Almost Optimal Dynamically-Ordered Multi-Channel Accessing for Cognitive Networks

Bowen Li*, Panlong Yang*, Xiang-Yang Li†, Shaojie Tang†, Yunhao Liu‡, Qihui Wu*
* Institute of Communication Engineering, PLAUST
† Department of Computer Science, Illinois Institute of Technology
‡ Department of Computer Science and Engineering, TsingHua University

***Abstract*—For cognitive wireless networks, one challenge is that the status of the channels' availability and quality is difficult to predict and quantify. 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 ordering. 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 conducted extensive simulations to compare the performance of our method with traditional MAB-based approach. Our simulation results show that our scheme improves the throughput by more than 30% and speed up the learning process by more than 100%.**

I. Introduction

In cognitive radio networks, user is regulated to perform spectrum sensing before it transmits over a channel, so as to protect primary user's communication [1]. Due to hardware limitations, cognitive user can only sense a small portion of the spectrum band at a time[1]. Thus, properly arranging sensing and accessing policy is critical for improving system throughput as well as reducing access delay. A major challenge in achieving optimal opportunistic channel accessing is the difficulty of predicting the channel status and quality accurately. Online learning schemes, due to the adaptivity and efficiency inherently for dynamic wireless network, have received much attention [2].

Assuming cognitive user could only sense/access one channel at each time slot, existing online channel sensing and accessing solutions often model the learning process as a multi-armed bandit (MAB) problem [3]. Although the *one channel per slot* scheme is somehow reasonable in periodical and synchronized spectrum sensing system, it fails to exploit instantaneous opportunities among channels, i.e. multichannel diversity. Such diversity is widespread in dynamic spectrum access system, since the available channels are commonly 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 [4]. Meanwhile, compared with the duration of an access time slot, the channel sensing time is typically very short, e.g., the sensing time is about 10ms, and the access duration is typically about 2s in TV band [5].

Motivated by these facts, we investigate the online channel sensing and accessing schemes, where cognitive user is allowed to sense multiple channels sequentially during each time slot. 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. In this way, both long-term statistics and short-term diversity among different licensed channels can be jointly explored and exploited. Note that in our model, the number of channels being sensed in each time slot is a random variable, while for all the previous work applying MAB in dynamic spectrum access [6]–[8], the number of channels sensed in each time slot is a fixed constant, typically one. This distinguishing feature makes the traditional MAB models cannot be used to solve our problem directly.

In this work, 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. The throughput reward of choosing an arm in a slot is linearly proportional to the remaining transmission time. Observe that the number of arms using this simple mapping 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 challenge in devising an efficient online learning policy, as traditional MAB solutions [9], [10] would result in exponential throughput loss with the increasing number of channels. Moreover, the rewards from different arms are no longer independent in our model, because multiple channels (up to K channels) could be sensed in one time slot. Consequently, previous results under the assumption of independent arms are no longer applicable to our model, which is the second challenge in analyzing the performance of our scheme.

The main contributions of our paper are as follows. Firstly, we apply the classic UCB1 algorithm [10] to handle the online sequential sensing/accessing problem and analyze the regret value, where the regret is the difference between the expected reward gained by a genie-based optimal choice (i.e., always using the optimal sequential sensing order derived with full channel statistics), and the reward obtained by a

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

given policy. We show that both regret and storage overhead are exponentially increasing with the number of channels N. We then propose an improved policy that we refer to as *UCB1 with virtual sampling* (UCB-VS) by considering the dependencies between arms, which significantly improves the convergence of the learning process. Finally, we develop a novel algorithm for such sequencing multi-armed bandit problem, called *sequencing confidence bound* (SCB). We show that the regret is not only logarithmic in time (i.e., order-optimal rate) but also polynomial in the number of channels. Meanwhile, the storage overhead is reduced from $O(N^K)$ to $O(N)$.

The rest of the paper is organized as follows. We present our system model and problem formulation in Section II. Our novel online sequential channel sensing and accessing policy is presented in Section III. Extensive simulation results are reported in Section IV. We conclude our work in Section V.

II. System Model and Problem Formulation

Consider a cognitive radio network with potential channel set $\Omega = \{1, 2, \ldots, N\}$. Each cognitive user is operated in *constant access time* (CAT) mode, i.e., user would have a constant duration T once it obtains a communication chance. We denote the duration of each communication chance as a time slot. 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. We assume channel state is stable during slot time, and independently changes across slots, since the interval time between adjacent communication chances is relatively long in multi-user networks (as discussed in [11]). We consider that the channel idle probability $\theta_i \in [0, 1]$ ($i \in \Omega$) 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 $\theta_{[1]} \geq \theta_{[2]} \geq \ldots \geq \theta_{[N]}$.

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 τ_s. We denote Ψ as the set of all possible sensing orders. Each element in Ψ, that is $\vec{\psi}_m := (s_1^m, s_2^m, \ldots, s_K^m)$, is a permutation of the K channels, where K is the maximum number of steps in each decision slot, and s_k^m denotes the ID of k^{th} channel in $\vec{\psi}_m$. 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 the user stops at step k (i.e, $a_{s_k} = 1$ in current slot), it could obtain an immediate data transmission reward $R(T - k\tau_s)$.

We define the deterministic policy $\pi(j)$ at each time, mapping from the observation history $\mathcal{F}_{j-1}$ to a sequence of channels $\vec{\psi}(j)$ for the j^{th} time slot. The problem is how to make sequential decision on sensing order selection among multiple choices, offering stochastic rewards with unknown distribution. Our main goal is to devise a learning policy maximizing the accumulated throughput, i.e.,

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

where μ is the expected reward in one slot time according to an order $\vec{\psi}$. Let $\alpha = \frac{\tau_s}{T}$. The expected per-slot reward when choosing order $\vec{\psi}_m$ is given by

$$\mu_m = E\left[r_{\vec{\psi}_m}\right] = \sum_{k=1}^{K} \left\{ (1 - k\alpha)\theta_{s_k^m} \prod_{\kappa=1}^{k-1} \left(1 - \theta_{s_\kappa^m}\right) \right\} \quad (1)$$

Here, $r_{\vec{\psi}_m}$ is the normalized immediate reward obtained using order $\vec{\psi}_m$. Without special emphasis, 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 get

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

where $\rho_\pi(L)$ is the *regret* after L slots, which is the difference between the reward with optimal sensing order (obtained by a genie) and the reward achieved by the given policy. $\mu^* = \max_m \{\mu_m\}$ is the expected per slot reward in optimal sensing order.

III. 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 [10], and the second one is *UCB1 with virtual sampling* (UCB1-VS), which is an improved version of UCB1 by exploring the dependency among arms. We analyze the performance of such intuitive methods. Both storage overhead and regret are exponentially increasing with the number of channels N. We then develop a novel algorithm for such SMAB problem, i.e. sequencing confidence bound (SCB), which needs only $O(N)$ in storage overhead. Moreover, we prove that the regret of SCB is $O(NK \log L)$, which is in polynomial order of N and strictly in logarithmic order of time slots.

A. Solutions Based on UCB1

1) Intuitive UCB1 Algorithm: An intuitive approach to solve the sequencing multi-armed bandit problem is to use the UCB1 policy proposed by Auer et al. [10]. In supporting sensing order selection, two variables are used for each candidate order $\vec{\psi}_m$ ($1 \leq m \leq M$): $\hat{\mu}_m(j)$ is the averaged value of all the obtained rewards of sensing/accessing with order $\vec{\psi}_m$ up to slot j, and $n_m(j)$ is the number of times that $\vec{\psi}_m$ has been chosen up to slot j. They are both initialized to zero and updated according to the following rules:

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

UCB1 algorithm

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

Fig. 1. UCB1 algorithm description

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

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

The regret of the UCB1 policy is bounded according to the following theorem.

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

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

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

Proof: See ([10], Theorem 1). ■

According to Equ. (5), we conclude that the regret under UCB1 policy is upper bounded in the order $O(M \log L)$. As $M = \begin{pmatrix} N \\ K \end{pmatrix} K!$, it can be rewritten as $O(N^K \log L)$. Intuitively, although the UCB1 policy achieves zero-regret (i.e., $\lim_{L\to\infty} \frac{\rho_\pi(L)}{L} = 0$), it performs poorly in the sequencing multi-armed bandit problem, especially when the number of channels is large.

2) Improved UCB1-VS Algorithm: As the reward in our sequencing multi-armed bandit problem is order-related, the orders with identical sub-sequence would result in similar rewards. This basic finding provides us an important hint that we could improve learning efficiency by exploring dependency among arms, e.g., obtaining information about multiple arms by playing a single arm.

The UCB1-VS is developed from UCB1, where only the update process is revised with *Virtual Sampling*. Specifically, suppose that a user selects an order $\vec{\psi}_m = (s_1^m, s_2^m, \ldots, s_K^m)$ in a slot and finds that channel s_k^m is idle, then:

- Update statistics of all the sensing order starting with $s_1^m, s_2^m, \ldots, s_k^m$, using reward $1 - k\alpha$;
- Update statistics of all the sensing order starting with s_k^m, using reward $1 - \alpha$.

Moreover, in the special case that $K = N$ (i.e., user is capable of sensing all channels in a slot time) and all channels are sensed to be busy, we can conclude that all sensing orders would lead to zero reward in this slot.

Clearly, with virtual sampling, the learning process could be greatly accelerated while the zero-regret property still holds. As analytical result of the precise regret by this UCB1-VS scheme is hard to achieve, we evaluate its performance via extensive simulations in Sec. IV.

B. A Novel Algorithm for Sequencing Bandit Problem

Although the UCB1 based solutions achieve optimal logarithmic regret over time, they are exponentially increasing with the number of channels. Moreover, as the choices are made according to order-specific statistics, the required storage overhead for supporting decision-making is also exponentially growing with the number of channels. Consequently, when the number of channels for dynamic accessing is large, e.g., more than 50 in the TV band [5], the order-specific methods result in poor performance in regret and unacceptable storage overhead. In this subsection, we propose a novel learning policy for sequencing channel sensing and accessing, in which decisions are made according to channel-related statistics. As a result, the storage overhead is linear with the number of channels. We also proved that the regret of our proposed algorithm is in polynomial order of channels.

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

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

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

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 $\vec{\psi}_m$ with the maximum $SCB_m(j)$, where $SCB_m(j)$ is defined by

$$SCB_m(j) = \sum_{k=1}^{K} \left\{ (1 - k\alpha)\theta_{s_k^m}^u(j) \prod_{\kappa=1}^{k-1} \theta_{s_\kappa^m}^u(j) \right\} \quad (8)$$

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.2.

Note that $\vec{\psi} = \arg\max_{\vec{\psi}_m \in \Psi} SCB_m(j)$ is really simple to achieve in practice. In fact, the channel sequence with descending order of current channel upper confidence bound $\theta_i^u(j)$ will achieve maximum SCB.

SCB algorithm

Initialize: for all $1 \leq i \leq N$: $\hat{\theta}_i = 0$, $n_i^s = 0$; $S_0 = \{1, 2, \ldots, N\}$; $l = 1$, $k = 1$;
while $S \neq \emptyset$ **do**
  Sense random channel $i \in S_0$
  Update $\hat{\theta}_i$, n_i^s accordingly
  $k = k + 1$, $S_0 = S_0 \setminus \{i\}$
  if $a_i^l = 1$ **then**
    $l = l + 1$, $k = 1$; access the idle channel
  else if $k = K + 1$ **then**
    $l = l + 1$, $k = 1$; wait for next slot
  end if
end while
for $j = l$ to L **do**
  Sequentially sensing/accessing with $\vec{\psi}$ where

$$\vec{\psi} = \arg \max_{\vec{\psi}_m \in \Psi} SCB_m(j)$$

  Update $\hat{\theta}_i$, n_i^s accordingly
end for

Fig. 2. SCB algorithm description

2) Analysis of Regret: Traditionally, the regret of a policy 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, our proposed approach requires more finely analysis, since it focuses on the basic elements of each arm (i.e. the sub-sequences in each sensing order). Actually, we analyze the number of times each sub-optimal channel being sensed in each step, and sum up this expectation over all channels and then over all steps. Our analysis provides an upper bound 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

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

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

The detailed proof of this theorem is omitted here due to page limitation. As $N \geq K$ and $K \geq 1$, we conclude that $3N - 2K - 1 \geq 0$. Thus, the right part of regret expression $N - \frac{K+1}{2} - \frac{\alpha(K+1)(3N-2K-1)}{6} < N$. As a result, our policy achieves with a regret upper bounded in the order of $O(NK \log L)$, which is in polynomial order to number of channels and strictly in logarithmic order to time.

IV. 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.

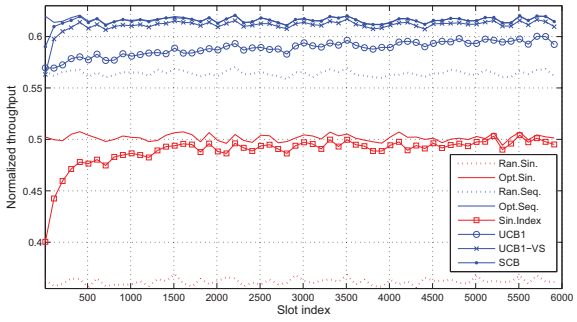

Fig. 3. Learning progress analysis

Eight policies are running under the same environment for performance comparison, where UCB1, UCB1-VS, and SCB are our proposed sequential online learning policies, *Single Index* is an order-optimal *one channel per slot* online learning policy which is first presented by Lai et al in [6]. *Randomized Single Channel* chooses one random channel for sensing/accessing at each slot, and *Optimal Single Channel* is a genie-based policy that user always senses/accesses the channel with highest idle probability in each slot. Correspondingly, user would sequentially sense/access with a randomly chosen sensing order at each slot under *Randomized Sequence*, and would always use the optimal sensing order for sequential sensing/accessing under *Optimal Sequence*.

We derive the normalized throughput as a function of slot index in Fig.3. The results we are averaged from 1500 rounds of independent experiments, where each lasts 6000 time slots. Our experiment setting is 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 of current experiment round. Here, $N = 3$ and the normalized sensing cost $\alpha = 0.2$. It clearly shows 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* that always using the optimal channel; 2) all the learning policies converge to the optimal solution under either sequential sensing scheme or one channel per slot scheme; and 3) our proposed SCB policy outperforms all other three online policies in both expected throughput and learning speed.

In Fig.4, we further compare the performance of the learning policies with different N. Comparing the results in the case that $N = 3$ (i.e., the left part) with that in the case $N = 5$ (i.e., the right part), we obtain following observations. Firstly, as the number of channels increases, user could obtain more throughput gain through learning. This is because 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

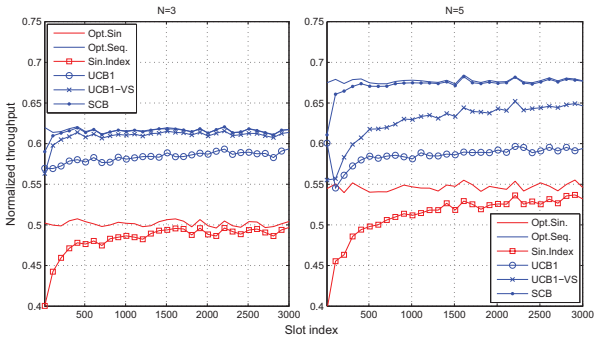

Fig. 4. Throughput reward with different N

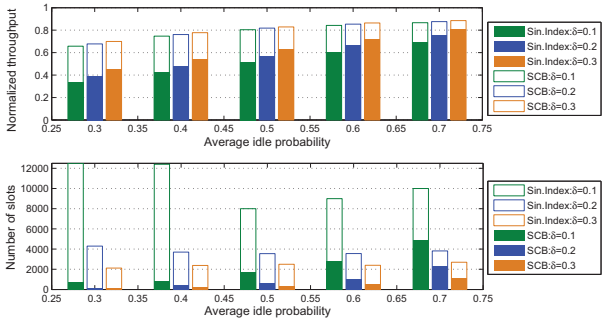

Fig. 5. Impact of idle probability

accelerate the learning progress over traditional UCB1. These conform to the analysis we stated in Section.III.

Finally, in Fig. 5, we study the impact of channel idle probability on the efficiency of learning policies. In this part, $N = 5$ and $\alpha = 0.1$. Two parameters, i.e., $\bar{\theta}$ and δ, are used to control the generation of channel idle probability, where channel idle probabilities are randomly generated in the range $\left[\bar{\theta} - \delta, \bar{\theta} + \delta, \right]$ at the beginning of each round. We compare our proposed SCB policy with existing MAB-based online policy *Single Index* [6] in both normalized throughput and learning speed. It clearly shows in upper part of Fig. 5 that SCB outperforms *Single Index* in all cases by exploiting instantaneous diversity among channels. Meanwhile, the throughput gain decreases as $\bar{\theta}$ and δ increases. This indicates that our scheme would benefit more in the spectrum scarcity scenario, e.g., it shows nearly two times throughput over *Single Index*) when $\bar{\theta} = 0.3$. The averaged throughput gain over all the considered scenarios is more than 30%. Further, we study the learning speed of these two policies in the lower part of this figure. We denote the number of slots user experienced before achieving "σ-learning-progress" ($0 < \sigma < 1$) as t_σ, and use it to quantify the learning speed of the online learning policies. Specifically, t_σ^{scb} and $t_\sigma^{sin.index}$ are defined as $t_\sigma^{scb} \doteq \min_j \left\{ \frac{E[r_{scb}(j) - r_{seq.rand.}(j)]}{E[r_{seq.opt.}(j) - r_{seq.rand.}(j)]} = \sigma \right\}$ and $t_\sigma^{sin.index} \doteq \min_j \left\{ \frac{E[r_{sin.index}(j) - r_{sin.rand.}(j)]}{E[r_{sin.opt.}(j) - r_{sin.rand.}(j)]} = \sigma \right\}$ respectively. We choose a typical value of σ, i.e. $\sigma = 0.9$, to evaluate the learning speed. It is clearly shown that SCB scheme greatly reduced the time cost for achieving 90% learning progress, e.g., less than half even when $\bar{\theta} = 0.7$, which means that SCB accelerates the learning process by more than 100%. The results also show that the learning speeds of the two policies are strictly increasing with δ, where δ characterizes the deviation of channel statistics. Meanwhile, the learning speed of SCB is increasing with θ, perhaps due to the fact that less channels would be observed in a slot when θ increases.

V. Conclusion

In this work, we investigated online learning of optimal sequential channel sensing and accessing. We first introduced the classic UCB1 algorithm in solving our problem. We concluded that using this classic algorithm, both the storage and regret are exponentially increasing with the number of channels. Then, an improved algorithm, i.e. UCB1-VS, was presented, which accelerated learning process by exploring dependency between orders. Finally, we proposed the SCB algorithm with storage overhead linear to the number of channels, and the regret in $O(NK \log L)$.

Acknowledgment

The research of authors is partially supported by NSF CNS-0832120, NSF CNS-1035894, NSF of China under Grant No. 60932002, 61003277, 61170216, 61172062, 973 Program of China under grant No. 2009CB320400, 2010CB328100, 2010CB334707, 2011CB302705, Tsinghua National Laboratory for Information Science and Technology (TNList), program for Zhejiang Provincial Key Innovative Research Team, and program for Zhejiang Provincial Overseas High-Level Talents (One-hundred Talents Program). Jiangsu NSF (Grant No. BK2010102).

References

[1] C. Wang, X.-Y. Li, S. Tang, and C. Jiang, "Multicast capacity scaling laws for multihop cognitive networks," in *IEEE Transactions on Mobile Computing*, Sep. 2011, pp. 262–271.

[2] P. Xu, X.-Y. Li, S. Tang, and J. Zhao, "Efficient and strategyproof spectrum allocations in multichannel wireless networks," *IEEE Trans. Computers*, vol. 60, no. 4, pp. 580–593, 2011.

[3] A. Mahajan and D. Teneketzis, "Multi-armed bandit problems," 2009.

[4] M. Mishra and A. Sahai, "How much white space is there?" EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-3, Jan. 2009.

[5] "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/

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

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

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

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

[10] 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.

[11] Y. Liu, Y. He, M. Li, J. Wang, K. Liu, L. Mo, W. Dong, Z. Yang, M. Xi, J. Zhao, and X.-Y. Li, "Does wireless sensor network scale? a measurement study on greenorbs," in *INFOCOM*, 2011, pp. 873–881.