IEEE Transactions on Information Theory*, 2007.
[23] T. Shu and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," ser. MobiCom '09. NY, USA: ACM, 2009, pp. 37–48.
[24] R. Fan and H. Jiang, "Channel sensing-order setting in cognitive radio networks: A two-user case," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 4997–5008, Nov. 2009.
[25] J. Zhao and X. Wang, "Channel sensing order in multi-user cognitive radio networks," in *IEEE Dynamic Spectrum Access Network (DySPAN)*, 2012.
[26] 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 Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1648–1659, 2011.

**Bowen Li** (S'11) received his B.S. degree in wireless communication and Ph.D degree in communication and information system from Institute of Communication Engineering at the PLA University of Science and Technology, China, in 2007 and 2012 respectively. He is currently a Postdoctoral Fellow in the School of Software and TNLIST, Tsinghua University. His main research interests include stochastic optimization, energy efficient wireless design and backscatter communication networks.



**Panlong Yang** (M'02) received his B.S. degree, M.S. degree, and Ph.D. degree in communication and information system from Nanjing Institute of Communication Engineering, China, in 1999, 2002, and 2005 respectively. Dr. Yang is now an associate professor in the PLA University of Science and Technology. Dr. Yang has published more than 50 papers in the areas of mobile ad hoc networks, wireless mesh networks and wireless sensor networks.

**Jinlong Wang** (M'05) received his B.S. degree in wireless communications, M.S. degree and Ph.D. degree in communications and electronic system from Institute of Communications Engineering, Nanjing, China, in 1983, 1986 and 1992, respectively. He is currently a professor at the PLA University of Science and Technology, China. He is also the co-chairman of IEEE Nanjing Section. His current research interests include wireless communication, cognitive radio and soft-defined radio.

**Qihui Wu** (M'08) received his B.S. degree in communications engineering, M.S. degree and Ph.D. degree in communications and information system from Institute of Communications Engineering, Nanjing, China, in 1994, 1997 and 2000, respectively. He is currently a professor at the PLA University of Science and Technology, China. His current research interests are algorithms and optimization for cognitive wireless networks, soft-defined radio and wireless communication systems.

**Shao-Jie Tang** (S'09) is an Assistant Professor in the Department of Computer and Information Science (Research) at Temple University. He received his Ph.D degree from Department of Computer Science at Illinois Institute of Technology in 2012. He received B.S. in Radio Engineering from Southeast University, P.R. China in 2006. He is a member of IEEE. His main research interests focus on wireless networks (including sensor networks and cognitive radio networks), social networks, pervasive computing, mobile cloud computing and algorithm analysis and design.

**Xiang-Yang Li** (SM'08) received a Bachelor degree at Department of Computer Science and a Bachelor degree at Department of Business Management from Tsinghua University, P.R. China, both in 1995. He received M.S. (2000) and Ph.D. (2001) degree at Department of Computer Science from University of Illinois at Urbana-Champaign. He has been an Associate Professor (since 2006) and Assistant Professor (from 2000 to 2006) of Computer Science at the Illinois Institute of Technology. His research interests include the cyber physical systems, wireless sensor networks, game theory, and algorithms.

**Yunhao Liu** (SM'06) received the B.S. degree in automation from Tsinghua University, Beijing, China, in 1995, and the M.S. and Ph.D. degrees in computer science and engineering from Michigan State University, in 2003 and 2004, respectively. Being a member of Tsinghua National Lab for Information Science and Technology, he holds Tsinghua EMC Chair Professorship. Yunhao is the Director of Key Laboratory for Information System Security, Ministry of Education, and Professor at School of Software, Tsinghua University. His research interests include pervasive computing, peer-to-peer computing, and sensor networks.