Channel Exploration and Exploitation with Imperfect Spectrum Sensing for Multiple Users

Zuohong Xu[1], Zhou Zhang[2](✉), Ye Yan[2], and Shilian Wang[1]

[1] National University of Defense Technology, Changsha, China
[2] Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin, China
zt.sy1986@163.com

Abstract. In this paper, the fundamental problem of multiple secondary users (SUs) contending for opportunistic spectrum sensing and access over multiple channels in cognitive radio networks is investigated, when sensing is imperfect and each SU can access up to a limited number of channels at a time. For each channel, the busy/idle state is independent from one slot to another. The availability information of channels is unknown and has to be estimated by SUs during channel sensing and access process. Learning loss, also referred as regret, is thus inevitable. To minimize the loss, we model the channel sensing and access process as a multi-armed bandit problem, and contribute to proposing policies for spectrum sensing and access among multiple SUs under both centralized and distributed framework. Through theoretical analysis, our proposed policies are proved with logarithmic regret asymptotically and in finite time, and their effectiveness is verified by simulations.

Keywords: Multi-user channel sensing and access · Distributed multi-armed bandit problem · Logarithmic regret

1 Introduction

As new wireless devices and applications have been rapidly deployed, the past decade has witnessed a growing demand for wireless radio spectrum resources [1]. However, the traditional static spectrum allocation policy has been reported that most of the licensed spectrum is severely under-utilized [2]. In this regard, the concept of cognitive radio (CR) is proposed and has received great attention to alleviate the spectrum shortage problem due to its great capacity for spectrum exploitation [3,4]. In a CR network, all users are categorized as primary users (PUs) and secondary users (SUs), where PUs have the licence and strict priority to use the channel in frequency and SUs have to explore and exploit channels in an opportunistic manner. In particular, when a SU detects that a PU is occupying a given channel, it releases the channel and switches to another. If no channel is available, a SU waits until a channel is available.

B. Li et al. (Eds.): IoTaaS 2019, LNICST 316, pp. 261–277, 2020.
https://doi.org/10.1007/978-3-030-44751-9_23

However, in a practical network environment, the information of primary channels is usually unknown to SUs, and thus channels have to be fully explored by SUs to learn the information. At the same time, the prior observations could be exploited to gain potential rewards by accessing the sensed idle channels. The requirement for on-line learning this information results in hardness for balancing a fundamental trade-off between channel exploration and exploitation. To measure the trade-off performance, the loss due to on-line learning until time t, also represented by the regret $R(t)$, is defined as the expected difference between the reward of a genie-aided rule with known statistical information of the channels and the actual reward of a specific channel access policy. To well balance the trade-off, the problem of single SU performing opportunistic spectrum access (OSA) has been studied, and formulated as a multi-armed bandit problem (MABP). Extended from those efforts, MABP with multiple players is used to formulate the OSA problem for multiple SUs [5–7]. In such a problem, multiple SUs contend for primary channels access. To minimize the regret and equivalently maximize the average system throughput, design of on-line learning policy is much desired enabling SUs to estimate the channel information and access without collisions.

Several existing studies are seminal for studying our work. Under a centralized framework, the research in [8] proposes a stochastic game framework modelling time-varying process of competition evolution for spectrum opportunities among SUs. In each stage, there is a central spectrum moderator that auctions the available resources for SUs, and a best-response learning algorithm that improves SUs bidding policy is proposed. However, the centralized framework could not be used in cognitive networks when multiple SUs operate autonomously. Motivated by that, under a distributed framework, works in [9–11] have developed algorithms under scenarios of multiple SUs and channels with perfect spectrum sensing. Specifically, in [9] Anandkumar proposes a randomized distributed policy that utilizes the collision feedback under a slotted CR network. It is proved that under any uniformly-good learning and access policy, the proposed policy can achieve order-optimal regret. The total regret is logarithmic with slot time. Liu and Zhao [10] present a time-division fair share (TDFS) policy which yields asymptotically logarithmic regret with respect to slot time. In addition, an index-type policy with coordination mechanism is proposed in [11] which achieves regret of $O(\ln t)$ uniformly over time t. To summarize, all these existing works introduce the distributed framework under perfect sensing, and each SU can sense and access only one channel.

Moreover, some studies take imperfect spectrum sensing into accounts. In works [12,13], a scenario with imperfect sensing and channel access limitations has been investigated. Specifically, author in work [13] models the channel sensing and access process as a bi-level MAB problem, upon which several sensing and access policies are proposed with logarithmic regret asymptotically and in finite time. Additionally, [7] deals with the OSA problem for infrastructure-less CR networks, where multiple SUs collect a priori reward by sensing and accessing one channel at one time. Therein, a policy called QoS-UCB is proposed with at most logarithmic order regret.

Different from existing works, this paper considers the scenario with multiple SUs where each user can sense multiple channels and access only limited channels under imperfect sensing, which has not yet been investigated. To the best of our knowledge, this work is the first research which considers the scenario with several features in a whole view: (i) there are multiple SUs contending for the channels with imperfect channel sensing, if more than one SU access the same channel, collision occurs and no data is successfully transmitted; (ii) each SU senses up to a limited number of channels at one time, and accesses a portion of the sensed channels[1]; (iii) SUs have no knowledge on channel availability and other SUs activities, and no exchange information is assumed among SUs.

The rest of this paper is organized as follows. System model and problem formulation are described in Sect. 2. In Sect. 3, a centralized learning and access policy for multiple SUs is proposed, and in Sect. 4 a distributed learning and access policy for multiple SUs is proposed. Theoretical analysis is made in both sections. Numerical results are then illustrated verifying the performance of our proposed policies.

2 System Model and Problem Formulation

Consider a time slotted system as used in [9,13], where time is partitioned into slots, denoted by T. Let U be the number of SUs and $N \geq U$ be the number of orthogonal licensed channels for PUs. In each channel, e.g. channel i and slot j, PUs are active with probability $1 - \theta_i$, where θ_i represents idle probability. We assume idle probability satisfies $\theta_1 > \theta_2 > ... > \theta_N$, and unknown by SUs. In this work, we define $\mathcal{M}^*$ as a set of channels $\{1, 2, ..., U \cdot M\}$. Time varying channel model is considered, and for channel i, $S_i(j) = 1$ and $S_i(j) = 0$ means the channel idle and busy at slot j, respectively. The state of each channel varies independently from a slot to another. For channel access of a slot, a reward can be obtained by a SU, which is defined as the information bits successfully transmitted in a slot. For simple expression, a reward in each slot is normalized.

In such a system, SUs access the shared spectrum in an opportunistic manner. We briefly introduce the sense and access process as follows. In a time slot, each SU selectively senses M ($M < N$) channels and subsequently access up to K ($K \leq M$) sensed idle channels. Denote sensing results of N channels by SU u in slot j as $\mathbf{X}_u(j) = (X_{u,1}(j), X_{u,2}(j), ..., X_{u,N}(j))$, where $X_i(j) = 1$ and $X_i(j) = 0$ indicate that channel i has been sensed idle and busy at slot j, respectively. Taking sensing errors into accounts, we denote P_d as the detection probability of channel i (i.e., the probability of detecting the PU active if there is PU activity), and P_f as the false-alarm probability of channel i (i.e., the probability of mistakenly estimating that the PU is active when there is no PU activity). For channel i at slot j, the probability that sensed idle is expressed as

[1] In a wireless CR sensor network, sensors usually have the capacity to sense more than one channel at one time; and in view of hardware constraints or limited power supply for wireless devices [16], the number of channels that can be sensed and accessed in each time slot is typically limited.

$f(\theta_i) = (1 - P_f)\theta_i + (1 - P_d)(1 - \theta_i)$. Furthermore, if channel i is sensed idle and accessed at slot j, the conditional reward is obtained from channel access, calculated as $\mathbb{E}[S_i(j)|X_i(j) = 1] = \frac{(1-P_f)\theta_i}{f(\theta_i)}$. $\mathbb{E}[\cdot]$ denotes expectation. In every slot, SU updates observed information for N channels. In particular, for SU u, the information of sensed time is denoted by $\mathbf{T}_u(t) = (T_{u,1}(t), T_{u,2}(t), ..., T_{u,N}(t))$, and reward information is denoted by $\mathbf{Y}_u(t) = (Y_{u,1}(t), Y_{u,2}(t), ..., Y_{u,N}(t))$. More specifically, $T_{u,i}(t)$ represents the number of slots in which channel i has been sensed until slot t by SU u, within it $Y_{u,i}(t)$ represents the number of slots in which channel i has been sensed idle.

As multiple SUs operate in a CR system, collisions happen when more than one SU accesses the same channel simultaneously, resulting in zero reward. For multi-user access model, both centralized and distributed OSA frameworks are investigated. An illustration of sense and access policy under a centralized framework and a distributed framework are shown in Fig. 1. In particular, under a centralized CR framework, at the beginning of slot t, each SU, e.g. SU u selects M channels for sensing, and obtains the sensing results $\mathbf{X}_u(t) = (X_{u,1}(t), X_{u,2}(t), ..., X_{u,N}(t))$. Then SUs report the results to a central agent, which updates the information of $\mathbf{T}_u(t)$ and $\mathbf{Y}_u(t)$. Subsequently, some sensed-idle channels are scheduled by the agent to SUs for channel access. On the other hand, under a distributed learning framework, a central agent does not exist. Each SU records the information $\mathbf{T}_u(t)$ and $\mathbf{Y}_u(t)$, and accesses the sensed idle channels in an autonomous manner. At the end of each slot, each SU receives an acknowledgement (ACK) feedback.

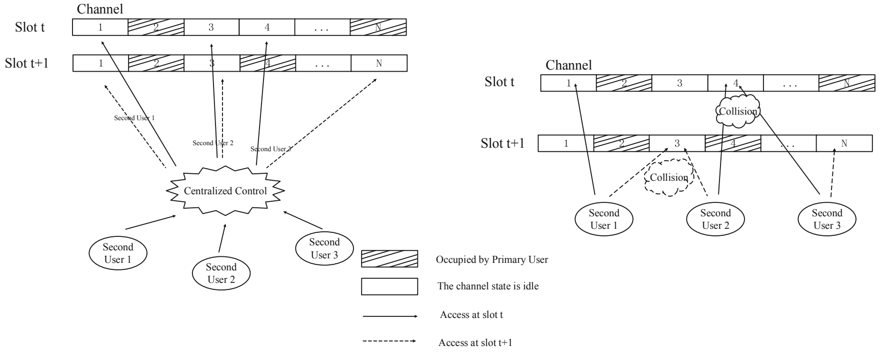

Fig. 1. An illustration of sensing and accessing policies under centralized and distributed frameworks.

Notation: For any two functions $f(n), g(n)$, $f(n) = O(g(n))$ if there exits a constant c such that $f(n) \leq c \cdot g(n)$ for all $n \geq n_0$ for a fixed $n_0 \in \mathbb{N}$, where $\mathbb{N}$ represents natural number set. Moreover, $|\cdot|$ represents the cardinality of a set.

3 Centralized Sensing and Access for Multiple SUs

In this section, we consider the scenario that multiple SUs perform joint channel sensing and access under the control of a central agent. As a fundamental metric to measure the learning policy performance, an optimal channel sensing and access policy under the ideal case where the channel information is known by the central agent, is worthwhile. In reference to the work [15], it is found that the channel set $\mathcal{M}^*$ is the optimal channel set to be sensed in each slot, which maximizes the expected system throughput. We call this sense and access policy a genie-aided rule, and regard the expected reward of this policy as the benchmark. The maximal expected reward is expressed as

$$U^*(t) = \sum_{j=1}^{t} \mathbb{E}\left[\max_{\mathcal{K}(j) \subset \mathcal{I}_{\mathcal{M}^*}(j), |\mathcal{K}(j)| \leq K} \sum_{i \in \mathcal{K}(j)} \mathbb{E}\left[S_i(j) | X_i(j) = 1\right] \right], \tag{1}$$

where $\mathcal{I}_{\mathcal{M}^*}(j)$ denotes the set of sensed idle channels under the set $\mathcal{M}^*$ at slot j, $\mathcal{K}(j)$ denotes the set of channels accessed in slot j. The inner expectation is to calculate the conditional reward of channel i while the outer expectation is for the set $\mathcal{K}(j)$.

3.1 Single Channel Access at a Slot ($\mathbf{K = 1}$)

First, we consider that each SU can simultaneously sense M channels and access only one channel. Under this scenario, we propose a policy, denoted by ρ^{CENT} presented in Algorithm 1. Let ϕ denote the sensing and access policy for all U SUs, and $\phi(j)$ denote the set of channels accessed by SUs at slot j. Compared with the genie-aided rule, the regret of our proposed policy is expressed as

$$R(t, \phi) = U^*(t) - \mathbb{E}\left[\sum_{j=1}^{t} \sum_{i=1}^{N} \mathbb{E}\left[S_i(j) | X_i(j) = 1\right] \cdot \mathbb{I}\left[i \in \phi(j)\right] \right], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function, and when channel i is accessed at slot j, $\mathbb{I}[i \in \phi(j)]$ equals to 1.

Now we analyse the performance of Algorithm 1 in terms of the regret, and prove that the regret is upper-bounded logarithmically over time. Observing that the learning loss occurs when SUs do not choose the optimal channels to access, the regret comprises two components. One is from the case where a SU chooses non-optimal channels to sense and accesses the sensed idle channels. The other is from the case where a SU senses the optimal channels but not accesses the optimal sensed idle channels. By theoretical analysis, the first component is proved having an upper bound logarithmic in time as shown in Lemma 1, and the second component is also similarly bounded as shown in Lemma 2.

Algorithm 1. Single Channel Access under Centralized Learning Framework

```
For time slot $l = 1 : \lceil \frac{N}{UM} \rceil$, SU $u = 1, 2, ..., U$ senses $M$ channels $\{(u-1)M + 1 + (l-1)UM : uM + (l-1)uM \mod N\}$, respectively. All SUs report the sensing result $\mathbf{X}_u(l)$, $u = 1, 2, ..., U$ to the central agent.
The central agent updates $\mathbf{T}_u$ and $\mathbf{Y}_u$, $u = 1, 2, ..., U$, and then randomly selects up to $K$ sensed idle channels for SUs to access.
for each time $t$ do
    The central agent estimates $\theta_i, (i = 1, 2, ..., N)$ by $\hat{\theta}_i(t) = \frac{\frac{Y_i(t-1)}{T_i(t-1)} + P_d - 1}{P_d - P_f}$, sorts channels in descending order according to indexes $\hat{\theta}_i(t) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_i(t-1)}}$, and chooses $M \cdot U$ channels of largest indexes, denoted by $\mathcal{M}(t)$.
    The central agent schedules SU $u$ to sense channels set $\{(u-1)M + 1 + (l-1)UM : uM + (l-1)uM\}$, and then the SU reports the sensing result to the central agent.
    The central agent selects the set of sensed idle channels $\mathcal{I}(t)$, and updates $\mathbf{T}_u(t)$ and $\mathbf{Y}_u(t)$.
    if $|\mathcal{I}(t)| \geq K$ then
        it chooses $K$ largest channels in $\mathcal{I}(t)$, and allocates them for SUs to access.
    else if $0 < |\mathcal{I}(t)| < K$ then
        it allocates channels for SUs $u = 1, 2, ..., |\mathcal{I}(t)|$ to access.
    end if
end for
```

Lemma 1. *For Algorithm 1, the expected number of slots where any channel $i \notin \mathcal{M}^*$ is sensed by SUs until time t has an upper bound derived as*

$$\mathbb{E}[T_i(t)] \leq \frac{8 \ln t}{(\theta_{UM} - \theta_i)^2 (P_d - P_f)^2} + 1 + \frac{MU\pi^2}{3}, \tag{3}$$

where $T_i(t)$ represents the number of slots that channel i has been sensed.

Proof. Recall that idle probability satisfies $\theta_1 > \theta_2 > ... > \theta_N$, for any channel $i \notin \mathcal{M}^*$, we have

$$T_i(t) = 1 + \sum_{j = \lceil \frac{N}{UM} \rceil + 1}^{t} \mathbb{I}\left[i \notin \mathcal{M}^* \right]. \tag{4}$$

Similar to the Appendix D in the work [13], it can conclude that

$$\begin{aligned} & T_i(t) \leq l + \\ & \sum_{k=1}^{UM} \sum_{j=1}^{t} \sum_{t_1=1}^{j} \sum_{t_2=l}^{j} \mathbb{I}\left[\hat{\theta}_k(t_1) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln j}{t_1}} \leq \hat{\theta}_i(t_2) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln j}{t_2}} \right]. \end{aligned} \tag{5}$$

By doing expectation for both sides of (5) and using Chernoff-Hoeffding bound, we obtain that

$$\mathbb{E}\left[T_i(t)\right] \leq \frac{8 \ln t}{(\theta_{UM} - \theta_i)^2 (P_d - P_f)^2} + 1 + \frac{MU\pi^2}{3}. \tag{6}$$

From above, we can see that the expected number of slots that the channels sensed not within the set $\mathcal{M}^*$ grows as $O(\ln t)$ with finite t and $t \to \infty$.

We denote $T'(t)$ as the number of slots where optimal set is sensed but wrong channel is accessed. In the following, we present Lemma 2 which provides an upper bound for the expectation $\mathbb{E}[T'(t)]$.

Lemma 2. *For Algorithm 1, the expectation of $T'(t)$ is bounded by*

$$\mathbb{E}\left[T'(t)\right] \leq \sum_{i=1}^{UM-1} \sum_{k=i+1}^{UM} \left[\frac{8\ln t}{(\theta_i - \theta_k)^2 (P_d - P_f)^2} + 1 + \frac{\pi^2}{3}\right]. \tag{7}$$

Proof. By definition of $T'(t)$, a slot where optimal set is sensed but wrong channel is accessed happens when optimal channel set is sensed, i.e. $\mathcal{M}(j) = \mathcal{M}^*$, but central agent has a wrong top UM-order of the indexes $\hat{\theta}_i(j) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(j-1)}{T_i(j-1)}}$. Under such circumstance, the estimated order of indexes $\hat{\theta}_i(t) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_i(t-1)}}$ for the first $U \cdot M$ channels is not correct. Therefore, it suffices to analyse the bound for the event where any two channels, e.g. channel i and channel k with $i < k, i, k \in \mathcal{M}^*$ are estimated in wrong order.

Recall that $\theta_i > \theta_k$. The event happens when $\hat{\theta}_i(t) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_i(t-1)}} < \hat{\theta}_k(t) + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_k(t-1)}}$. Similar to the proof in Theorem 1, the result is derived.

In accordance with two lemmas above, two components contributing to the regret are with upper bound logarithmic over time. And apparently we can conclude the regret bound in the following theorem.

Theorem 1. *The regret $R(t)$ of Algorithm ρ^{CENT} satisfies $O(\ln t)$.*

Proof. For the centralized scenario, regret has a bound by the sum of two components as shown below.

$$R(t) \leq \triangle \sum_{i=UM+1}^{N} \mathbb{E}\left[T_i(t)\right] + \triangle \mathbb{E}\left[T^{'}(t)\right], \tag{8}$$

where $\Delta \triangleq \mathbb{E}\left[\max_{i \in \mathcal{M}^*} \frac{\theta_i(1-P_f)}{f(\theta_i)} X_i(j)\right]$ is the bound for expected reward loss in each slot.

In particular, $\sum_{i=UM+1}^{N} \mathbb{E}\left[T_i(t)\right]$ represents the number of slots where SUs do not choose the optimal channels to sense while $\mathbb{E}\left[T'(t)\right]$ represents the number of slots that SUs sense the optimal channels but access non-optimal sensed idle channels. Combining results from (3) and (7), the regret satisfies the following inequality

$$
\begin{aligned}
R(t) \leq & \triangle \ln t \sum_{i=UM+1}^{N} \frac{8}{\left(\theta_{UM}-\theta_i\right)^2 \left(P_d-P_f\right)^2} \\
& + \triangle \ln t \sum_{i}^{UM-1} \sum_{k=i+1}^{UM} \frac{8}{\left(\theta_i-\theta_k\right)^2 \left(P_d-P_f\right)^2} \\
& + \triangle (N-UM) \left(\frac{UM\pi^2}{3}+1\right) + \triangle \binom{UM}{2} \left(\frac{\pi^2}{3}+1\right).
\end{aligned} \tag{9}
$$

By definition of $O(\ln t)$ in the notation above, the conclusion is derived.

3.2 Multiple Channel Access at a Slot ($K > 1$)

Then we consider the case where SUs simultaneously access multiple channels at a slot. After SUs report the sensing result to central agent, the central agent schedules the sensed idle channels. If the number of sensed idle channels is less than $U \cdot K$, all sensed idle channels will be allocated to SUs; otherwise, $U \cdot K$ channels with best expected reward are selected for SUs. Under this case, the expected reward of the genie-aided rule can be calculated in (1), while the regret is derived in (2). To design a policy, Line 7 to Line 11 of Algorithm 1 should be modified as follows: if $|\mathcal{I}(t)| \geq U \cdot K$, within set $\mathcal{I}(t)$ central agent schedules $U \cdot K$ channels with best expected reward for SUs to access; otherwise, central agent allocates all channels in $\mathcal{I}(t)$ to SUs. In an extreme case where all channels are sensed busy, no channel is accessed. In this scenario, the regret $R(t)$ of Algorithm ρ^{CENT} satisfies $O(\ln t)$, and the proof process is similar to Theorem 1.

4 Distributed Sensing and Access for Multiple SUs

Different from centralized framework, in this section, we consider a distributed framework where no information exchange or prior agreement is assumed among multiple SUs. Two challenges are thus faced. On one hand, sensing results cannot be shared by SUs, resulting in a slow convergence of estimation process in respect to channel information. On the other hand, multiple SUs accessing the same channel in one slot causes transmission collisions, resulting in additional throughput loss. To overcome these challenges, a distributed sensing and access policy with minimal regret is desired.

As follow, we propose a distributed policy by which each SU selects the channel set for sensing in a randomized manner, driven by collision feedback after channel access. In particular, different from the centralized case, SUs randomly choose one channel set for sensing, then choose sensed idle channels for access in a slot. At the end of this slot, SUs will receive a collision feedback indicating whether collisions happen. Only those SUs who receive collision feedback will randomize the channel set for sensing and access in the next slot. For simplicity, we name the proposed distributed policy $\rho^{\mathrm{RANDOMIZE}}$.

4.1 Single Channel Access ($K = 1$)

First, we consider the genie-aided policy for optimal channel sensing and access when channel information is known by SUs. Before introducing our policies, it is necessary to analyse the benchmark, i.e., the optimal expected reward of genie-aided rule. Different from the benchmark of centralized cases, SUs make decisions based on their own sensing observations. The expected reward of distributed genie-aided rule can be given as

$$\begin{aligned} U^*(t) = &\sum_{j=1}^{t} \max_{\mathcal{M}_u(j), u=1,\ldots,U} \sum_{u=1}^{U} \\ &\mathbb{E}\left[\max_{\mathcal{K}_u(j) \subset \mathcal{I}_{\mathcal{M}_u(j)}, |\mathcal{K}_u(j)| \leq K} \sum_{i \in \mathcal{K}_u(j)} \mathbb{E}\left[S_i(j) | X_i(j) = 1\right] \right], \end{aligned} \tag{10}$$

where $\mathcal{M}_u(j)$ denotes the set of channels sensed by SU u at slot j, $\mathcal{I}_{\mathcal{M}_u(j)}$ denotes the set of sensed idle channels in $\mathcal{M}_u(j)$ at SU u in slot j, $\mathcal{K}_u(j)$ denotes the channel set that SU u accesses in slot j.

Then, we consider how to get the best reward by genie-aided rule in the following lemma.

Lemma 3. *When SUs sense the channel set* $\{\mathcal{M}_u^*\}_{u=1,2\ldots,U}$*, where* $\mathcal{M}_u^* = \{u, M(U-u)+(u+1), \ldots, M(U-u)+(u+M-1)\}$*, and access the sensed idle channel of the smallest index, the maximal reward* $U^*(t)$ *is achieved.*

Proof. Recall that $K = 1$, the expected reward in (10) can be rewritten as

$$U^*(t) = \sum_{j=1}^{t} \max_{u=1,\ldots,U} \sum_{u=1}^{U} W_M\left(\theta_{u,1}, \theta_{u,2}, \ldots, \theta_{u,M}\right) \cdot (1 - P_f), \tag{11}$$

where $W_n(x_1, \cdots, x_n) = x_1 + (1 - f(x_1))\, x_2 + \cdots + \prod_{i=1}^{n-1} (1 - f(x_i))\, x_n$. It can be proved that the function $W_n(x_1, x_2, \cdots, x_n)$ is an increasing function of variables $(x_1, x_2, \cdots, x_n)$.

To maximize the expected reward $U^*(t)$, the problem is transformed into how to allocate $U \cdot M$ channels to each SU. Assume that channel 1 is allocated to SU 1, Eq. (11) is written as

$$\begin{aligned} U^*(t)/t = \theta_1 + &\max_{u=1,\ldots,U} \\ &\left((1 - f(\theta_1))\, W_{M-1}(\theta_{1,2}, \cdots, \theta_{1,M}) + \sum_{u=2}^{U} W_M(\theta_{u,1}, \cdots, \theta_{u,M}) \right). \end{aligned} \tag{12}$$

Subsequently, we further consider how to allocate remaining channels to SU 1 and other SUs. Notably, any $i \geq 2$, we have $1 - f(\theta_1) < 1 - f(\theta_2) < 1$. Thus, $U^*(t)/t$ will become larger if optimal channels are allocated to other SUs. In particular, if the channel is allocated to SU 1, the reward $W_{M-1}(\theta_{1,2},...,\theta_{1,M})$ will be discounted by a coefficient $1-f(\theta_1)$, which contributes less to $U^*(t)$). Therefore, it is optimal to allocate channels $\{M(U-1)+2,\cdots,UM\}$ (i.e., those channels with smallest availability probability in $\mathcal{M}^*$) to SU 1 and the other channels $\{2,3,\cdots,M(U-1)+1\}$ to other SUs. By analogy, the optimal allocation rule for all SUs is derived, under which for each SU u, the optimal sensing channel set should be $\{u, M(U-u)+(u+1),\cdots,M(U-u)+(u+M-1)\}$. Here concludes the proof.

In accordance with Lemma 3, the expected reward of genie-aided rule is calculated as

$$U^*(t)=t\cdot\sum_{u=1}^{U}W_M\left(\theta_u,\theta_{(U-u)M+(u+1)},\cdots,\theta_{(U-u)M+(u+M-1)}\right)\cdot(1-P_f), \quad (13)$$

and the regret is derived as

$$R(t,\{\phi_u\}_{u=1,2,\cdots,U})=U^*(t)- \\ \mathbb{E}\left[\sum_{j=1}^{t}\sum_{u=1}^{U}\sum_{i=1}^{N}\mathbb{E}[S_i(j)|X_i(j)=1]\mathbb{I}[\phi_u(j)=i]\prod_{v=1,v\neq u}^{U}\mathbb{I}[\phi_v(j)\neq\phi_u(j)]\right], \quad (14)$$

where $\{\phi_u\}$ represents the access policy and $\phi_u(j)$ represents the channels accessed in slot j at SU u. The term $\mathbb{I}[\phi_v(j)\neq\phi_u(j)]$ means that if there are multiple SUs choosing the same channel, collision happens and no reward is received.

In the following, to minimize the regret, we present our proposed policy in Algorithm 2 and analyse its regret bound. Note that there are multiple SUs contending for channels, collisions exist resulting in regret increase. Through analysis, the regret consists of two parts: one comes from the case where SUs do not sense or access non-optimal sensed idle channels, and the other comes from multi-user collisions.

For the first part contributing to the regret, it contains two situations, denoted by situation 1 and situation 2, respectively. In situation 1, SU u senses or accesses channels not within $\mathcal{M}^*$, while in situation 2, the channels sensed or accessed by SU u within $\mathcal{M}^*$ but not within $\mathcal{M}^*_u$. The reason for situation 2 existing is that SU u has a wrong estimation of optimal channel order, and thus SU is likely to choose wrong channels to sense and access.

Considering situation 1, the expected number of slots $T_{u,i}(t)$ where a SU u senses and accesses non-optimal channel i is derived in Lemma 4.

Lemma 4. *For Algorithm 2, the expected number of slots where any channel $i\notin\mathcal{M}^*$ is sensed by SU u until time t has an upper bound*

$$\mathbb{E}\left[T_{u,i}(t)\right]\leq\frac{8\ln t}{(\theta_{UM}-\theta_i)^2(P_d-P_f)^2}+1+\frac{M\pi^2}{3}. \quad (15)$$

Proof. The proof of inequality (15) is similar to that of Lemma 1.

Algorithm 2. Distributed Learning with Randomization by SU u

```
Set $Flag = 0$. SU $u$ senses all $N$ channels using $\lceil \frac{N}{M} \rceil$ slots. At each slot, it selects “a sensed idle channel to access, and then update $\mathbf{T}_u$ and $\mathbf{Y}_u$.
for at slot $t$ do
    SU $u$ estimates $\theta_i$, $(i = 1, 2, ..., N)$ by $\hat{\theta}_{u,i}(t) = \frac{\frac{Y_{u,i}(t-1)}{T_{u,i}(t-1)} + P_d - 1}{P_d - P_f}$, and then sorts channels in a descending order by indexes $\hat{\theta}_{u,i} + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_{u,i}(t-1)}}$. The order is recorded as $\mathcal{N}_u^{'}$.
    if $Flag = 1$ then
        $Sel_u(t) \Leftarrow Unif(1, 2, ..., U)$.
    else
        $Sel_u(t) \Leftarrow Sel_u(t-1)$.
    end if
    SU $u$ calculates index of sensing channels set as $\mathcal{A}_u(t) = \{Sel_u(t), M(U - Sel_u(t)) + (Sel_u(t) + 1), \cdots, M(U - Sel_u(t)) + (Sel_u(t) + M - 1)\}$
    Update channel sensing set by $\mathcal{M}_u(t) \Leftarrow \mathcal{N}_u^{'}(\mathcal{A}_u(t))$. $Flag \Leftarrow 0$.
    Update $\mathbf{T}_u(t)$ and $\mathbf{Y}_u(t)$, and obtains the set of sensed idle channels $\mathcal{I}_u(t)$.
    if $\mathcal{I}_u(t)$ is non-empty then
        SU $u$ accesses channel $i$ with largest $\hat{\theta}_{u,i} + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_{u,i}(t-1)}}$, and waiting for ACK feedback.
        if SU $u$ receives the ACK then
            $Flag \Leftarrow 0$.
        else
            $Flag \Leftarrow 1$.
        end if
    else
        Do not access any channel at slot $t$.
    end if
end for
```

In reference to Lemma 4, we can derive the bound for the expected time of situation 1 by taking the slots that all SUs sense and access non-optimal channels into consideration. Then, we analyse the expected number of slots for situation 2. Denote $T_u'(t)$ as the expected number of slots where SU u has a wrong estimation of optimal channel order until time t, and the conclusion is derived in Lemma 5.

Lemma 5. *For Algorithm 2, the expected number of slots where a SU u does not estimate the correct order of optimal channels until t has an upper bound*

$$\mathbb{E}\left[T_u'(t)\right] \leq \sum_{i=1}^{UM-1} \sum_{k=i+1}^{UM} \left[\frac{8\ln t}{(\theta_i - \theta_k)^2 (P_d - P_f)^2} + 1 + \frac{\pi^2}{3}\right]. \qquad (16)$$

Proof. The proof of inequality (16) is similar to that of Lemma 2.

In Lemma 5, we derive the bound for the expected number of slots where SU u has a wrong estimation of optimal channel order. Notably, when situation

1 exists, there must be a wrong estimation of optimal channel order at SU u. Therefore, we can derive an upper bound for the expected time of situation 2 by summing up $\mathbb{E}\left[T'_u(t)\right]$ at all SUs.

Subsequently, we analyse the second part contributing to regret, which comes from multi-user collisions. Define $M(t)$ as the number of collisions faced by SUs in optimal channels until time t. The bound for the expectation of $M(t)$ is given in the following lemma.

Lemma 6. *For Algorithm 2, the expectation of $M(t)$ is bounded by*

$$\begin{aligned}\mathbb{E}[M(t)] &\leq \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right] \cdot \binom{2U-1}{U} \\ &\leq U \cdot \sum_{i=1}^{UM-1} \sum_{k=i+1}^{UM} \left[\frac{8\ln t}{(\theta_i-\theta_k)^2(P_d-P_f)^2} + 1 + \frac{\pi^2}{3}\right] \cdot \binom{2U-1}{U}.\end{aligned} \tag{17}$$

Proof. Define *good events* as the events that all SUs have correct order of top $U \cdot M$ channels, while other events are defined as *bad events*. Each event consists of a consecutive time slots. Denote b as the number of *bad events* until t. $B(k)$, $k = 1, 2, ..., b$ represents the *good event* and $B^c(k)$, $k = 1, 2, ..., b$ represents the *bad event* between two adjacent *good events*. Denote $M(B(k))$ and $M(B^c(k))$ as the collision number during the event $B(k)$ and $B^c(k)$, respectively. In each slot, if there are multiple SUs accessing the same channel, collision number will increase 1.

Subsequently, we analyse collision number $M(t)$ until t. In each slot, either a *good event* or a *bad event* happens, so $M(t)$ contains two parts, one is the collisions under *good event*, which can be expressed as $\sum_{k=1}^{b} M(B(k))$, the other one is the collisions under *bad event*, which can be expressed as $\sum_{k=1}^{b} M(B^c(k))$. Therefore, $M(t)$ is written as

$$M(t) = \sum_{k=1}^{b} M(B(k)) + \sum_{k=1}^{b} M(B^c(k)). \tag{18}$$

In a *good event*, collision number is bounded by

$$\sum_{k=1}^{b} \mathbb{E}\left[M(B(k))\right] \overset{(a)}{\leq} \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right] \cdot \left(\binom{2U-1}{U} - 1\right). \tag{19}$$

where inequality (a) comes from the Theorem 3 in work [9], $\binom{2U-1}{U} - 1$ represents the probability of having an orthogonal configuration over optimal channels by all SUs in a slot under the perfect knowledge of channel information.

In a *bad event*, by definition of $T'_u(t)$, we have the following inequality

$$\sum_{k=1}^{b} \mathbb{E}\left[M(B^c(k))\right] \leq \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right]. \tag{20}$$

Concluded from (18) to (20), the expectation of $M(t)$ is bounded by

$$\begin{aligned}\mathbb{E}[M(t)] &= \sum_{k=1}^{b} \mathbb{E}[M(B(k))] + \sum_{k=1}^{b} \mathbb{E}[M(B^c(k))] \\ &\leq \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right] \cdot \binom{2U-1}{U}.\end{aligned} \tag{21}$$

In reference to (16) and (21), (17) is concluded.

In accordance with Lemmas 4, 5 and 6, we derive the regret bound in the following theorem.

Theorem 2. *The regret* $R(t, \{\phi_u\}_{u=1,2,\ldots,U})$ *of Algorithm 2 satisfies* $O(\ln t)$*.*

Proof. Recall that the regret consists of two parts: one comes from the case where SUs do not sense or access non-optimal sensed idle channels, and the other comes from multi-user collisions. The regret resulting from the first case can be concluded from Lemmas 4 and 5, while the regret resulting from the second case can be concluded from Lemma 6. Therefore, the regret is bounded as

$$\begin{aligned}& R(t, \{\phi_u\}) \\ & \leq \eta \cdot \left(\sum_{u=1}^{U} \sum_{i \notin \mathcal{M}^*} \mathbb{E}\left[T_{u,i}(t)\right] + U \cdot K \cdot \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right] \right) + \eta \cdot U \cdot K \cdot \mathbb{E}[M(t)] \\ & \leq \eta \cdot \left(\sum_{u=1}^{U} \sum_{i \notin \mathcal{M}^*} \left[\frac{8 \ln t}{(\theta_{UM} - \theta_i)^2 (P_d - P_f)^2} + 1 + \frac{M\pi^2}{3} \right] \right. \\ & \quad \left. + U^2 \cdot K \cdot \sum_{i=1}^{UM-1} \sum_{k=i+1}^{UM} \left[\frac{8 \ln t}{(\theta_i - \theta_k)^2 (P_d - P_f)^2} + 1 + \frac{\pi^2}{3} \right] \cdot \left(\binom{2U-1}{U} + 1 \right) \right).\end{aligned}$$

where $\eta = \max_{u=1,\cdots,U} \mathbb{E}\left[\max_{i \in \mathcal{M}^*_u} \frac{\theta_i (1-P_f)}{f(\theta_i)} X_i(j) \right]$. The term $\sum_{u=1}^{U} \sum_{i \notin \mathcal{M}^*} \mathbb{E}\left[T_{u,i}(t)\right]$ represents the number of time slots that all SUs sense or access channels not belonging to $\mathcal{M}^*$, while the term $U \cdot K \cdot \sum_{u=1}^{U} \mathbb{E}\left[T'_u(t)\right]$ represents the maximal number that channel order is incorrectly estimated and thus incorrectly accessed by all SUs. $U \cdot K \cdot \mathbb{E}[M(t)]$ describes the worst case that all SUs access the same K channels in a slot, under which all SUs have no reward.

From the expression of $R(t, \{\phi_u\})$, by definition of $O(\ln t)$, the conclusion is derived.

4.2 Multiple Channel Access ($K > 1$)

Then we consider the case where SUs simultaneously sense and access multiple channels at a slot. For each SU, if the number of sensed idle channels is less than K, all sensed idle channels will be accessed; otherwise, K channels with best expected rewards are accessed. Under this case, the expected reward of the genie-aided rule can be calculated in (11), while the expected reward of regret is derived in (14). To design a policy, Line 12 to Line 13 of Algorithm 2 can be modified as follows: if $|\mathcal{I}_u(t)| \geq K$, SU u accesses up to K channels with K-th largest $\hat{\theta}_{u,i} + \frac{1}{P_d - P_f}\sqrt{\frac{2\ln(t-1)}{T_{u,i}(t-1)}}$ in $\mathcal{I}_u(t)$; if $|\mathcal{I}_u(t)| \leq K$, SU u accesses all channels in $\mathcal{I}_u(t)$. Then SU u waits for ACK feedback. For the scenario where SUs simultaneously access multiple channels at a slot, the regret $R(t)$ of Algorithm $\rho^{\text{RANDOMIZE}}$ also satisfies $O(\ln t)$, and the proof process is similar to that of Theorem 2.

5 Numerical Results

In this part, we perform simulations, varying the number of channels and SUs to verify the effectiveness of the proposed algorithms. Consider a cognitive radio network with U SUs and N channels, each SU can sense M channels, and access K channels. We set $P_d = 0.8$ and $P_f = 0.3$. The information of channels is listed in Table 1.

Table 1. Experimental parameters.

N	θ_i
N = 5	(0.5296, 0.4001, 0.9817, 0.1931, 0.2495)
N = 6	(0.1647, 0.7506, 0.4402, 0.9408, 0.8242, 0.6610)
N = 7	(0.8811, 0.5390, 0.3468, 0.9522, 0.7823, 0.0471, 0.7968)
N = 8	(0.6923, 0.5430, 0.3544, 0.8753, 0.5212, 0.6759, 0.8783, 0.9762)
N = 20	(0.0965, 0.1320, 0.9221, 0.9861, 0.5352, 0.0598, 0.2348, 0.3532, 0.8612, 0.0154, 0.0430, 0.1690, 0.6891, 0.7317, 0.6477, 0.4709, 0.5870, 0.2963, 0.7847, 0.1890)

First we perform simulations for ρ^{CENT}, where the results are shown in Fig. 2. We consider various scenes with $U = 2, 3$, $N = 5, 6, 7, 8$ and $K = 1, 2$. Since our proposed algorithm is bounded by $O(\ln t)$, we explore the relationship between time and normalized regret $R(t)/\ln t$. From Fig. 2, it is easy to see that $R(t)/\ln t$ is finitely bounded. Further observation finds that when the number of channels increases, $R(t)/\ln(t)$ becomes larger. This is because as the number of channels increases, the number of non-optimal channels increases, so SUs utilize more time to sense and access non-optimal channels, consequently resulting in larger regret and slower sensing speed. Additionally, as K increases, $R(t)/\ln(t)$ becomes larger. This is because when the estimation of channel order is not correct, SUs are more likely to access the non-optimal channels.

Then, we present simulations for $\rho^{\text{RANDOMIZE}}$. We consider different scenarios with $U = 2, 3$, $N = 5, 6, 7, 8$, $K = 1, 2$. From Fig. 3, it is seen that the

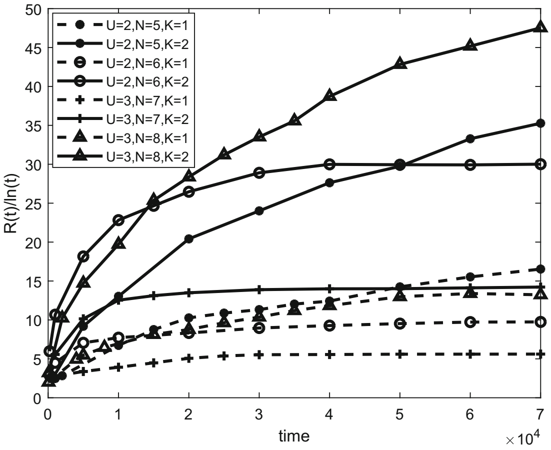

Fig. 2. Average $R(t)/\ln t$ of Algorithm ρ^{CENT}.

normalized regret $R(t)/\ln(t)$ tends to be finitely bounded as time goes, which verifies that the regret is bounded by $O(\ln t)$. From the result we find that there is no transmission loss as time goes infinitely, which means that all SUs will access the optimal channels and converge to a collision-free configuration. Similar to ρ^{CENT}, with the increasing number of channels N, regret increases. As expected from the comparison between Figs. 2 and 3, we can see centralized allocation policy has a lower regret than that of distributed allocation policy.

Further, we increase the number of channels ($N = 20$), and compare the regret with various M, K and fixed U in Fig. 4. It is seen that as time increases, $R(t)/\ln(t)$ is asymptotically limited. With fixed number of SUs and the number

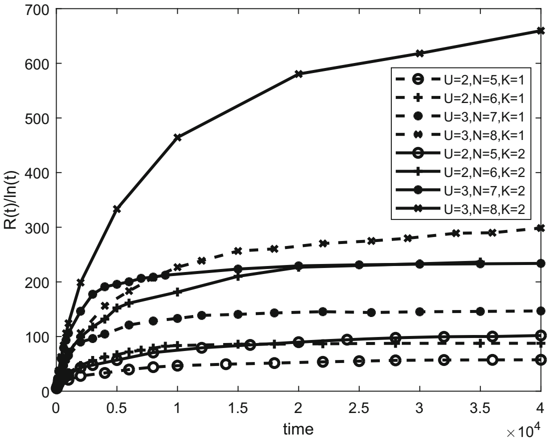

Fig. 3. Average $R(t)/\ln t$ of Algorithm $\rho^{\mathrm{RANDOMIZE}}$.

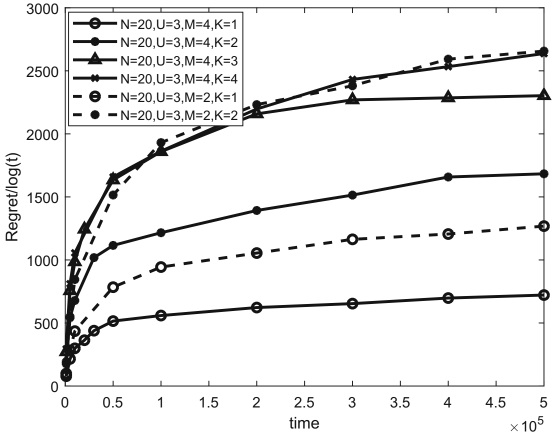

Fig. 4. Average $R(t)/\ln t$ of Algorithm $\rho^{\mathrm{RANDOMIZE}}$ with different M and K.

of channels M, $R(t)/\ln(t)$ increases with increasing number of K. It is explained that before correctly estimating the order of channels, accessing more channels will bring out more collisions and higher probability for non-optimal channels. Another important observation is that with fixed U and K, sensing more channels simultaneously can contribute to lower $R(t)/\ln(t)$. The reason behind is that if SU can sense more channels at one time, it can not only estimate channels more quickly, but has broader chance to get access of sensed idle channels.

6 Conclusion

In this paper, we propose policies for both centralized and distributed learning of channel information for multiple SUs under imperfect sensing in a CR network. Algorithm ρ^{CENT} considers the scenario under the centralized framework while Algorithm $\rho^{\mathrm{RANDOMIZE}}$ adapts the collision feedback to randomize the channel set for SUs under the distributed framework. Both algorithms make SUs converge to a collision-free configuration, ensuring that the regret is logarithmic asymptotically and in finite time. Theoretical analysis and simulations are presented to illustrate the efficiency of proposed algorithms.

References

1. Tanab, M.E., Hamouda, W.: Resource allocation for underlay cognitive radio networks: a survey. IEEE Commun. Surv. Tutor. **19**(2), 1249–1276 (2016)
2. Chen, Y., Oh, H.: A survey of measurement-based spectrum occupancy modeling for cognitive radios. IEEE Commun. Surv. Tutor. **18**(1), 848–859 (2016)
3. Le, T.N., Chin, W.L., Chen, H.H.: Standardization and security for smart grid communications based on cognitive radio technologies–a comprehensive survey. IEEE Commun. Surv. Tutor. **19**(1), 125–166 (2017)

4. Sexton, C., Kaminski, N.J., Marquez, J.M., Marchetti, N., Dasilva, L.A.: 5G: adaptable networks enabled by versatile radio access technologies. IEEE Commun. Surv. Tutor. **19**(2), 688–720 (2017)
5. Rai, V., Diad, I., Tholeti, T., Kalyani, S.: Spectrum access in cognitive radio using a two-stage reinforcement learning approach. IEEE J. Sel. Top. Signal Process. **12**(1), 20–34 (2018)
6. Kumar, R., Darak, S.J., Hanwal, M.K., Sharma, A.K., Tripathis, R.K.: Distributed algorithm for learning to coordinate in infrastructure-less network. IEEE Commun. Lett. **23**(2), 362–365 (2018)
7. Modi, N., Mary, P., Moy, C.: QoS driven channel selection algorithm for cognitive radio network: multi-user multi-armed bandit approach. IEEE Trans. Cogn. Commun. Netw. **3**(1), 49–66 (2017)
8. Fu, F., Schaar, M.C.D.: Learning to compete for resources in wireless stochastic games. IEEE Trans. Veh. Technol. **58**(4), 1904–1919 (2009)
9. Anandkumar, A., Michael, N., Tand, K., Swami, A.: Distributed algorithms for learning and cognitive medium access with logarithmic regret. IEEE J. Sel. Areas Commun. **29**(4), 731–745 (2011)
10. Liu, K., Zhao, Q.: Distributed learning in multi-armed bandit with multiple players. IEEE Trans. Signal Process. **58**, 5667–5681 (2010)
11. Liu, H., Liu, K., Zhao, Q.: Learning in a changing world: restless multi-armed bandit with unknown dynamics. IEEE Trans. Inf. Theory **59**(3), 1902–1916 (2013)
12. Nguyen, T.V., Shin, H., Quek, T.Q.S., Win, M.Z.: Sensing and probing cardinalities for active cognitive radios. IEEE Trans. Signal Process. **60**(4), 1833–1848 (2012)
13. Zhang, Z., Jiang, H., Tan, P., Slevinsky, J.: Channel exploration and exploitation with imperfect spectrum sensing in cognitive radio networks. IEEE J. Sel. Areas Commun. **31**(3), 429–441 (2013)
14. Wang, K., Chen, L., Liu, Q., Wang, W., Li, F.: One step beyond myopic probing policy: a heuristic lookahead policy for multi-channel opportunistic access. IEEE Trans. Wireless Commun. **14**(2), 759–769 (2015)
15. Zhang, Z., Jiang, H.: Cognitive radio with imperfect spectrum sensing: the optimal set of channels to sense. IEEE Wirel. Commun. Lett. **1**(2), 133–136 (2012)
16. Ahmad, A., Ahmad, S., Rehmani, M.H., Hassan, N.U.: A survey on radio resource allocation in cognitive radio sensor networks. IEEE Commun. Surv. Tutor. **17**(2), 888–917 (2015)
17. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Mach. Learn. **47**, 235–256 (2002)