Thompson Sampling Based Dynamic Spectrum Access in Non-Stationary Environments

Shuai Ye, Tianyu Wang, *Member, IEEE*, and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—In dynamic spectrum access (DSA), unlicensed secondary users (SUs) estimate the idle probabilities of primary channels by using historical sensing results and opportunistically access the channel with the highest idle probability for transmission. Due to the rapid traffic changes and irregular user mobility, the primary channels can be highly dynamic and their idle probabilities are generally time-varying. In this paper, we investigate DSA in non-stationary environments. Specifically, we consider two channel state models, i.e., the non-stationary Bernoulli model with a time-varying mean and the non-stationary Markovian model with a time-varying transition matrix. For the single-SU scenario, we propose a Thompson sampling based method with a change detection technique, which is capable of detecting the variations of channel statistics and adjusting the channel access strategy accordingly. For the multi-SU scenario, we propose a Thompson sampling based collision alleviation method to coordinate the transmissions of SUs, which does not need any prior protocol or information exchange among SUs. Numerical results show that the proposed methods outperform the existing ones in terms of successful transmission ratio in various network settings.**

***Index Terms*—Dynamic spectrum access, multi-armed bandit, non-stationary environment, Thompson sampling.**

I. Introduction

WORLDWIDE measurements of spectrum usage have revealed that a large portion of the spectrum is underutilized due to static spectrum management policy [1]. While at the same time, there is an increasing demand for spectrum resources due to the emerging wireless applications such as the Internet of Things and virtual reality [2]. Therefore, to improve the efficiency of spectrum utilization, a more flexible spectrum management policy, namely dynamic spectrum access (DSA), has gained a lot of attention.

In DSA, unlicensed secondary users (SUs) can opportunistically access the channels of primary network when licensed primary users (PUs) are absent [3, 4]. Since the availability of primary channels is unknown to SUs, each SU has to sense primary channels before opportunistic access [5–7]. Due to hardware limitations, each SU can only sense a limited number of channels in each time slot [8]. With incomplete spectrum sensing, SUs try to identify the channel with the highest idle probability to acquire the highest chance of opportunistic transmission. However, when SUs are geographically close to each other, multiple SUs may access the same channel at the same slot, which leads to collisions between secondary transmissions. Therefore, an efficient coordination mechanism is needed to avoid collisions between SUs. The channel statistical learning and the coordination among SUs are interrelated with each other, which makes the design of a DSA policy highly challenging.

Much effort has been devoted to DSA, in which a multi-player multi-armed bandit (MPMAB) problem is formulated [9–15]. Specifically, a SU is seen as a player interacting with a bandit machine with multiple arms. Each arm represents a specific primary channel with an unknown idle probability. In each round, the player pulls an arm, which represents spectrum sensing in each transmission slot, and receives a reward that represents the gain from the possible secondary transmission. Thus, the goal is to maximize the cumulative reward of all players within a given number of rounds.

Most existing works consider a stationary environment where the idle probability of each primary channel is time-invariant [16–20]. However, in practical networks, primary channels can be highly dynamic due to the rapid changes of primary traffic and the irregular mobility of wireless users. The channel statistics are generally non-stationary and the corresponding idle probabilities are time-variant. Therefore, these existing DSA algorithms based on stationary assumptions may suffer from severe performance degradation in practical scenarios.

In this paper, we consider a more practical DSA scenario where the idle probabilities of primary channels are not only unknown but also time-varying. Specifically, we formulate a non-stationary MPMAB problem. For the single-SU scenario, we propose a Thompson sampling with change detection (TSCD) method to track the network dynamics without prior knowledge. For the multi-SU scenario, we propose a Thompson sampling based collision alleviation (TSCA) method to reduce collisions between SUs. The contributions of our work are summarized as follows:

- We introduce two non-stationary channel state models, i.e., the non-stationary Bernoulli model and the non-stationary Markovian model, to formulate the non-stationary characteristics in practical DSA.
- We propose a Thompson sampling (TS) based method to estimate the idle probabilities of primary channels, which achieves a promising tradeoff between the exploitation of the current best channel and the exploration of potential better channels.

Manuscript received May 31, 2022; revised December 2, 2022; accepted January 8, 2023. This work was partially supported by the National Natural Science Foundation of China under Grants 61931023 and U1936202. Part of this work has been presented at IEEE Global Communications Conference, Rio de Janeiro, Brazil, December 4-8, 2022. (*Corresponding author: Shaowei Wang.*)

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: DZ20230027@smail.nju.edu.cn; tianyu.alex.wang@nju.edu.cn; wangsw@nju.edu.cn).

TABLE I.
LIST OF MAIN NOTATIONS

$a_{m,t}$	sensing action of SU m at slot t
$A_m(t)$	the set of M best channels of SU m at slot t
$D_{k,w}$	change detection statistic
F_k	number of busy slots of channel k
$J_{m,k}$	number of successful transmissions of SU m in channel k
K	number of channels
$L_{m,k}$	number of collisions of SU m in channel k
M	number of SUs
n_k	number of sensing times of channel k
$p_{k,v}$	idle probability of channel k in segment v
$r_m(t)$	reward of SU m at slot t
$s_k(t)$	state of channel k at slot t
S_k	number of idle slots of channel k
T	time horizon
V	number of piecewise-stationary segments
w	number of recent observations
δ	threshold of the change detection technique
$\Delta_{k,v}$	change of idle probability of channel k in segment v
$\zeta_m(t)$	collision indicator of SU m at slot t
τ	segment length

- We propose a double change detection technique to detect the variations of channel statistics, which does not need prior information and achieves a tradeoff between timeliness and accuracy.
- We propose a distributed method based on TS to coordinate the transmissions of multiple SUs, which requires no dedicated control channels or pre-agreement of SUs.

The rest of the paper is organized as follows. In Section II, we discuss the related work. In Section III, we introduce two non-stationary channel state models and formulate DSA as a non-stationary MPMAB problem. In section IV, we present the proposed TSCD algorithm for the single-SU scenario. In Section V, we extend the TSCD algorithm to the multi-SU scenario by using the TSCA policy to alleviate the collisions between SUs. Numerical results are given in Section VI. Finally, we conclude the paper in Section VII. The main notations used in this paper are summarized in Table I.

II. RELATED WORK

In the Bernoulli model, the state of each channel follows a Bernoulli process with an unknown idle probability [10–16]. In [10], each SU senses primary channels randomly and calculates the channel ranking based on the sensing results. The total number of required sensing slots is shown to be determined by the idle probability gap between channels. In [11], SUs sequentially sense primary channels, considering both the availability and the transmission rate of each primary channel. The total number of required sensing slots is determined by the performance gap between the best and the second best channels. In [12], an upper confidence bound (UCB) method is introduced to DSA to efficiently balance the exploitation of the currently best channel and the exploration of potentially better channels. In [13], a modified UCB method is proposed to reduce the number of suboptimal accesses, which converges to the best channel more quickly as compared to the classical UCB.

In the Markovian model, the state of each channel evolves as a Markov chain [17–20]. In [17], a myopic algorithm achieves the optimal performance when the transition matrix of each channel is known a priori. In [18], the time slots are split into staggered exploration and exploitation epochs. During the exploration epochs, each SU randomly chooses a primary channel to access. During the exploitation epochs, each SU accesses the channel with the largest number of idle observations in the previous exploration epoch. In [19], SUs utilize the classic UCB to choose the sensing channel and stays on the channel until an idle state is observed, which results in a regenerative cycle. Observations inside the regenerative cycle are stored and the rest are discarded.

In the multi-SU scenario, the transmissions of SUs need to be coordinated to avoid transmission collisions. In [14], a pre-agreement based orthogonalization policy ensures that SUs access the estimated best channels in a round-robin fashion, which guarantees fairness between different SUs. In [21], SUs are assigned unique IDs and classified into groups according to a prior protocol for collision-free access. To get rid of the pre-agreement, each SU chooses one of the estimated best channels based on a randomly selected rank [15]. In [16], a coordination policy is proposed to further reduce the collisions of the rank-based policy, which allows SUs to keep sensing the same channel when a successful transmission occurs.

In the study of multi-armed bandit (MAB), non-stationary environments have received extensive attention in recent years. In [22], a discounting factor technique is introduced to the UCB method to weaken the relevance of past observations. In [23], the UCB algorithm is combined with a sliding window technique, where only the observations in the sliding window are taken into account. However, the algorithm cannot converge with a limited window size. In [24], a change detection technique is proposed to detect the changes in the reward distribution of each arm.

Recently, TS based methods have gained a lot of attention in the study of MAB problems [25–29]. They assume a prior distribution for the unknown parameter (e.g., the expected reward) of each arm. At each round, the player pulls an arm according to its posterior probability of being the best arm. Usually, the player can draw a sample from the posterior distribution of each arm and play the arm with the largest sample. Parameters of the posterior distribution are updated by using the obtained reward.

III. SYSTEM MODEL

We consider a slotted DSA network with $\mathcal{K} = \{1, 2, \ldots, K\}$ independent channels and $\mathcal{M} = \{1, 2, \ldots, M\}$ SUs, where $M \leq K$ is ensured to avoid SU congestion. The SUs are clock synchronized and the entire time horizon is given by T. In any slot $t \in [1, T]$, we denote by $s_k(t) \in \{0, 1\}$ the instant state of channel k, where $s_k(t) = 1$ represents the channel is idle and $s_k(t) = 0$ represents the channel is occupied.

In each slot t, each SU $m \in \mathcal{M}$ can sense only one channel $a_{m,t} \in \mathcal{K}$ and observes the corresponding channel

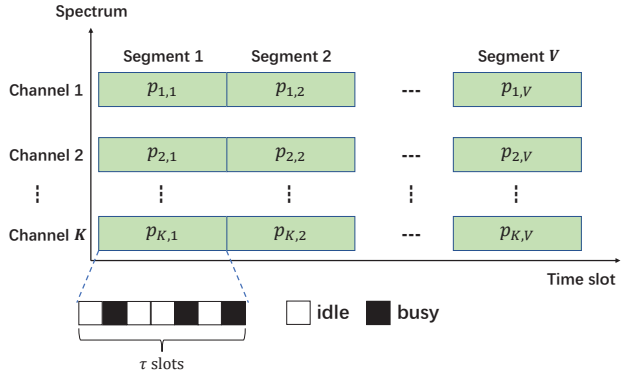


Fig. 1. Non-stationary channel state model.

state $s_{a_{m,t}}(t)$. We assume that the sensing result of each SU is always correct [30]. If $s_{a_{m,t}}(t) = 0$, SU m does not transmit and waits for the next slot. If $s_{a_{m,t}}(t) = 1$, SU m transmits on channel $a_{m,t}$. If no other SUs access channel $a_{m,t}$, SU m receives an ACK signal at the end of the slot. Otherwise, SU m suffers from a collision and receives a NACK signal. We denote by $\zeta_m(t) \in \{0, 1\}$ the collision indicator of SU m at slot t, where $\zeta_m(t) = 0$ represents a collision occurs and $\zeta_m(t) = 1$ represents a successful transmission. Thus, we have

$$\zeta_m(t) = \begin{cases} 0, & \exists m' \neq m : a_{m',t} = a_{m,t} \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

A. Non-Stationary Environment

An illustration of the considered non-stationary channel state model is presented in Fig. 1. In the non-stationary environments, the entire time horizon T is divided into V piecewise-stationary segments with each segment containing $\tau = T/V$ slots. In each segment $v \in [1, V]$, the idle probability of channel k is fixed and denoted by $p_{k,v}$, i.e., $\mathbb{P}[s_k(t) = 1] = p_{k,v}$, $(v-1)\tau < t \leq v\tau$. Note that $p_{k,v}$ and τ are unknown to the SUs.

In the non-stationary Bernoulli model, for any channel k, the channel state follows a Bernoulli process with a time-varying idle probability $p_{k,v}$, i.e,

$$s_k(t) = \begin{cases} 1, & \text{with} \quad p_{k,v} \\ 0, & \text{with} \quad 1 - p_{k,v}, \end{cases} \tag{2}$$

where $(v-1)\tau < t \leq v\tau$. In the non-stationary Markovian model, for any channel k in segment v, the channel state evolves according to a time-varying transition matrix

$$\rho_{k,v} = \begin{bmatrix} \rho_{k,v}^{00} & \rho_{k,v}^{01} \\ \rho_{k,v}^{10} & \rho_{k,v}^{11} \end{bmatrix}, \tag{3}$$

where $\rho_{k,v}^{ij}$ is the transition probability of channel k from state i to j. The idle probability $p_{k,v}$ is given by the stationary distribution with the transition matrix $\rho_{k,v}$, i.e., $p_{k,v} = \rho_{k,v}^{01}/(\rho_{k,v}^{01} + \rho_{k,v}^{10})$.

B. Problem Formulation

We define the reward of SU m at slot t as

$$r_m(t) = s_{a_{m,t}}(t)\zeta_m(t), \tag{4}$$

which represents whether SU m gets a successful transmission at slot t. The successful transmission ratio (STR) of the SUs is given by

$$\text{STR} = \frac{1}{MT}\mathbb{E}[\sum_{t=1}^{T}\sum_{m=1}^{M} r_m(t)]. \tag{5}$$

An ideal DSA policy is to assign the M SUs to sense the M channels with the largest $p_{k,v}$ values in each segment v. However, due to the lack of information about $p_{k,v}$ and τ, no practical policy can achieve the ideal performance. Here, we aim to maximize the STR considering such practical constraints of the SUs.

We note that the considered DSA can be formulated as a non-stationary MPMAB problem. Each SU is a player who pulls an arm $a_{m,t}$ and gets a reward $r_m(t)$ in round t. Regret is a commonly used metric for evaluating the performance of a DSA policy, which is defined as the cumulative reward gap between the ideal policy and the considered policy,

$$R(T) = \tau \sum_{v=1}^{V}\sum_{k\in\mathcal{K}_v^*} p_{k,v} - \mathbb{E}[\sum_{v=1}^{V}\sum_{t=(v-1)\tau+1}^{v\tau}\sum_{m=1}^{M} p_{a_{m,t},v}\zeta_m(t)], \tag{6}$$

where $\mathcal{K}_v^*$ is the set of M channels with the largest $p_{k,v}$ values in segment v. The SUs aim to minimize their regret by choosing their access channels online based on the historical observations.

IV. TSCD for Single SU Scenario

In this section, we first consider a simplified scenario with only one SU, i.e., $M = 1$, where the SU superscript m can be omitted without any confusion. Specifically, we identify the challenges of DSA for the single-SU scenario and propose the TSCD algorithm for the SU to choose the sensing channel a_t in each slot t.

A. Challenges

In the single-SU scenario, the considered problem in (6) is reduced to a single player MAB problem with non-stationary arms. There are two major challenges.

- **Exploitation-Exploration Tradeoff**: In each slot t, the SU is faced with a dilemma between exploitation and exploration, i.e., whether to sense the empirically best channel for the largest immediate reward or to explore other channels to learn their statistics for better transmissions in the future.
- **Non-Stationary Channels**: In the non-stationary environments, the channel statistics vary over time. The SU needs to keep track of the variations of channel statistics so as to maximize its long-term reward by accessing the currently best channel.

B. Change Detection Technique

Change detection (CD) aims to detect the variations of the idle probability $p_{k,v}$ by using a sequence of historical sensing observations. We denote by $\Delta_{k,v}$ the actual change in the idle probability of channel k at the end of segment v, i.e., $\Delta_{k,v} =$

$|p_{k,v+1} - p_{k,v}|$. A successful detection should raise an alarm about the change of $p_{k,v}$ in segment $v+1$ as early as possible.

We denote by n_k the number of sensing slots of channel k. The corresponding historical observations are denoted by $\mathcal{H}_k = \{h_{k,1}, h_{k,2}, \ldots, h_{k,n_k}\}$. The CD algorithm detects the change of $p_{k,v}$ by using the latest $2w$ observations in $\mathcal{H}_k$. The CD statistic of channel k is defined as the difference between the average values of the first half and the last half of the $2w$ observations, which is given by

$$D_{k,w} = \frac{\left|\sum_{i=n_k-w+1}^{n_k} h_{k,i} - \sum_{i=n_k-2w+1}^{n_k-w} h_{k,i}\right|}{w}. \tag{7}$$

In the stationary environments, the $2w$ observations follow the same distribution. The corresponding statistic $D_{k,w}$ is a zero-mean random variable and its variance decreases as the observation window $2w$ increases. However, in the non-stationary environments, a part of the $2w$ observations follow the distribution with parameter $p_{k,v}$ and the rest observations follow the distribution with parameter $p_{k,v+1}$, which results in a positive drift of $D_{k,w}$. Thus, we can set a threshold δ and the CD method raises an alarm when $D_{k,w} > \delta$.

We assume that the actual change $\Delta_{k,v}$ always exceeds the threshold δ, i.e., $\Delta_{k,v} \geq \delta$. For any $\Delta_{k,v} - \delta = c$ $(c > 0)$, we have [24]

$$\begin{aligned}\mathbb{P}[D_{k,w} > \delta] &\geq 1 - 2\exp(-w(\Delta_{k,v} - \delta)^2/2) \\ &\geq 1 - 2\exp(-wc^2/2),\end{aligned} \tag{8}$$

where the first inequality is derived by using the McDiarmids inequality. Therefore, when the $2w$ observations crosses two segments, the CD algorithm raises an alarm with probability at least $1 - 2\exp(-wc^2/2)$.

It has been shown that if the idle probability $p_{k,v}$ changes by amount $\Delta_{k,v} \geq \delta$, then $w \geq 1/2\delta^2$ is sufficient to detect the change [31]. Note that a large threshold requires a smaller window size. There exists a tradeoff between detection timeliness and detection accuracy. If δ is large, the CD method can detect changes with fewer observations but the minimum change that can be detected decreases. If δ is small, the CD method can detect smaller changes but requires more observations, which can result in a large detection delay.

To avoid an inappropriate setting of δ, we use a double change detection method. Specifically, we set two thresholds in our proposed TSCD algorithm, i.e., δ_1 and δ_2 with $\delta_1 > \delta_2$, and perform a sequential CD for δ_1 and δ_2. The corresponding CD statistics for these two thresholds are denoted by D_{k,w_1} and D_{k,w_2}, respectively. If a change is detected with threshold δ_1, i.e., $D_{k,w_1} > \delta_1$, we reset the historical observations $\mathcal{H}_k \leftarrow \varnothing$. If no changes are detected, we try a smaller threshold δ_2 to detect a smaller change of $p_{k,v}$. Also, if a change is detected, i.e., $D_{k,w_2} > \delta_2$, we reset the observations $\mathcal{H}_k$. The proposed TSCD algorithm can quickly detect large changes, while still being able to detect small changes with more observations.

C. Thompson Sampling Based Channel access

In each slot t, we apply TS to decide the channel a_t that the SU tries to access. In each piecewise-stationary segment v detected by the CD algorithm, the true idle probability $p_{k,v}$ of channel k is approximated by a random variable Θ_k following a beta distribution $\text{Beta}(S_k, F_k)$, the probability density function of which is given by

$$P(\Theta_k) = \frac{\Gamma(S_k + F_k)}{\Gamma(S_k)\Gamma(F_k)} \Theta_k^{S_k - 1}(1 - \Theta_k)^{F_k - 1}, \tag{9}$$

where Γ is the Gamma function, $S_k > 0$ and $F_k > 0$ are distribution parameters. For beta distribution $\text{Beta}(S_k, F_k)$, the mean of which is given by $S_k/(S_k + F_k)$ and the variance is given by $S_k F_k/[(S_k + F_k + 1)(S_k + F_k)^2]$ [32].

At slot t in segment v, the SU draws a sample $\theta_k(t)$ from $\text{Beta}(S_k, F_k)$ for each channel $k \in \mathcal{K}$, which is assumed to be an approximation of $p_{k,v}$. To maximize the chance of successful transmission, the SU chooses the channel with the maximum sampling result, i.e.,

$$a_t = \underset{k \in \mathcal{K}}{\text{argmax}}\, \theta_k(t). \tag{10}$$

Upon observing the state $s_{a_t}(t)$, the parameters of the corresponding beta distribution are updated as

$$S_{a_t} = S_{a_t} + s_{a_t}(t), \tag{11}$$

$$F_{a_t} = F_{a_t} + 1 - s_{a_t}(t). \tag{12}$$

Note that S_k represents the number of idle slots of channel k, and F_k represents the number of busy slots of channel k. Thus, the mean of the beta distribution $\text{Beta}(S_k, F_k)$ represents the frequency of idle slots. According to the law of large numbers, the mean converges to the true idle probability $p_{k,v}$ as the number of observations $n_k = S_k + F_k$ increases to infinity, i.e.,

$$\lim_{n_k \to \infty} \frac{S_k}{S_k + F_k} = p_{k,v}. \tag{13}$$

Also, the variance of the beta distribution decreases to zero as n_{a_t} increases to infinity,

$$\begin{aligned}0 &\leq \lim_{n_k \to \infty} \frac{S_k F_k}{(S_k + F_k + 1)(S_k + F_k)^2} \\ &\leq \lim_{n_k \to \infty} \frac{1}{4(S_k + F_k + 1)} = 0,\end{aligned} \tag{14}$$

where the second inequality is derived by using $(x+y)^2 \geq 4xy$. Therefore, the sample $\theta_k(t)$ drawn from the time-varying beta distribution converges to the true idle probability $p_{k,v}$, i.e., $\lim_{t \to \infty} \theta_k(t) = p_{k,v}$.

On the one hand, the TS method ensures that the sample $\theta_k(t)$ converges to the authentic idle probability. Thus, the channel with the highest idle probability is chosen by the SU with probability 1 as time goes to infinity, which maximizes the total reward in the long term. On the other hand, during the limited piecewise-stationary segment, random sampling of the beta distribution implies that each channel has a chance to be selected by the SU. Thus, the channel characteristics can be fully explored. Therefore, the TS method achieves a promising tradeoff between exploitation and exploration.

Algorithm 1 TSCD for Single-SU Scenario

initialize $S_k = 1$, $F_k = 1$, $n_k = 0$ and $\mathcal{H}_k = \varnothing$ for all $k \in \mathcal{K}$
for $t = 1, 2, \ldots, T$ **do**
  for each $k \in \mathcal{K}$ **do**
    Draw $\theta_k(t) \sim \text{Beta}(S_k, F_k)$
  end for
  sense channel a_t given by (10) and observe state $s_{a_t}(t)$
  update S_{a_t} and F_{a_t} according to (11) and (12)
  $n_{a_t} \leftarrow n_{a_t} + 1$, $h_{a_t,n_{a_t}} = s_{a_t}(t)$
  $\mathcal{H}_{a_t} \leftarrow \mathcal{H}_{a_t} \cup \{h_{a_t,n_{a_t}}\}$
  if $n_{a_t} \geq 2w_1$ & $D_{a_t,w_1} > \delta_1$ **then**
    $\mathcal{H}_{a_t} \leftarrow \varnothing$
    $S_{a_t} = F_{a_t} = 1$, $n_{a_t} = 0$
  else if $n_{a_t} \geq 2w_2$ & $D_{a_t,w_2} > \delta_2$ **then**
    $\mathcal{H}_{a_t} \leftarrow \varnothing$
    $S_{a_t} = F_{a_t} = 1$, $n_{a_t} = 0$
  end if
end for

D. Thompson Sampling with Change Detection

The proposed TSCD algorithm consists of two components, i.e., the CD part and the TS part. The CD strategy is responsible for detecting the variations of channel idle probability in the non-stationary environments. The TS method is responsible for balancing the exploration and exploitation of channel selection in each detected piecewise-stationary segment.

In each slot t, we first apply the TS method to decide the sensing channel a_t based on the historical sensing observations. The observed channel state $s_{a_t}(t)$ is fed back to both the TS method and the CD strategy. The TS method utilizes $s_{a_t}(t)$ to update the parameters of distribution $\text{Beta}(S_{a_t}, F_{a_t})$ for the next slot channel selection. The CD strategy utilizes $s_{a_t}(t)$ to detect the statistical change of channel a_t and raises an alarm to restart the TS method once a change is detected. Details of the proposed TSCD algorithm are summarized in Algorithm 1.

E. Regret Analysis

In this section, we present the regret analysis of our proposed TSCD algorithm. The best channel in segment v is given by $k_v^* = \text{argmax}_{k\in\mathcal{K}} p_{k,v}$. The idle probability gap between channel k and the best channel k_v^* is denoted by $\varepsilon_{k,v}$ and we have $\varepsilon_{k,v} = p_{k_v^*,v} - p_{k,v}$. Let $d(p\|q) = p\ln(\frac{p}{q}) + (1-p)\ln(\frac{1-p}{1-q})$ denote the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q. We define $\Delta_v = \max_{k\in\mathcal{K}} \Delta_{k,v}$ and $Q_v(w) = \frac{\sqrt{w\ln(2KT^2)}}{2\Delta_v} + \frac{9T}{2V}\sqrt{\frac{2}{w}}$.

Theorem 1: Running the Algorithm 1 with $\delta_1 = \sqrt{\ln(2KT^2)/w_1}$ and $\delta_2 = \sqrt{\ln(2KT^2)/w_2}$, we have

$$R(T) \leq \underbrace{\sum_{v=1}^{V}(\tilde{R}_v + C_v)}_{(a)} + \underbrace{\sum_{v=1}^{V} 2\min(T/V, Q_v(w))}_{(b)} + \underbrace{2V}_{(c)}, \quad (15)$$

where $\tilde{R}_v = \sum_{k\in\mathcal{K}} \frac{\varepsilon_{k,v}(\ln T + \ln\ln T)}{d(p_{k_v^*,v}\|p_{k,v})}$, C_v is a constant depending on the values of $p_{1,v}, p_{2,v}, \ldots, p_{K,v}$ and $w = \text{argmin}_{w\in\{w_1,w_2\}} Q_v(w)$. Term (a) bounds the regret in the detected piecewise-stationary segments, term (b) bounds the regret incurred by the CD delay and term (c) gives the upper bound of the regret associated with the false alarms and miss detections. The segment length τ is large enough for CD.

Proof of Theorem 1: We first present the proof of term (a). We demonstrate that the regret in each detected piecewise-stationary segment v is lower than $\tilde{R}_v + C_v$. Recall that the regret in segment v is given by

$$R_v(\tau) = \sum_{t=(v-1)\tau+1}^{v\tau} p_{k_v^*,v} - \mathbb{E}[\sum_{t=(v-1)\tau+1}^{v\tau} p_{a_t,v}], \quad (16)$$

where $\tau \leq T$. To bound the regret $R_v(\tau)$, we rewrite it as

$$R_v(\tau) = \sum_{k\in\mathcal{K}} \varepsilon_{k,v}\mathbb{E}[n_k]. \quad (17)$$

Thus, we only need to bound the expected number of selections $\mathbb{E}[n_k]$ of channel k since the gap $\varepsilon_{k,v}$ between the best channel k_v^* and channel k is fixed. For each channel k and $\epsilon > 0$, there exist a constant $Z(\epsilon, p_{k_v^*,v}, p_{k,v})$ such that [32]

$$\mathbb{E}[n_k] \leq (1+\epsilon)\frac{\ln T + \ln\ln T}{d(p_{k_v^*,v}\|p_{k,v})} + Z(\epsilon, p_{k_v^*,v}, p_{k,v}). \quad (18)$$

Substituting in equation (17), we get

$$R_v(\tau) \leq (1+\epsilon)\sum_{k\in\mathcal{K}} \frac{\varepsilon_{k,v}(\ln\tau + \ln\ln\tau)}{d(p_{k_v^*,v}\|p_{k,v})} + C_v, \quad (19)$$

where $C_v = \varepsilon_{1,v}Z(\epsilon, p_{k_v^*,v}, p_{1,v}) + \varepsilon_{2,v}Z(\epsilon, p_{k_v^*,v}, p_{2,v}) + \ldots + \varepsilon_{K,v}Z(\epsilon, p_{k_v^*,v}, p_{K,v})$. The fact that the inequality (19) holds for every $\epsilon > 0$ proves that the regret $R_v(\tau)$ in each detected segment v is lower than $\tilde{R}_v + C_v$.

Next, we give the proof of term (b). We denote by R_v^D the expected CD delay in segment v. If the actual change $\Delta_{k,v}$ of channel k exceeds the threshold δ_1 and the number of sensing slots n_k is larger than the window $2w_1$, then the expected CD delay R_v^D is bounded by [24]

$$R_v^D \leq \frac{\min(T/V, Q_v(w_1))}{1 - 2\exp(-w_1c^2/2)}. \quad (20)$$

By setting $c = \sqrt{2\ln(2T)/w_1}$ and $T \geq 2$, we further obtain that

$$\begin{aligned} R_v^D &\leq \frac{\min(T/V, Q_v(w_1))}{(1-1/T)} \\ &\leq 2\min(T/V, Q_v(w_1)). \end{aligned} \quad (21)$$

If $\Delta_{k,v} \geq \delta_2$ and $n_k \geq 2w_2$, by setting $c = \sqrt{2\ln(2T)/w_2}$ and $T \geq 2$, we can also achieve that $R_v^D \leq 2\min(T/V, Q_v(w_2))$. Summing the CD delay regrets of all segments, we obtain that $\sum_{v=1}^{V} R_v^D \leq \sum_{v=1}^{V} 2\min(T/V, Q_v(w))$.

Last, we prove that the regret incurred by the false alarms and miss detections is no more than $2V$. We denote the probability of raising a false alarm in each segment as P_{false}, which is bounded by [24]

$$P_{\text{false}} \leq 2wK(1 - (1 - 2\exp(-w\delta^2))^{T/2w}), \quad (22)$$

where $(w, \delta) \in \{(w_1, \delta_1), (w_2, \delta_2)\}$. Note that $(1-x)^a > 1-ax$ for any $a > 1$ and $0 < x < 1$. Using the fact in the above inequality, we have that $P_{\text{false}} \leq 2KT\exp(-w\delta^2)$. Since

$\delta_1 = \sqrt{\ln(2KT^2)/w_1}$ and $\delta_2 = \sqrt{\ln(2KT^2)/w_2}$, we further obtain that $P_{\text{false}} \le 1/T$. Considering the segment length $\tau \le T$, this result means that at most one false alarm is raised in each detected piecewise-stationary segment. In (8), we have demonstrated that the probability of achieving a successful change detection is at least $1-2\exp(-wc^2/2)$ in each detected segment, where $w \in \{w_1, w_2\}$. By setting $c = \sqrt{2\ln(2T)/w}$, we have that the probability of miss detection is at most $2\exp(-wc^2/2) = 1/T$. Thus, it is expected that only one miss detection will occur in each segment. Therefore, the SU gets less than $2V$ of regret caused by the false alarms and missed detections on all segments.

V. TSCD-TSCA for Multi-SU Scenario

In this section, we concentrate on the multi-SU DSA scenario. We identify the corresponding challenges and extend the TSCD algorithm to the multi-SU scenario, for which we use the TSCA policy to orthogonalize the SUs on different channels.

A. Challenges

If all the SUs apply the TSCD algorithm for the single-SU scenario, then collisions are inevitable since they will choose the same channel with the highest idle probability. In the multi-SU scenario, in addition to competing for the channel with high idle probability, the SUs need to cooperate with each other to avoid collisions. There are two challenges to achieving the tradeoff between competition and cooperation.

- **Distributed Coordination**: There are no prior protocols, central controller, or dedicated control channels for the SUs. For any SU $m \in \mathcal{M}$, there is no information of the instant channels selected by other SUs and thus the SU can only rely on its own collision indicator $\zeta_m(t)$.
- **Unsynchronized Estimation**: Due to the non-stationarity of the environments and the randomness of the TSCD algorithm, different SUs may have different estimations of channel idle probabilities. Thus, classical order-based schemes may lead to collisions since they assume that SUs have the same estimation of channel ranking [15, 16].

B. Thompson Sampling Based Collision Alleviation

At slot t, we denote by $\boldsymbol{\theta}_m(t) = [\theta_{m,1}(t), \ldots, \theta_{m,K}(t)]$ the samples that SU m draws for each channel $k \in \mathcal{K}$ by using the TSCD algorithm. Let $\sigma(i, \boldsymbol{\theta}_m(t))$ be the index of the i-th largest entry in $\boldsymbol{\theta}_m(t)$. The estimated M best channels of SU m are then given by

$$A_m(t) = \{\sigma(1, \boldsymbol{\theta}_m(t)), \ldots, \sigma(M, \boldsymbol{\theta}_m(t))\}. \tag{23}$$

SU m can choose a channel $a_{m,t} \in A_m(t)$ to access. The set $A_m(t)$ ensures that SU m can always access a channel with a high idle probability. However, due to the absent of coordination policy among the SUs, SU m may still encounter a collision when accessing channel $a_{m,t}$. Thus, we propose the TSCA policy to alleviate the collision.

Algorithm 2 TSCD-TSCA for Multi-SU Scenario

initialize $S_{m,k} = 1$, $F_{m,k} = 1$, $n_{m,k} = 0$, $\mathcal{H}_{m,k} = \varnothing$, $J_{m,k} = 1$, $L_{m,k} = 1$ for each SU $m \in \mathcal{M}$ and channel $k \in \mathcal{K}$
for each SU $m \in \mathcal{M}$ **do**
  for $t = 1, 2, 3, \ldots, T$ **do**
    for each $k \in \mathcal{K}$ **do**
      Draw $\theta_{m,k}(t) \sim \text{Beta}(S_{m,k}, F_{m,k})$
    end for
    calculate $A_m(t)$ as in (23)
    for each $k \in A_m(t)$ **do**
      Draw $\phi_{m,k}(t) \sim \text{Beta}(J_{m,k}, L_{m,k})$
    end for
    sense channel $a_{m,t}$ given by (24) and observe state $s_{a_{m,t}}(t)$
    update $S_{a_{m,t}}$ and $F_{a_{m,t}}$ according to (11) and (12)
    $n_{m,a_{m,t}} \leftarrow n_{m,a_{m,t}} + 1$, $h_{a_{m,t},n_{a_{m,t}}} = s_{a_{m,t}}(t)$
    $\mathcal{H}_{m,a_{m,t}} \leftarrow \mathcal{H}_{m,a_{m,t}} \cup \{h_{a_{m,t},n_{a_{m,t}}}\}$
    perform change detection as in Algorithm 1
    if $s_{a_{m,t}}(t) = 1$ **then**
      update $J_{m,a_{m,t}}$ and $L_{m,a_{m,t}}$ as in (25) and (26)
    end if
  end for
end for

We denote $\eta_{m,k}$ as the probability that SU m does not encounter a collision when choosing channel k to access, i.e., $\eta_{m,k} = \mathbb{P}[\zeta_m(t) = 1 | a_{m,t} = k]$. At each slot t, SU m is supposed to access the channel with the highest $\eta_{m,k}$ in $A_m(t)$. However, the probability $\eta_{m,k}$ of channel k is unknown to SU m. Here, we propose a TS based method to estimate $\eta_{m,k}$ by using the historical collision indicators $\zeta_m(t)$.

Specifically, for any SU $m \in \mathcal{M}$, $\eta_{m,k}$ is approximated by a random variable $\Phi_{m,k}$, which follows a beta distribution $\text{Beta}(J_{m,k}, L_{m,k})$. At slot t, SU m first draws a sample $\theta_{m,k}(t)$ from $\text{Beta}(S_{m,k}, F_{m,k})$ for each channel $k \in \mathcal{K}$ and calculates the corresponding M best channels $A_m(t)$. Then, SU m draws a sample $\phi_{m,k}(t)$ from $\text{Beta}(J_{m,k}, L_{m,k})$ for each channel $k \in A_m(t)$, which is assumed to be an approximation of $\eta_{m,k}$. To maximize the probability of successful transmission, SU m senses the channel with the maximum sampling result, i.e.,

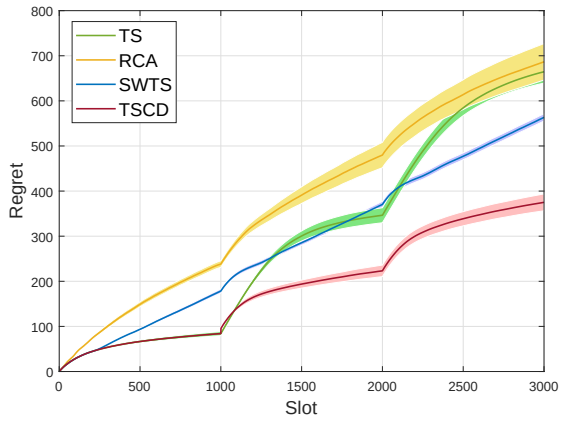
$$a_{m,t} = \underset{k \in A_m(t)}{\text{argmax}}\, \phi_{m,k}(t). \tag{24}$$

If $s_{a_{m,t}}(t) = 0$, which means channel $a_{m,t}$ is occupied by the primary network, SU m does not transmit and no $\zeta_m(t)$ is observed. Thus, we have no additional information on $\eta_{m,a_{m,t}}$, and the parameters $J_{m,a_{m,t}}$ and $L_{m,a_{m,t}}$ keep unchanged. If $s_{a_{m,t}}(t) = 1$, which means channel $a_{m,t}$ is idle, SU m transmits data and observes $\zeta_m(t)$. We update the beta distribution $\text{Beta}(J_{m,a_{m,t}}, L_{m,a_{m,t}})$ by

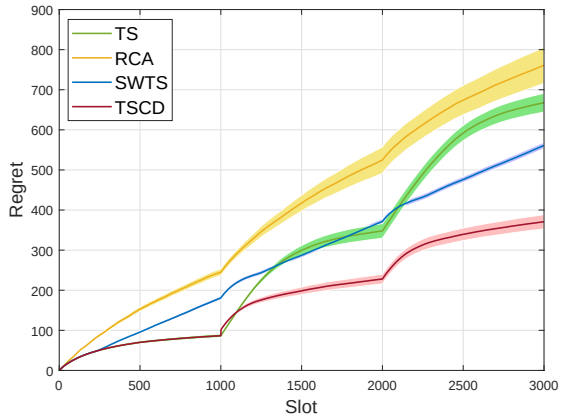
$$J_{m,a_{m,t}} = J_{m,a_{m,t}} + \zeta_m(t), \tag{25}$$

$$L_{m,a_{m,t}} = L_{m,a_{m,t}} + 1 - \zeta_m(t). \tag{26}$$

This article has been accepted for publication in IEEE Transactions on Cognitive Communications and Networking. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2023.3237578

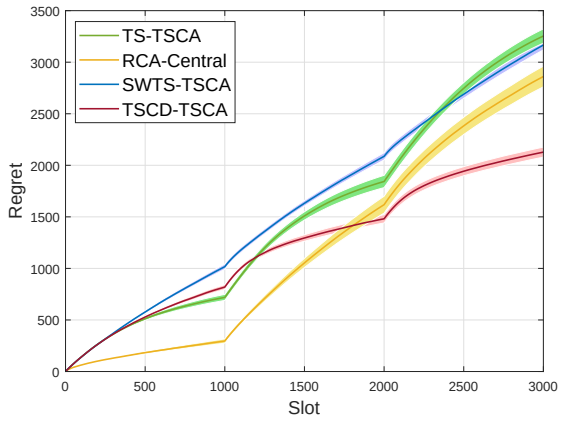


(a) Non-stationary Bernoulli model

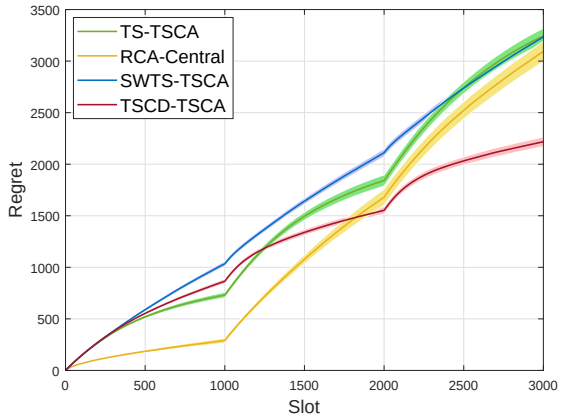


(b) Non-stationary Markovian model

Fig. 2. Regret along with its 95% confidence interval as a function of time slot for $K = 20$, $T = 3000$ and $\lambda = 0.3$.



(a) Non-stationary Bernoulli model



(b) Non-stationary Markovian model

Fig. 3. Regret along with its 95% confidence interval as a function of time slot for $K = 20$, $T = 3000$, $\lambda = 0.3$ and $M = 5$.

Thus, for any SU m and channel k, $J_{m,k}$ represents the total number of successful transmissions (i.e., $s_k(t) = 1, \zeta_m(t) = 1$), and $L_{m,k}$ represents the total number of collisions (i.e., $s_k(t) = 1, \zeta_m(t) = 0$). Therefore, for similar reasons as we explained for functions (11) and (12), the sample $\phi_{m,k}(t)$ drawn from Beta$(J_{m,k}, L_{m,k})$ converges to $\eta_{m,k}$ as the time slot goes, which implies the SUs will split and choose different channels to alleviate collisions. Details of the proposed TSCD-TSCA algorithm are summarized in Algorithm 2.

VI. Numerical Results

In this section, we present the numerical results of our proposed TSCD and TSCD-TSCA algorithms under different network settings, i.e., channel number K, SU number M and average idle probability $\lambda = \sum_{k=1}^{K}\sum_{v=1}^{V} p_{k,v}/KV$, which represents the average primary load. The length of a slot is set as 1ms and each segment contains $\tau = 1000$ slots. Thus, the idle probability of each channel remains unchanged within 1 second. For each channel $k \in \mathcal{K}$ in segment $v \in [1, V]$, the idle probability $p_{k,v}$ is sampled uniformly from $(0, 1)$.

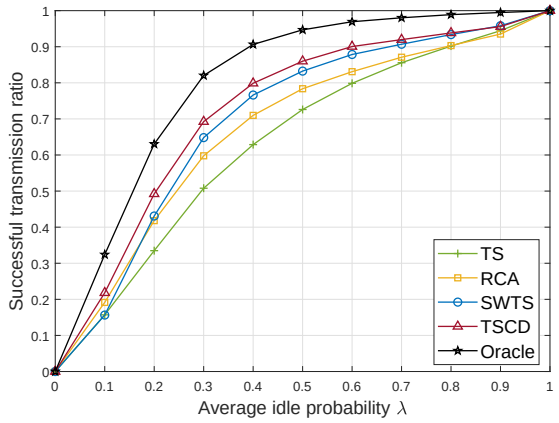
In the single-SU scenario, we compare the proposed TSCD algorithm with the regenerative cycle algorithm (RCA) [19], the TS [26] and the sliding window TS (SWTS) [29] algorithms. The parameters of these algorithms are tuned as suggested in the corresponding papers. Specifically, the window size of the SWTS algorithm is set to $2\sqrt{T\ln T/(V-1)}$. In addition, we also present the performance of an oracle as a baseline, which always senses the channel with the highest idle probability in each segment. For our proposed methods, the change detection parameters are set as $\delta_1 = 0.25$, $\delta_2 = 0.08$, $w_1 = 32$ and $w_2 = 156$ based on a series of experiments. All numerical results are averaged by 1000 Monte Carlo simulations.

In the multi-SU scenario, we combine the TS and the SWTS algorithms with the proposed TSCA policy. The RCA algorithm is extended to the multi-SU scenario by using a central controller [19]. Note that in this paper we consider DSA in a distributed scenario, where there is no central controller for the SUs. The central controller ensures that no collision occurs at each slot, but introduces additional signaling exchange and communication costs. The performance of an oracle is also presented, which assigns the M SUs to sense the M channels with the highest idle probabilities in each segment.

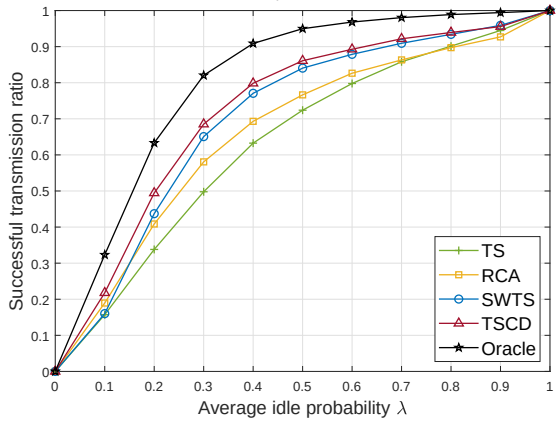
A. Regret Performance

Fig. 2 shows the regret along with its 95% confidence interval as a function of the time slot for $K = 20$, $T = 3000$ and $\lambda = 0.3$. Since the regret of the oracle is zero, we don't

This article has been accepted for publication in IEEE Transactions on Cognitive Communications and Networking. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2023.3237578

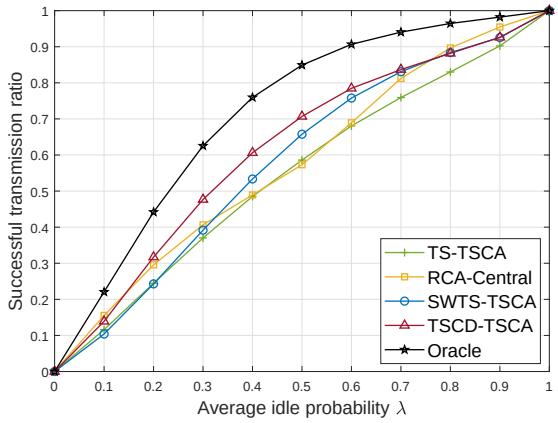


(a) Non-stationary Bernoulli model

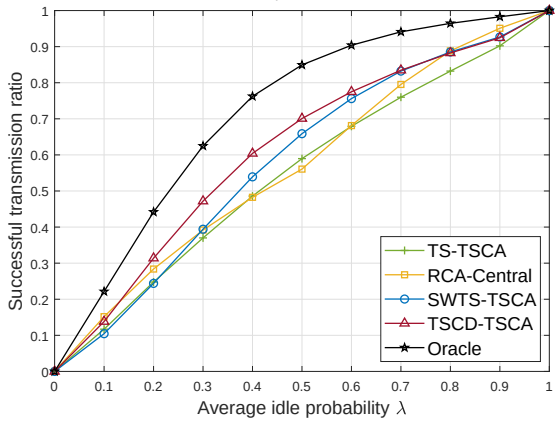


(b) Non-stationary Markovian model

Fig. 4. STR as a function of average idle probability λ for $K = 20$ and $T = 10000$.



(a) Non-stationary Bernoulli model



(b) Non-stationary Markovian model

Fig. 5. STR as a function of average idle probability λ for $K = 20$, $M = 5$ and $T = 10000$.

show the curve of the oracle in the figure. The trends are similar in the two non-stationary models. The regrets of the algorithms increase with the time slot and the proposed TSCD algorithm achieves the lowest regret, which can quickly detect the change of channel idle probability and converge to the best channel in each segment. The SWTS algorithm can also adapt to the non-stationary environments. However, it only uses recent observations in the sliding window, which results in information loss and reduces its capability to identify the best channel. The RCA algorithm and the TS algorithm cannot track the changes of channel idle probability and the regrets of them increase significantly with time.

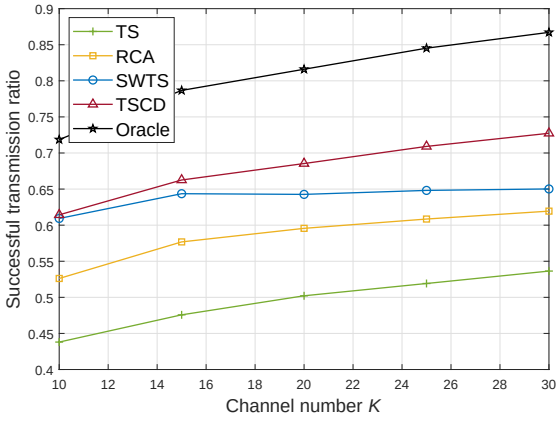
Fig. 3 shows the regret along with its 95% confidence interval as a function of the time slot for $K = 20$, $T = 3000$, $\lambda = 0.3$ and $M = 5$. The regrets of the algorithms increase with time and the proposed TSCD-TSCA algorithm achieves the lowest regret after the second segment ($t \geq 2000$). Due to the central controller, the RCA-Central algorithm initially shows the lowest regret. However, the RCA-Central algorithm discards the observations outside the regenerative cycle, which also has the problem of information loss. It cannot track the changes of channel statistics and its convergence to the best channel is limited due to the loss of information. Therefore, the regret of the RCA-Central algorithm increases significantly after the first segment.

The proposed TSCD-TSCA algorithm can detect the changes of idle probability and orthogonalize the SUs on the M channels with the highest idle probabilities in each segment. It shows a lower growth rate of regret than the RCA-Central algorithm. The regret of the TS-TSCA algorithm increases significantly with time since it cannot adapt to the changes of channel statistics. The regret of the SWTS-TSCA algorithm grows almost linearly in each segment as the loss of information leads to a decrease in its convergence.

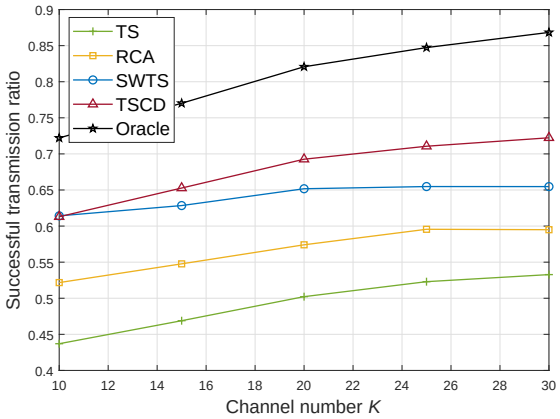
B. Successful Transmission Ratio

Fig. 4 shows the STRs of different single-SU DSA algorithms as a function of average idle probability λ for $K = 20$ and $T = 10000$. As we can see, the STRs increase with λ for all considered algorithms in the two non-stationary models, since a larger λ implies more transmission opportunities. When the primary channels are fully occupied, i.e., $\lambda = 0$, or fully unoccupied, i.e., $\lambda = 1$, there is no difference between the DSA algorithms. Thus, all considered algorithms achieve the same performance in those extreme scenarios. When primary channels are partially occupied, the proposed TSCD algorithm outperforms the other algorithms as it can detect the changes of channel statistics and maintains a balance between the exploitation of the currently best channel and the exploration of potential better channels.

This article has been accepted for publication in IEEE Transactions on Cognitive Communications and Networking. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2023.3237578

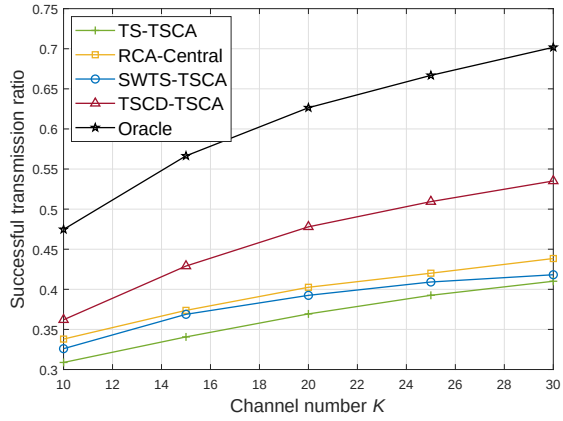


(a) Non-stationary Bernoulli model

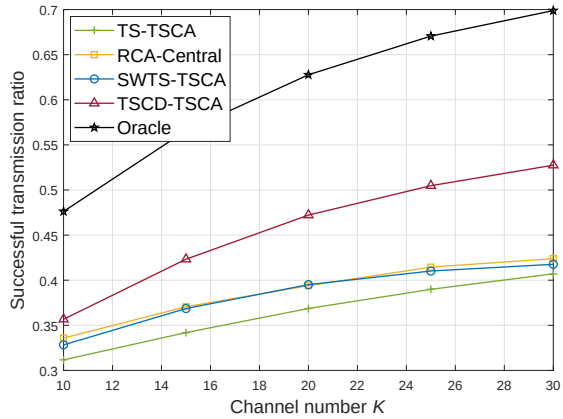


(b) Non-stationary Markovian model

Fig. 6. STR as a function of channel number K for $\lambda = 0.3$ and $T = 10000$.



(a) Non-stationary Bernoulli model



(b) Non-stationary Markovian model

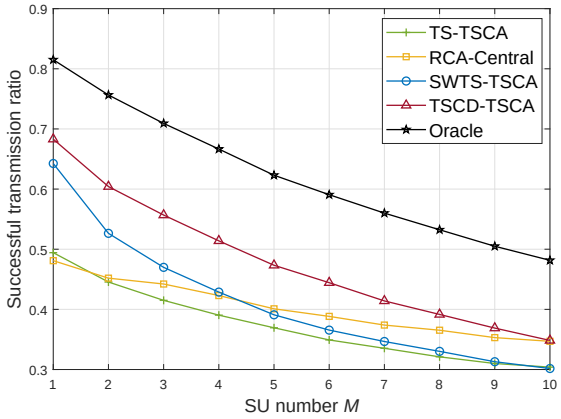
Fig. 7. STR as a function of channel number K for $\lambda = 0.3$, $M = 5$ and $T = 10000$.

Fig. 5 shows the STRs of different multi-SU DSA algorithms as a function of average idle probability λ for $K = 20$, $M = 5$ and $T = 10000$. In the multi-SU scenario, the STR of a DSA algorithm depends not only on the channel statistic learning but also on the coordination among SUs. When $\lambda < 0.1$ or $\lambda > 0.8$, there is no obvious difference in the statistical learning of each algorithm. The coordination policy plays a more important role in the STR performance. Thus, the RCA-Central algorithm achieves a higher STR than the proposed TSCD-TSCA algorithm. When $0.1 < \lambda < 0.8$, the importance of statistical learning begins to manifest. The proposed algorithm can quickly identify the best channels in each segment and coordinate the SUs to access these channels without collision, which achieves the highest STR.

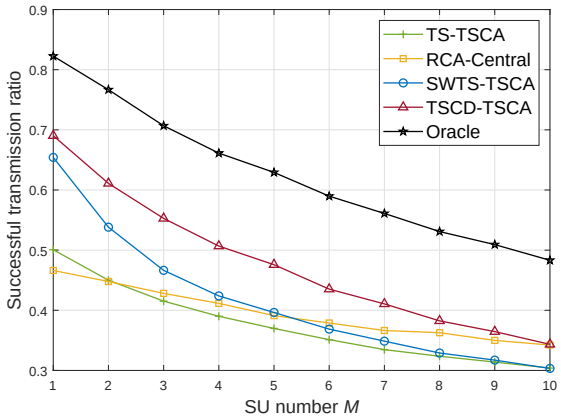
Fig. 6 shows the STRs of different single-SU DSA algorithms as a function of channel number K for $\lambda = 0.3$ and $T = 10000$. The proposed TSCD algorithm achieves the highest STR among all considered algorithms except for the oracle. Based on the largest order statistic, the mean idle probability of the best channel increases with the number of channels. Thus, the STR of the SU should increase with K. However, the increasing number of channels also increases the difficulty of identifying the best primary channel, as the sensing capability of the SU stays the same. Note that the window size $2\sqrt{T\log T/(V-1)}$ of the SWTS algorithm is independent of channel number. When the channel number becomes large, the increasing difficulty introduced by extra channels offsets the additional transmission opportunities and the curve of the SWTS algorithm flattens out.

Fig. 7 shows the STRs of different multi-SU DSA algorithms as a function of channel number K for $\lambda = 0.3$, $M = 5$ and $T = 10000$. The trends are similar to that in the single-SU scenario. In both channel state models, the STRs of the algorithms increase with K and the proposed TSCD-TSCA algorithm achieves the highest STR. Compared with other algorithms, the proposed TSCD-TSCA algorithm can better deal with the additional exploration cost brought by the increasing channel number. Thus, the gap between our proposed algorithm and other algorithms increases with K.

Fig. 8 shows the STRs of different multi-SU DSA algorithms as a function of SU number M for $\lambda = 0.3$, $K = 20$ and $T = 10000$. The average STR of each SU decreases with M. The proposed TSCD-TSCA algorithm can quickly orthogonalize the SUs on the M best channels in each segment and thus achieves the highest STR except for the oracle. The increasing SU number increases the difficulty of coordination, as the total number of channels remains unchanged. Due to the central controller, the RCA-Central algorithm ensures that no collision occurs among the SUs regardless of the SU number M. Thus, the RCA-Central algorithm shows a lower decrease



(a) Non-stationary Bernoulli model



(b) Non-stationary Markovian model

Fig. 8. STR as a function of SU number M for $\lambda = 0.3$, $K = 20$ and $T = 10000$.

rate of STR than other algorithms.

VII. CONCLUSIONS

In this paper, we investigated the dynamic spectrum access problem in non-stationary environments and proposed two algorithms for the single-SU and the multi-SU scenarios, respectively. In the single-SU scenario, the proposed TSCD algorithm can effectively detect the changes of idle probabilities without prior knowledge. Besides, the proposed TSCD algorithm achieves a promising tradeoff between the exploitation of the currently best channel and the exploration of potentially better channels, which enables the SU to quickly converge to the best channel in each piecewise-stationary segment. In the multi-SU scenario, we propose a TS based collision alleviation policy to orthogonalize the SUs on different channels in a distributed manner, where there is no information exchange among the SUs. Numerical results show that the proposed algorithms yield a higher successful transmission ratio for the SUs than the existing ones with various network parameters such as the average idle probability, the number of channels and SUs.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers, whose invaluable comments helped improve the presentation of this paper substantially.

REFERENCES

[1] F. Li *et al.*, "Advances and emerging challenges in cognitive Internet-of-Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5489–5496, Aug. 2020.

[2] A. J. Onumanyi, A. M. Abu-Mahfouz, and G. P. Hancke, "Cognitive radio in low power wide area network for IoT applications: Recent approaches, benefits and challenges," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7489–7498, Dec. 2020.

[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[4] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[5] Q. Zhao *et al.*, "Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[6] M. Zandi, M. Dong, and A. Grami, "Distributed stochastic learning and adaptation to primary traffic for dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1675–1688, Mar. 2016.

[7] 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.

[8] S. M. Zafaruddin *et al.*, "Distributed learning for channel allocation over a shared spectrum," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2337–2349, Oct. 2019.

[9] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, Nov. 2010.

[10] M. K. Hanawal and S. J. Darak, "Multiplayer bandits: A trekking approach," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2237–2252, May 2022.

[11] A. Javanmardi, M. A. Qureshi, and C. Tekin, "Decentralized dynamic rate and channel selection over a shared spectrum," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3787–3801, Jun. 2021.

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

[13] 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.

[14] 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.

[15] 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.

[16] L. Besson and E. Kaufmann, "Multi-player bandits revisited," in *Proc. Algorithmic Learn. Theory*, Lanzarote, Spain, Apr. 2018.

[17] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

[18] 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.

[19] 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.

[20] N. Modi, P. Mary, and C. Moy, "Qos driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 49–66, Mar. 2017.

[21] A. Magesh and V. 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.

[22] A. B. H. Alaya-Feki, E. Moulines, and A. LeCornec, "Dynamic spectrum access with non-stationary multi-armed bandit," in *IEEE Workshop Signal Process. Adv. Wireless Commun. SPAWC*, Jul. 2008, pp. 416–420.

[23] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. 22nd Int. Conf. Algorithmic Learn. Theory*, Oct. 2011.

[24] Y. Cao *et al.*, "Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit," in *Proc. Int. Conf. Artif. Intell. Stat.*, Naha, Okinawa, Japan, Apr. 2019.

[25] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. Adv. Neural Inf. Proces. Syst.*, Granada Spain, Dec. 2011.

[26] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," in *Proc. Int. Conf. Artif. Intell. Stat.*, Scottsdale, AZ, USA, Apr. 2013.
[27] 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.
[28] 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.
[29] F. Trovo *et al.*, "Sliding-window Thompson sampling for non-stationary settings," *J. Artif. Intell. Res.*, vol. 68, pp. 311–364, May 2020.
[30] F. Awin, E. Abdel-Raheem, and K. Tepe, "Blind spectrum sensing approaches for interweaved cognitive radio system: A tutorial and short course," *IEEE Commun. Surv. Tuts.*, vol. 21, no. 1, pp. 238–259, 1st Quart. 2019.
[31] P. Auer, P. Gajane, and R. Ortner, "Adaptively tracking the best bandit arm with an unknown number of distribution changes," in *Proc. Annu. Conf. Learn. Theory*, Phoenix, Arizona, Jun. 2019.
[32] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. Annu. Conf. Learn. Theory*, Edinburgh, Scotland, Jun. 2012.

Shuai Ye received the BS degree in 2020 from Nanjing University, Nanjing, China, where he is currently pursuing the PhD degree at the School of Electronic Science and Engineering. His current research interests include dynamic spectrum access and online learning.

Tianyu Wang received the BS degrees in physics and a double major in computer software from Peking University, Beijing, China, in 2011, and the Ph.D. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2016. He is currently an assistant professor in the School of Electronic Science and Engineering at Nanjing University, China. His current research interest focuses on network slicing and machine learning in wireless networks.

Shaowei Wang received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. He joined the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, as a Faculty Member, in 2006, where he is currently a Full Professor. From 2012 to 2013, he was a Visiting Scholar/a Professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include communications and networking, operations research, and machine learning.