A Distributed Online Learning Method for Opportunistic Channel Access with Multiple Secondary Users

Shuai Ye and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
E-mail: dz20230027@smail.nju.edu.cn, wangsw@nju.edu.cn

ICC 2023 - IEEE International Conference on Communications | 978-1-5386-7462-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC45041.2023.10278615

***Abstract*—In this paper, we investigate the problem of distributed dynamic spectrum access with multiple secondary users (SUs), where the availabilities of licensed channels are unknown to SUs. If multiple SUs access the same channel simultaneously, none of them would transmit successfully because of collision. As a result, each SU has to learn unknown channel statistics and coordinate with others based on its local observations. We develop an online learning based channel access method, which identifies the best channel quickly by using Thompson sampling and orthogonalizes SUs on different channels efficiently without prior information. Numerical results show that the proposed method achieves the highest probability of identifying the best channel compared to the existing ones.**

***Index Terms*—Dynamic spectrum access, multi-armed bandit, online learning.**

I. Introduction

Emerging applications such as the Internet of Things and smart manufacturing require more spectrum bands to support the transmissions of thousands of wireless devices, which consequently leads to spectrum scarcity [1], [2]. On the other hand, actual measurements of spectrum usage reveal that a large portion of the licensed spectrum is underutilized due to static spectrum management policy. In contrast to the static policy, dynamic spectrum access (DSA) allows unlicensed secondary users (SUs), e.g., wireless sensor nodes, to opportunistically access licensed channels if they are not occupied by licensed primary users (PUs) [3]. Therefore, the DSA is considered as a promising solution to overcome the inefficient utilization of spectrum and meet the transmission demands of various wireless devices [4].

The availabilities of licensed channels are unknown to SUs. To avoid causing interference to PUs, each SU has to sense the channels before transmitting on them. SUs access licensed channels only when the sensing results are idle. Due to hardware limitations, each SU can only choose a limited number of channels to sense per time slot [5], [6]. The historical sensing results are utilized to estimate the idle probability of each channel. When multiple SUs access the same channel simultaneously, a collision occurs and none of the SUs transmit successfully at this slot. Thus, an efficient collision avoidance mechanism is needed to coordinate the transmissions of SUs. The channel statistics learning and the coordination among SUs directly impact the spectrum access efficiency and the throughputs of SUs.

This work was partially supported by the National Natural Science Foundation of China under Grants 61931023 and U1936202.

In the literature, such an online channel selection problem falls into an online learning framework, referred to as multi-player multi-armed bandit (MPMAB), where SUs are seen as players facing a bandit machine with multiple arms. Each arm represents a particular licensed channel with unknown channel statistics. At each round, the player would pull an arm, which represents the SU sensing a channel, and receives a reward related to the state of the arm and the actions of other players. Existing DSA schemes can be classified into two classes according to whether there is communication between SUs: centralized methods [7] and distributed methods [8]–[13]. For the centralized methods, SUs report their sensing results to a central controller at each slot. The central controller estimates the idle probability of each channel and decides the sensing and access action of each SU. The centralized methods ensure that no collision occurs, but incur high communication costs. For the distributed methods, each SU estimates the unknown channel statistics and coordinates with others based on its local observations. We focus on the distributed setting since it's more challenging than the centralized setting.

In [8], [9], SUs spend a finite number of slots on randomly sensing licensed channels to estimate the unknown channel idle probability. The total number of slots required for random sensing depend on the idle probability gap between two successive channels. In [10], an index method named upper confidence bound (UCB) is employed to identify the best channel without preliminary knowledge, which efficiently balances the exploitation and exploration of licensed channels. In [14], a modified UCB method is proposed to reduce the number of slots of sensing suboptimal channels, which shows better convergence than the classic UCB.

In [10], [11], a pre-agreement based coordination policy lets SUs sense the estimated best channels in a round-robin fashion to avoid collisions. To get rid of the pre-agreement, each SU senses one of the estimated best channels based on a randomly selected *rank* [12]. In [13], a collision avoidance mechanism inspired by the game 'Musical Chair' is proposed to reduce the number of collisions incurred by the rank policy, which lets the SU continue to sense the same channel once it achieves a

successful transmission on that channel. The above-mentioned policies assume that the number of SUs is known in advance. In a distributed scenario such as ad-hoc network, SUs may not know how many other SUs are also sensing the channels. In [8], the total number of SUs is estimated by using the number of collisions encountered by SUs.

In this paper, we consider the multi-SU DSA problem in a completely distributed scenario where there are no central controller or pre-agreement for SUs. We propose an online learning based channel access method, which utilizes Thompson sampling to identify the best channel without preliminary knowledge and orthogonalizes SUs on different channels without knowing the total number of SUs. Numerical results show that our proposed method outperforms the existing ones in terms of identifying the best channel.

II. System Model

We consider a slotted DSA network with time horizon T. The licensed spectrum is divided into $\mathcal{K} = \{1, 2, \ldots, K\}$ independent channels with equal bandwidth. Let $i_k(t) \in \{0, 1\}$ denote the state of channel k in time slot $t \in [1, T]$, where $i_k(t) = 1$ represents channel k is idle and $i_k(t) = 0$ represents channel k is occupied by PUs. The state of channel k is independently and identically distributed across time with idle probability μ_k, i.e.,

$$i_k(t) = \begin{cases} 1, & \text{with} \quad \mu_k, \\ 0, & \text{with} \quad 1 - \mu_k. \end{cases} \tag{1}$$

Without loss of generality, we assume $\mu_1 > \mu_2 > \ldots > \mu_K$. Let M denote the number of SUs and $M \leq K$ is assumed to avoid congestion. Note that the idle probabilities of the licensed channels are unknown to the SUs.

In any slot t, each SU $m \in [1, M]$ follows a sensing-then-transmission procedure to avoid causing interference to PUs. Specifically, each SU chooses a subset of the licensed channels to sense at the beginning of each slot. If the sensed channels are occupied, the SU does not access and waits for the next slot. If the sensed channels are idle, the SU transmits on the channels and receives an acknowledgment for the transmitted packet at the end of the slot. Due to limited hardware capabilities and power constraints, we assume that each SU m can sense only one channel $a_{m,t} \in \mathcal{K}$ at slot t. There is no prior protocol or information exchange among the SUs. Each SU m estimates the unknown channel idle probability by its local sensing observation $i_{a_{m,t}}(t)$.

A collision occurs when more than one SU accesses the same channel simultaneously and the SUs all fail to transmit data. We denote the collision indicator of SU m at slot t as $\eta_m(t) \in \{0, 1\}$, where $\eta_m(t) = 1$ represents a successful transmission and $\eta_m(t) = 0$ represents that there exists a SU $m' \in [1, M], m' \neq m$ that senses the same channel as SU m. Thus, the multi-SU DSA can be formulated as an MPMAB problem where the reward of SU m at slot t is defined as

$$r_m(t) = i_{a_{m,t}}(t)\eta_m(t), \tag{2}$$

which represents whether SU m transmits successfully.

In an ideal scenario where the channel idle probabilities are known in advance and the M SUs always sense the M channels with the highest idle probabilities without collision, the expected cumulative reward is given by $T\sum_{k=1}^{M}\mu_k$. It's clear that no practical policy can achieve the ideal performance. The regret of a policy is defined as the reward gap between the ideal policy and the considered DSA policy,

$$R(T) = T\sum_{k=1}^{M}\mu_k - \sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[r_m(t)]. \tag{3}$$

Regret $R(T)$ reflects the rate of convergence to the best channels. The goal is to design a policy that minimizes the regret under any given problem instance.

III. Proposed Algorithm

A. Best Channel Identification

Thompson sampling (TS) is an online learning algorithm based on Bayesian theory, which shows empirically better performance than the UCB methods [15]. The basic idea of TS is to assume a prior distribution on the unknown idle probability of each channel and at each slot, select a channel according to its posterior probability of being the best channel. Parameters of the corresponding posterior distribution are updated by using the observed channel state. We improve the classic TS by a top-two rule to identify the best channel $k^* = \text{argmax}_{k\in\mathcal{K}}\mu_k$ from the channel set $\mathcal{K}$.

Each SU m represents its uncertainty about the idle probability of channel k by a prior distribution $f_{m,k} = \text{Beta}(S_{m,k}, F_{m,k})$, the probability density function of which is given by

$$f_{m,k}(\theta_k) = \frac{\Gamma(S_{m,k} + F_{m,k})}{\Gamma(S_{m,k})\Gamma(F_{m,k})}\theta_k^{S_{m,k}-1}(1-\theta_k)^{F_{m,k}-1}, \tag{4}$$

where Γ is the Gamma function, $S_{m,k}$ and $F_{m,k}$ are parameters of Beta distribution. The mean of $f_{m,k}$ is $\frac{S_{m,k}}{S_{m,k}+F_{m,k}}$ and the variance is $\frac{S_{m,k}F_{m,k}}{(S_{m,k}+F_{m,k}+1)(S_{m,k}+F_{m,k})^2}$ [16]. We denote the posterior probability of channel k being optimal as $\alpha_{m,k}$, which can be computed as

$$\alpha_{m,k} = \int_{\theta\in R} f_{m,k}(\theta)\prod_{j\in\mathcal{K}, j\neq k} F_{m,j}(\theta)d\theta, \tag{5}$$

where $F_{m,j}(\theta)$ is the cumulative distribution function of θ_j.

At each slot t, SU m draws a sample $\theta_k(t)$ from the distribution $f_{m,k}$, which is considered as an approximation of the true idle probability μ_k, and senses the channel with the largest sampling value,

$$a_{m,t} = \underset{k\in\mathcal{K}}{\text{argmax}}\,\theta_k(t). \tag{6}$$

Parameters of the distribution $f_{m,a_{m,t}}$ are updated as

$$\begin{aligned} S_{m,a_{m,t}} &= S_{m,a_{m,t}} + i_{a_{m,t}}(t), \\ F_{m,a_{m,t}} &= F_{m,a_{m,t}} + 1 - i_{a_{m,t}}(t). \end{aligned} \tag{7}$$

Note that $S_{m,k}$ represents the total number of slots in which channel k is sensed idle and $F_{m,k}$ represents the total slots in

which channel k is sensed occupied. Thus, according to the law of large numbers, the mean of $f_{m,k}$ converges to the true idle probability μ_k as the number of observations of channel k increases to infinity,

$$\lim_{S_{m,k}+F_{m,k}\to\infty} \frac{S_{m,k}}{S_{m,k}+F_{m,k}} = \mu_k. \tag{8}$$

Also, the variance of $f_{m,k}$ decreases to zero with the increasing number of observations,

$$\begin{aligned} 0 &\leq \lim_{S_{m,k}+F_{m,k}\to\infty} \frac{S_{m,k}F_{m,k}}{(S_{m,k}+F_{m,k}+1)(S_{m,k}+F_{m,k})^2} \\ &\leq \lim_{S_{m,k}+F_{m,k}\to\infty} \frac{1}{4(S_{m,k}+F_{m,k}+1)} = 0. \end{aligned} \tag{9}$$

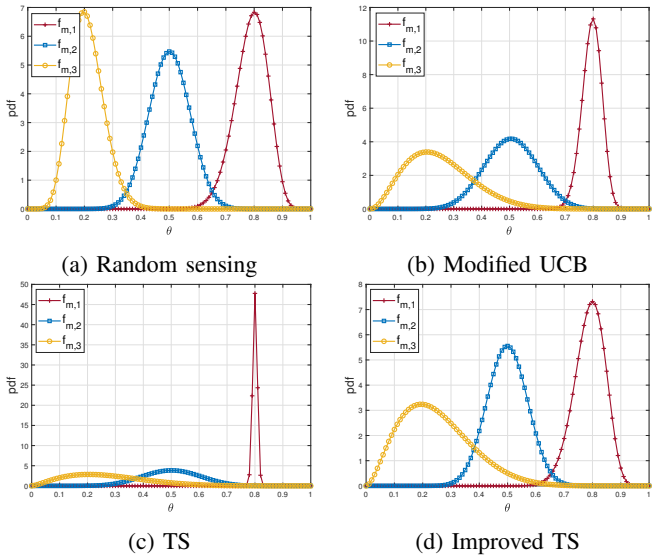
Thus, the true idle probability μ_k can be accurately estimated by the sample $\theta_k(t) \sim f_{m,k}$.

As more observations of channel k are gathered, the distribution $f_{m,k}$ becomes tighter around the true idle probability μ_k. Eq. (5) reveals that the posterior probability of the best channel k^* increases to 1 over time and the posterior probabilities of others decrease to 0. Thus, we set a threshold δ and once there exists a channel $\hat{k}$ whose posterior probability exceeds the threshold δ, i.e., $\alpha_{m,\hat{k}} \geq \delta$, we assume that $\hat{k}$ is the best channel k^*. The probability of identifying the best channel is at least δ with such a criterion, i.e., $\mathbb{P}(\hat{k} = k^*) \geq \delta$.

For the classic TS, once channel $\hat{k}$ is estimated with a reasonably high posterior probability, it will be selected in almost all the subsequent slots according to Eq. (6), resulting in insufficient information collection of other channels and a low convergence rate of $\alpha_{m,\hat{k}}$. For example, if $\alpha_{m,\hat{k}} = 0.95$, then the TS method senses a channel other than $\hat{k}$ roughly once every 20 slots. The distribution $f_{m,\hat{k}}$ is well centralized around the idle probability $\mu_{\hat{k}}$ while the distributions of other channels have large variances. The marginal utility of sensing channel $\hat{k}$ decreases with time.

To overcome the insufficient sensing of other channels, we improve the classic TS by a top-two rule [17]. In each slot t, SU m senses the estimated best channel $I = \text{argmax}_{k\in\mathcal{K}}\, \theta_k(t)$ with probability β. To avoid the algorithm focusing on one channel, with probability $1-\beta$, SU m senses an alternative channel $J = \text{argmax}_{k\in\mathcal{K}, k\neq I}\, \theta_k(t)$, which is posterior optimal except for channel I. On the one hand, the improved TS allocates fraction β of slots to sensing the estimated best channel $\hat{k}$ to guarantee its estimation accuracy. On the other hand, the improved TS allocates more slots to sensing the channels that are difficult to distinguish from $\hat{k}$ and fewer slots to sensing the channels that are clearly inferior, which facilitates the convergence of $\alpha_{m,\hat{k}}$.

We use a simple numerical experiment to better explain the insights of the improved TS. Consider a DSA network with $K = 3$ channels. The corresponding idle probabilities are $0.8, 0.5$ and 0.2, respectively. The prior distribution of channel k is $f_{m,k} = \text{Beta}(1,1), \forall k \in \mathcal{K}$, which is the uniform distribution on $(0,1)$. Fig. 1 shows the posterior distribution of each channel when $\alpha_{m,1} \geq 99.9\%$. The total slots required are $144, 166, 2329$ and 115 for the random sensing, the modified UCB, the TS, and the improved TS, respectively. As is shown in Fig. 1a, the random sensing gives each channel the same chance to be selected, which wastes too many slots on sensing clearly inferior channel 3. The modified UCB and the TS spend almost all slots on sensing channel 1. The insufficient observations of channel 2 and 3 lead to large variances of distribution $f_{m,2}$ and $f_{m,3}$, which consequently results in the low convergence rate of $\alpha_{m,1}$. Note that given the same number of observations, it is more difficult to certify that $\mu_2 < \mu_1$ than that $\mu_3 < \mu_1$, since μ_2 is closer to μ_1 and the observations from channel 2 have higher variance than the observations from channel 3. As is shown in Fig. 1d, the improved TS spends more slots on sensing channel 2 compared to channel 3 since the former is more difficult to distinguish from the best. Thus, the improved TS needs the fewest slots to identify the best channel 1.



(a) Random sensing (b) Modified UCB

(c) TS (d) Improved TS

Fig. 1: Posterior probability density function of each channel

B. Coordination

Once the estimated best channel $\hat{k}$ is determined, the proposed channel access strategy enters a coordination phase, where SU m senses channel $\hat{k}$ at each slot. If SU m achieves a successful transmission during the coordination phase, it senses channel $\hat{k}$ till the end of the time horizon. We call this situation SU m locked on channel $\hat{k}$. We consider two circumstances that SU m may encounter during the coordination phase. One is that channel $\hat{k}$ has already been locked on by some SU. Thus, SU m keeps encountering collisions. It excludes channel $\hat{k}$ from the set $\mathcal{K}$ and continues to identify the next best channel. The other is that SU m is the first one who identifies channel $\hat{k}$ as the best and others are still in the best channel identification procedure. We derive the probability of a SU other than m sensing channel $\hat{k}$ and design the length of the coordination phase accordingly.

Suppose that $\hat{k}$ is also the best channel for a SU $m^{'} \in [1, M], m^{'} \neq m$. During the best channel identification proce-

dure, in any slot t, the probability of SU $m^{'}$ sensing channel $\hat{k}$ is given by [17]

$$\begin{aligned}\mathbb{P}(a_{m^{'},t}=\hat{k}) &= \alpha_{m^{'},\hat{k}}(\beta+(1-\beta)\sum_{k\neq\hat{k}}\frac{\alpha_{m^{'},k}}{1-\alpha_{m^{'},k}}),\\ &\geq \alpha_{m^{'},\hat{k}}(\beta+(1-\beta)\frac{(K-1)(1-\alpha_{m^{'},\hat{k}})}{K+\alpha_{m^{'},\hat{k}}-2}),\end{aligned} \tag{10}$$

where the second inequality is derived by using the Cauchy-Schwarz inequality. The posterior probability $\alpha_{m^{'},\hat{k}}$ increases to δ as the slot goes on. The probability that both SU m and $m^{'}$ identify $\hat{k}$ as the best channel is at least δ^2. Thus, in each slot t, SU m assumes that a SU $m^{'}$ senses channel $\hat{k}$ with probability at least $p_{\min}$, which is given by

$$p_{\min}=\delta^3(\beta+(1-\beta)\frac{(K-1)(1-\delta)}{K+\delta-2}). \tag{11}$$

Denote the total slots of the coordination phase as T_c. The estimated idle probability of channel $\hat{k}$ is denoted as $\hat{\mu}_{\hat{k}}=\frac{S_{m,\hat{k}}}{S_{m,\hat{k}}+F_{m,\hat{k}}}$. $T_c\hat{\mu}_{\hat{k}}$ is the estimated number of idle slots in the coordination phase, i.e., the number of slots SU m accessing channel $\hat{k}$. SU m achieves a successful transmission only when there are no other SUs accessing channel $\hat{k}$ simultaneously. We consider the case that $\hat{k}$ is also the best channel for the other $M-1$ SUs. Thus, with probability at most $(1-p_{\min})^{M-1}$, the $M-1$ SUs don't access channel $\hat{k}$ at each slot. The minimum length of the coordination phase is given by

$$T_c=\lceil\frac{(1-p_{\min})^{1-M}}{\hat{\mu}_{\hat{k}}}\rceil. \tag{12}$$

In a distributed DSA network, the number of SUs M may not known a prior and is estimated by using the number of collisions C_l and the number of idle slots T_l that SU m encountered in the previous best channel identification procedure [8],

$$\hat{M}=\min(\text{round}(\frac{\log(\frac{T_l-C_l}{T_l})}{\log(1-1/K)}+1),K), \tag{13}$$

where $\hat{M}$ is the estimated number of SUs. Once channel $\hat{k}$ is determined, SU m calculates T_c based on $p_{\min}$, $\hat{\mu}_{\hat{k}}$ and $\hat{M}$.

C. TS based Channel Access

The proposed TS based channel access algorithm consists of two phases: best channel identification phase and coordination phase. In the best channel identification phase, each SU m employs the improved TS to decide the sensing channel $a_{m,t}$ in each slot. If the posterior probability $\alpha_{m,\hat{k}}$ of a channel $\hat{k}$ exceeds the threshold δ, SU m records channel $\hat{k}$ as the current best channel and calculates the length of the coordination phase T_c as in (12). In the coordination phase, SU m keeps sensing channel $\hat{k}$. Once SU m achieves a successful transmission, it senses channel $\hat{k}$ till the end. If SU m keeps encountering collisions during the coordination phase. It excludes channel $\hat{k}$ from the channel set $\mathcal{K}$, i.e., $\mathcal{K}\leftarrow\mathcal{K}\setminus\{\hat{k}\}$, and continues to learn the next best channel.

Algorithm 1 TS Based Channel Access Running at SU m

Input $\beta,\delta,p_{\min},\mathcal{K}$

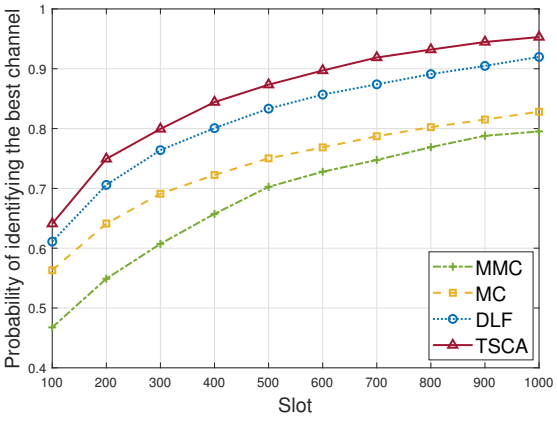
initialize $\hat{k}=0$, $T_l=0,C_l=0,S_{m,k}=1$, $F_{m,k}=1$ $\alpha_{m,k}=\frac{1}{K}$ for $k\in\mathcal{K}$
for $t=1,2,\ldots,T$ **do**
  if $\hat{k}==0$ **then**
    Sample $\theta_k(t)\sim$ Beta$(S_{m,k},F_{m,k}),k\in\mathcal{K}$
    Sample $B\sim$ Bernoulli(β)
    if $B==1$ **then**
      $a_{m,t}=I=\text{argmax}_{k\in\mathcal{K}}\,\theta_k(t)$
    else
      $a_{m,t}=J=\text{argmax}_{k\in\mathcal{K},k\neq I}\,\theta_k(t)$
    end if
    Observe $i_{a_{m,t}}(t)$ and $\eta_m(t)$
    $T_l=T_l+i_{a_{m,t}}(t)$, $C_l=C_l+\eta_m(t)$
    Update $S_{m,a_{m,t}}$ and $F_{m,a_{m,t}}$ as in (7)
    Calculate $\alpha_{m,k}$ for $k\in\mathcal{K}$ as in (5)
    if $\max_{k\in\mathcal{K}}\alpha_{m,k}\geq\delta$ **then**
      Set $\hat{k}=\text{argmax}_{k\in\mathcal{K}}\alpha_{m,k}$
      Calculate $\hat{M}$ and T_c as in (13) and (12)
    end if
  else
    $a_{m,t}=\hat{k}$, $T_c=T_c-1$
    Observe $i_{a_{m,t}}(t)$ and $\eta_m(t)$
    if $\eta_m(t)$==1 **then**
      Lock on channel $\hat{k}$ till the end
    else if $T_c==0$ **then**
      $\mathcal{K}\leftarrow\mathcal{K}\setminus\{\hat{k}\}$
      $T_l=0,C_l=0$, $\hat{k}=0$
    end if
  end if
end for

Details of the proposed TS based channel access algorithm are summarized in Algorithm 1.

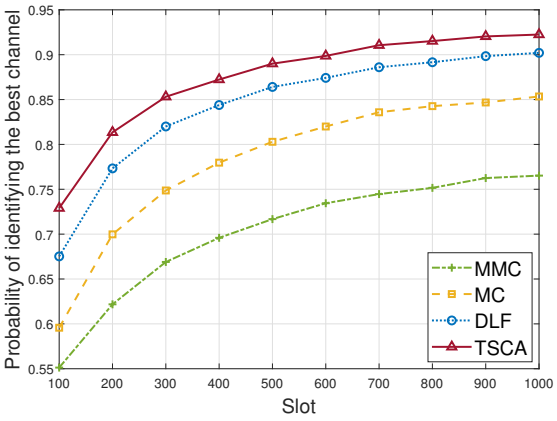
IV. Numerical Results

We present the numerical results in two simulation cases. In case 1, we set the number of channels $K=10$ with the corresponding idle probabilities ranging from 0.08 to 0.85. The idle probability gap between two successive channels is set to 0.07, as required in [8]. In case 2, the idle probability of each channel is sampled uniformly from $(0,1)$. In both cases, there are $M=4$ SUs. All numerical results are averaged over 5000 Monte Carlo simulations.

We compare the proposed TS based channel access (TSCA) algorithm with the distributed learning with fairness (DLF) [10], the Musical Chairs (MC) [8], and the modified MC (MMC) [13] algorithms. The DLF and the MMC algorithms utilize the UCB and the modified UCB for channel statistical learning, respectively. They require the number of SUs M known a prior for coordination. The MC algorithm lets the SUs sense the channels randomly, which does not require prior knowledge of M but requires the idle probability gap between



(a) Case 1



(b) Case 2

Fig. 2: Probability of identifying the best channel as a function of time slot for $T = 1000$.

channels to set the slots of random sensing. For our proposed TSCA algorithm, we set $\beta = 0.2$ and $\delta = 0.85$.

A. Posterior Probability

Fig. 2a shows the probability of identifying the best channel as a function of the time slot in case 1 for $T = 1000$. Specifically, the experiment illustrates the maximum posterior probability $\max_{k \in \mathcal{K}} \alpha_{m,k}$ that each algorithm can achieve within a given number of slots. More observations of channel statistics are collected as the number of slots increases and the idle probability of each channel can be accurately estimated. Thus, the probabilities of all algorithms increase with time. The proposed TSCA algorithm achieves the highest probability of identifying the best channel. The DLF and the MMC algorithms focus on sensing the current best channel at the cost of refining their knowledge of other channels, which reduces their confidence about the best channel. The MC algorithm spends too many slots on sensing clearly inferior channels and thus requires more slots to reach the same probability level as the proposed TSCA algorithm.

Fig. 2b shows the probability of identifying the best channel as a function of the time slot in case 2 for $K = 10$ and $T = 1000$. The idle probability gap in case 2 is smaller than

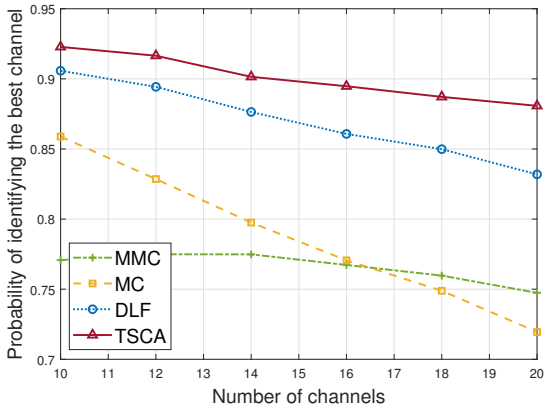
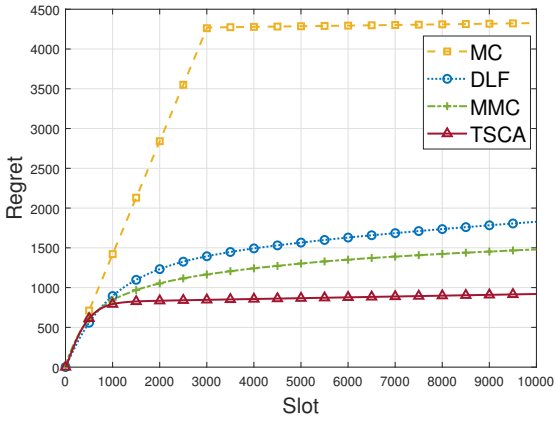
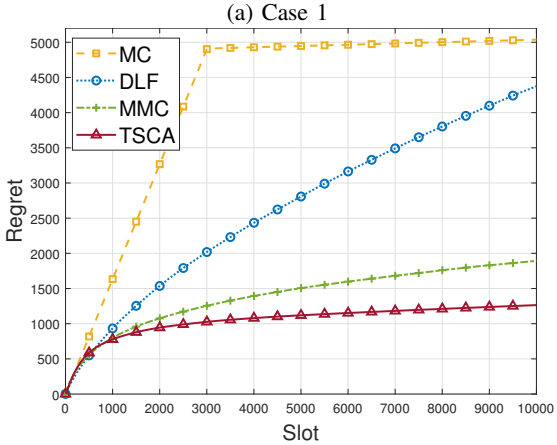


Fig. 3: Probability of identifying the best channel as a function of channel number K for $T = 1000$ in case 2.



(a) Case 1



(b) Case 2

Fig. 4: Regret as a function of time slot for $T = 10000$ and $K = 10$.

that in case 1, resulting in increased difficulty in identifying the best channel. The proposed TSCA algorithm allocates fewer slots to sensing the channels that are far from optimal and more slots to sensing the channels that are harder to distinguish from the best. Thus, the proposed TSCA algorithm needs the fewest slots to reach a given probability level as compared to others.

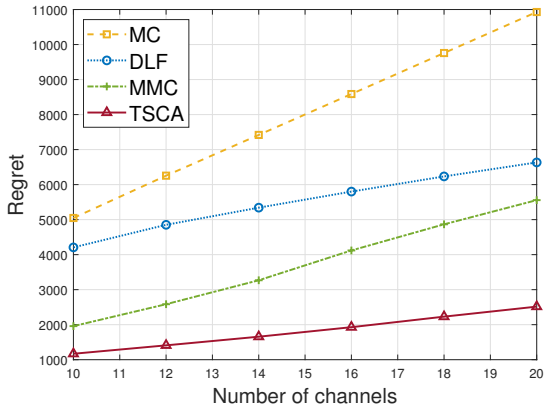


Fig. 5: Regret as a function of channel number K for $T = 10000$ in case 2.

Fig. 3 shows the probability of identifying the best channel as a function of channel number K in case 2 for $T = 1000$. Additional channels increase the exploration cost of each algorithm. The information available per channel decreases as the number of channels. Thus, the probabilities of the algorithms decrease with K. The probability of the MC algorithm decreases more rapidly than the other algorithms since it gives each channel the same number of selections. The proposed TSCA algorithm can fast exclude the clearly inferior channels and shows the highest probability level.

B. Regret

Fig. 4a shows the regret as a function of the time slot in case 1 for $T = 10000$. The proposed TSCA algorithm achieves the lowest regret since it can quickly identify the channels with high idle probability and utilize them for transmission. Due to random sensing, the regret of the MC algorithm grows linearly in the initial learning phase and flattens out in the subsequent coordination phase. Although the DLF and the MMC algorithms can efficiently orthogonalize the SUs on different channels with known M, they require more slots to identify the best channel and thus show higher regrets than the proposed TSCA algorithm.

Fig. 4b shows the regret as a function of the time slot in case 2 for $K = 10$ and $T = 10000$. The trends are similar to that in case 1 while the regret of each algorithm is higher than that in case 1 due to the increased difficulty in identifying the best one. The proposed TSCA algorithm yields the lowest regret since it can orthogonalize the SUs on the identified best channels quickly.

Fig. 5 shows the regret as a function of channel number K in case 2 for $T = 10000$. Due to the additional exploration cost, the algorithms require more slots to identify the best channel. Thus, the regrets of the algorithms increase with K. The proposed TSCA algorithm achieves the lowest regret since it can fast exclude the clearly inferior channels. The MC algorithm senses each channel randomly. The regret of it increases rapidly with the number of channels.

V. CONCLUSION

In this paper, we studied the distributed DSA problem with multiple SUs and proposed the TSCA algorithm to coordinate the transmissions of the SUs. The proposed TSCA algorithm allocates less effort to sensing clearly inferior channels and more effort to sensing channels that are close to the best, which identifies the best channel quickly without prior information. Besides, the TSCA algorithm efficiently orthogonalizes the SUs on different channels without requiring the number of SUs known a prior. Numerical results show that the proposed algorithm outperforms the state-of-the-art ones in terms of identifying the best channel and regret.

REFERENCES

[1] H. Albinsaid *et al.*, "Multi-agent reinforcement learning-based distributed dynamic spectrum access," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1174–1185, Jun. 2022.
[2] Y. Zhang *et al.*, "Asymptotic analysis and precoding design of integrated access and backhaul in full-duplex mmWave networks," *China Commun.*, vol. 19, no. 5, pp. 24–45, May 2022.
[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.
[4] X. Sheng and S. Wang, "Online primary user emulation attacks in cognitive radio networks using Thompson sampling," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8264–8273, Dec. 2021.
[5] A. Magesh and V. Veeravalli, "Decentralized heterogeneous multi-player multi-armed bandits with non-zero rewards on collisions," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2622–2634, Apr. 2022.
[6] J. Dai and S. Wang, "Clustering-based spectrum sharing strategy for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 228–237, Jan. 2017.
[7] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.
[8] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits - a musical chairs approach," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016.
[9] M. Hanawal and S. Darak, "Multiplayer bandits: A trekking approach," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2237–2252, May 2022.
[10] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6184–6193, Dec. 2014.
[11] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.
[12] A. Anandkumar *et al.*, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.
[13] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Proc. Algorithmic Learn. Theory*, Lanzarote, Spain, Apr. 2018.
[14] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1214–1227, Mar. 2015.
[15] M. Zhou, T. Wang, and S. Wang, "Spectrum sensing across multiple service providers: A discounted Thompson sampling method," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2402–2406, Dec. 2019.
[16] Z. Kuai and S. Wang, "Thompson sampling-based antenna selection with partial CSI for TDD massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7533–7546, Dec. 2020.
[17] D. Russo, "Simple Bayesian algorithms for best-arm identification," *Oper. Res.*, vol. 68, no. 6, pp. 1625–1647, Apr. 2020.