IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 5, 2020, accepted October 14, 2020, date of publication October 28, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034360

Proactive Dynamic Channel Selection Based on Multi-Armed Bandit Learning for 5G NR-U

YANPENG SHI[1], QIMEI CUI[1], (Senior Member, IEEE), WEI NI[2], (Senior Member, IEEE), AND ZESONG FEI[3], (Senior Member, IEEE)

[1]National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Digital Productivity and Services Flagship, Commonwealth Scientific and Industrial Research Organization, Sydney, NSW 2122, Australia
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Qimei Cui (cuiqimei@bupt.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant L182038, and in part by the National Youth Top-Notch Talent Support Program.

ABSTRACT With an increasing demand of mobile data traffic in fifth-generation (5G) wireless communication systems, new radio-unlicensed (NR-U) technology has been regarded as a promising technology to address the exponential growth of data traffic by offloading the traffic to unlicensed bands. Nevertheless, how to efficiently share the unlicensed spectrum resource among the NR and Wi-Fi systems is a key challenge to be addressed, especially in a dynamic network environment. In this article, we investigate a distributed channel access mechanism and focus on the channel selection for NR-U users to decide the optimal unlicensed channel for uplink traffic offloading. We formulate the selection problem as a non-cooperative game, which is proven to be an exact potential game. However, the Nash equilibrium (NE) point is hard to achieve, due to the unknown dynamic environment. Based on multi-armed bandit learning techniques, an online learning distributed channel selection algorithm (OLDCSA) is proposed and proven to have similar performance to the NE point. Finally, simulation results reveal that our proposed algorithm outperforms the existing random selection by 16.45 % on average and is close to the exhaustive search in the dynamic unknown environment.

INDEX TERMS New radio-unlicensed (NR-U), dynamic multiple-channel selection transmission, exact potential game, multi-armed bandit learning.

I. INTRODUCTION

With the continuous rapid growth of the mobile Internet and the Internet of Things, the huge traffic volumes pose a major challenge to network capacity [1], which has motivated the innovation to improve spectrum utilization by offloading traffic to unlicensed bands. According to [2]–[4], efficient utilization of the unlicensed spectrum has been regarded as a promising solution to meet the requirements of the high data rate and large capacity in 5G systems.

The Third Generation Partnership Project (3GPP) has started discussions and works about the deployments of LTE, LTE-A and NR in the unlicensed spectrum. Licensed-Assisted Access (LAA), which utilizes the LTE carrier aggregation technology to aggregate carriers across both the licensed and unlicensed bands, was first introduced in LTE Rel-13 [5]. A primary cell (PCell) of a licensed carrier will provide reliable control information, to guarantee the quality of service, reliability and mobility. A secondary cell (SCell) bears the large amount of traffic transmitted via unlicensed spectrum under the control of the licensed frequency band to serve the purpose of authorized band traffic offloading. In enhanced LAA (eLAA) Release-14, uplink (UL) transmissions are also extended and supported [6]. Along with the specification of New Radio (NR) in 5G, the 3GPP has carried out the study on NR Unlicensed (NR-U) in Release-16 exploiting the unlicensed spectrum supporting abundant frequency bands, such as 2.4GHz, 3.5GHz, 5GHz, 6GHz and even 60GHz [7]. The multi-bands give the possibility to select an appropriate channel in the unlicensed band for transmission. In this article, we focus on the unlicensed spectrum performance of the uplink transmission which is adapted for both eLAA and NR-U.

The unlicensed spectrum is shared among operators, which could cause interference to each other. Thus, channel access mechanisms such as Listen-Before-Talk (LBT) and

The associate editor coordinating the review of this manuscript and approving it for publication was Christian Esposito.

Channel-Clear Assessment (CCA), are designed before transmission to strive for fair and friendly coexistence between different technologies operating on the unlicensed spectrum. Besides, there are other regulatory requirements specified in [5], including limits on the maximum channel occupancy time and maximum transmit power. There have been some researches analyzing the performance of coexistence system. A Markov chain model of the throughput performance has been proposed in [8] and the delay analysis is presented in [9]. Optimizations of the system achievable throughput and spectrum efficiency (SE) are shown in [10], [11]. In [10], small base stations (SBS) adaptively adjust the back-off window according to the available spectrum bandwidth and Wi-Fi traffic. Both licensed and unlicensed spectrum bands are jointly allocated to optimize the spectral efficiency in [11]. The authors of [12] propose a user offloading method, which optimizes the throughput of both the LAA and Wi-Fi system. A fairness-based LAA and resource scheduling scheme are presented in [13] to improve the system throughput and the SE of coexisting systems. Apart from throughput, energy efficiency and effective capacity are also considered for LAA/Wi-Fi coexistence systems in [14]–[16]. However, these existing studies are only limited to a single unlicensed Wi-Fi channel, ignoring the potential gains brought by channel selection and a switch between multiple unlicensed channels. In practice, the LAA system still also operates static channel allocation. For example, Verizon and T-Mobile mostly operate on Wi-Fi 36, 40 and 44 channels. Similarly, AT&T operates on channels 149, 153, 157 and 165 in the 5 GHz band [17].

A large amount of unlicensed spectrum resources, more than 700 MHz, can be used by LAA system to carry out traffic offerings [5]. A heuristic dynamic unlicensed component carrier selection (UCCS) algorithm was proposed to minimize inter-cell interference and and improve the overall network performance, which estimates the quality of each free channel based on User Equipment (UE) feedback and selects the best channel [18]. The authors of [19] developed a cooperative Nash bargaining game (NBG) and a one-sided matching game to meet the quality-of-service (QoS) needs of LTE-U users for a network with multiple operators. Furthermore, quality of experience (QoE) was considered and a virtual coalition formation game (VCFG) was formulated to solve the unlicensed band selection problem in [20]. In [21], the authors optimized duty cycle to address the problem of coexistence between LTE-U and Wi-Fi in multi-channel scenarios based on Q-learning. A 2-Dimension Hopfield Neural Network (2D-HNN)-based algorithm was proposed to achieve fairness considering both the LTE-U data rate and the QoS requirements of WiFi networks [22]. Unlike the works based on ON/OFF coexistence scheme [19]–[22], a listen-before-talk (LBT)-based mechanism was considered in a multi-operator scenario [23], [24]. The authors in [23] jointly considered channel selection, carrier aggregation and fractional spectrum access for unlicensed LTE (U-LTE) networks in a centralized management and the work [24] exploited deep learning based on long short-term memory (LSTM) in resource allocation. Hence, in the existing literature, the solutions to the unlicensed band selection problem among SBSs and WAPs have been based on the assumption that the small-cell base stations (SBSs) have full knowledge of the environment and complete information about actions taken by other SBSs. Moreover, the environment is required to be static during the convergence of the algorithms. In recent years, machine learning (ML), such as supervised, unsupervised, and reinforcement learning, has been increasingly adopted in wireless communications to address the complex environment problems and achieved good performance [25], [26]. Thus, we have designed an online learning distributed channel selection scheme in the dynamic heterogeneous network under the incomplete information for uplink traffic offloading. The main contributions of this article are summarized as follows:

- We present a new user decision channel access mechanism for uplink traffic offloading in a dynamic heterogeneous networks and formulate a new optimization problem to maximize the sum-rate of NR-U users.
- We convert the channel selection optimization problem to a non-cooperative game, in which the utility function of each user is defined as the expectation of its achievable transmission rate. We prove that the game is an exact potential game, and has at least one pure strategy NE point.
- Leveraging the multi-armed bandit learning theory, an online learning distributed channel selection algorithm OLDCSA) is proposed to tackle the unknown, dynamic network environment and the incomplete information constraints of the potential game. The proposed algorithm can asymptotically converge to the NE of the game. The simulation results show that our proposed algorithm achieves on average 16.12 % higher sum-rate than the random selection in dynamic uncertained environment.

The remainder of the paper is organized as follows. We describe the multi-channel coexistence network model and formulate the sum-rate maximization problem in Section II. In Section III, we construct the non-cooperative game framework and investigate the properties of the Nash equilibrium. In Section IV, we propose the new learning algorithm and study the expected regret loss compared with the NE points. In Section V, simulation results and discussions are provided. Finally, the conclusions are drawn in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

As shown in Fig. 1, we consider a heterogeneous network where N NR-U UEs share M available unlicensed sub-channels in the 5GHz unlicensed band with other devices (OUEs), like Wi-Fi APs and STAs. All the N NR-U UEs (NUEs) exist within the coverage of the gNB and in the

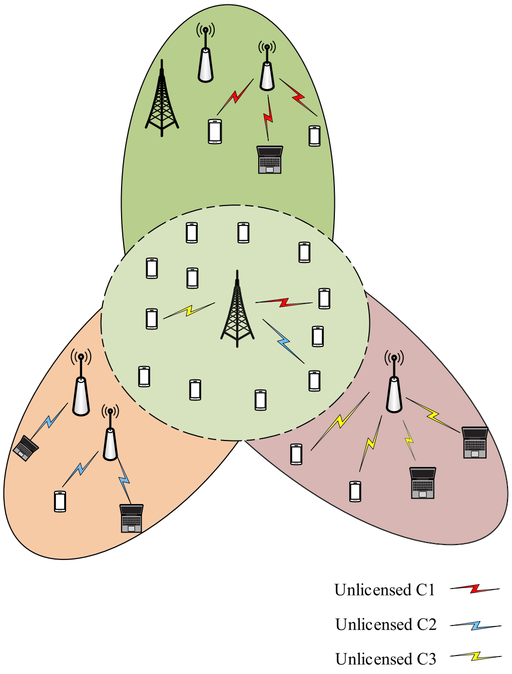


FIGURE 1. The LAA-LTE/Wi-Fi coexistence system.

junction area of the other systems, as depicted in the dotted area. The sub-channels are orthogonal, thus, the devices on different sub-channels do not cause interference. In contrast, the NUEs and OUEs on the same channel compete for uplink transmission opportunities. Due to the different device distributions and channel condition on unlicensed channels, the channel selection of NUEs affect the long-term system performance. The design objective of the coexistence system is to maximize the long-term sum rates of the NUEs.

As typically done in practical systems, we divide the time domain into mini-slots, referred to as "Decision Mini-slot (DMS)" to perform channel selection and data transmission. NUEs dynamically select the channel to access and receive the transmission information for learning to select the channel on the next DMS (see Fig. 2). Without loss of generality, we suppose that the network and channel condition are unchanged within a DMS.

B. PROBLEM FORMULATION

We consider the channel access scheme which involves both CSMA/CA and binary exponential back-off [27] for the OUEs and NUEs to compete for the access to channel.

For the considered dynamic channel access system, we define a_n as the channel selection of NUE n, N_m^w as the number of the OUEs in channel m, $\mathcal{N}_m^g$ as the set of users who select channel m, and N_m^g as the number of NUEs in channel m. Thus, $\mathcal{N}_m^g = \{n \,|\, a_n = m, n \in \mathcal{N}\}$ and $N_m^g = |\mathcal{N}_m^g|$. Then,

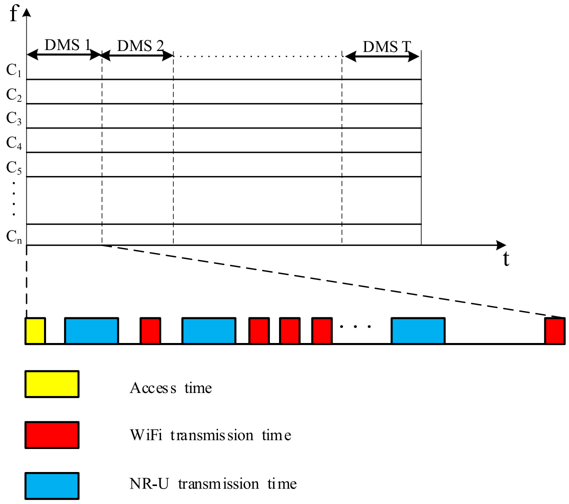


FIGURE 2. The division of time domain.

the normalized transmission rate of user n at the decision mini-slot t, is given by:

$$r_n(t) = f\left(N_m^w, N_m^g\right) r_{mn}(t) \quad st. \quad a_n = m. \tag{1}$$

where $r_{mn}(t)$ is the instantaneous transmission rate of UE n associated with channel m at decision mini-slot t. As depicted in [8] and [9], we use $f\left(N_m^w, N_m^g\right)$ to represent the transmission efficiency function of a NUE, as given in (2), as shown at the bottom of the page, for simplicity, where P_b (the idle probability of the channel), P_s^W (the probability that the channel is occupied by a successful transmission of WiFi), P_s^L (the probability that the channel is occupied by a successful transmission of NR-U), P_c^W (the probability that the channel is occupied by a collision between Wi-Fi transmissions), P_c^L (the probability that the channel is occupied by a collision between NR-U transmissions) and $P_c^{W,L}$ are functions of N_m^w and N_m^g. Besides, T_θ is the duration of an idle slot and T_s^L represents the time duration of a successful transmission of the NR-U users.

Then, the expected rate of user n selecting channel m can be written as

$$r_n = \mathbb{E}\left[r_n(t)\right] = \frac{1}{T}\sum_{t=1}^{T} f\left(N_m^w, N_m^g\right) r_{mn}(t) \tag{3}$$

and the system throughput, which is defined as the aggregate throughput of all the NR-U UEs, is given by

$$\begin{aligned} U_s(\boldsymbol{a}) &= \sum_{n=1}^{N} \mathbb{E}\left[r_n(t)\right] \\ &= \frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T} f\left(N_{a_n}^w, \sum_{j=1}^{N} I(a_j, a_n)\right) r_{a_n n}(t), \end{aligned} \tag{4}$$

$$f\left(N_m^w, N_m^g\right) = \frac{P_s^L T_s^L}{P_f T_\theta + P_s^W T_s^W + P_s^L\left(T_s^L + \delta + DIFS\right) + P_c^W T_c^W + \left(P_c^L + P_c^{W,L}\right)\left(T_s^L + \delta + DIFS\right)} \tag{2}$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_N)$ is a channel selection of the NUEs and $I(a_j, a_n)$ indicates whether user j selects the same channel as user n or not.

$$I(a_j, a_n) = \begin{cases} 1, & \text{if } a_j = a_n; \\ 0, & \text{if } a_j \neq a_n. \end{cases} \tag{5}$$

The objective of the system is to find the optimal selection vector $\boldsymbol{a}_{\text{opt}}$ to maximize the system capacity, i.e. ,

$$\boldsymbol{a}_{\text{opt}} = \arg\max_{\boldsymbol{a}} \ U_s(\boldsymbol{a}). \tag{6}$$

It is challenging to solve problem (6), due to the constraints of dynamic and unknown system environment information.

III. GAME-THEORY DISTRIBUTED CHANNEL SELECTION

Since there is no control center for information exchange, the users can behave selfishly to maximize its individual throughput. This motivates us to formulate a non-cooperative game to find the optimal channel selection $\boldsymbol{a}_{\text{opt}}$. The properties of the game are first investigated.

A. DYNAMIC CHANNEL ACCESS GAME MODEL

The channel selection game in a dynamic environment can be denoted as

$$\mathcal{G} \triangleq \left\{\mathcal{N}, \{a_n\}_{n\in\mathcal{N}}, \{u_n\}_{n\in\mathcal{N}}\right\}, \tag{7}$$

where the NUEs $n \in \mathcal{N}$ are the players in the game; each user's strategy space can be expressed as $\{a_n\} \triangleq \mathcal{A}_n = \{1, 2, \ldots, m\}, \forall n \in \mathcal{N}$ and u_n is the utility of user n defined as its throughput. Similar to (1), user n's utility at DMS t can be expressed as:

$$\begin{aligned} u_n(a_n, \boldsymbol{a}_{-n}, t) &= r_n(t) \\ &= f\left(N^w_{a_n}, \sum_{j=1}^{N} I(a_j, a_n)\right) r_{a_n n}(t) \end{aligned} \tag{8}$$

We assume that each channel supports the same transmission rate among all users. In other words, all channels yield the same bandwidth and same transmission rate for each user, although different users may experience different channel conditions. This condition holds which is hold in some practical systems, e.g., IEEE 802.16d/e standard [28], [29].

The finite rate channel model is applied to represent the time fluctuations, where the achievable transmission rate of each decision mini-slot can only be chosen from the relevant set. Specifically, the rate set on channel m is denoted as $R_m = \{r_{m1}, r_{m2}, \ldots, r_{mK}\}$. The corresponding rate-state probabilities are $P_m = \{p_{m1}, p_{m2}, \ldots, p_{mK}\}$. Then, the expected utility of user n is given by

$$\begin{aligned} u_n(a_n, \boldsymbol{a}_{-n}) &= \frac{1}{T}\sum_{t=1}^{T} r_n(t) \\ &= \frac{1}{T}\sum_{t=1}^{T} f\left(N^w_{a_n}, \sum_{j=1}^{N} I(a_j, a_n)\right) r_{a_n}(t) \\ &= \sum_{k=1}^{K} p_{a_n k} f\left(N^w_{a_n}, \sum_{j=1}^{N} I(a_j, a_n)\right) r_{a_n k}. \end{aligned} \tag{9}$$

In this formulated channel dynamic selection game $\mathcal{G}$, each user acts as a game player aiming to maximize its utility function. Hence, we proceed with the selection strategies of users in the game.

B. ANALYSIS OF NE

In this subsection, we prove that there exist NE points in the channel selection game $\mathcal{G}$.

Definition 1: If and only if no player can improve its utility function by deviating unilaterally, a channel selection profile $\boldsymbol{a}^* = (a_1^*, a_2^*, \ldots, a_N^*)$ is a pure strategy NE [30]:

$$u_n(a_n^*, \boldsymbol{a}^*_{-\boldsymbol{n}}) \geq u_n(a_n, \boldsymbol{a}^*_{-\boldsymbol{n}}) \quad \forall n \in \mathcal{N}, \ a_n \in \mathcal{A}_n \setminus \{a_n^*\} \tag{10}$$

Definition 2: A strategic form game $(\mathcal{N}, (\mathcal{A}_i), (u_i))$ is an ***exact potential game*** *(EPG) if there exists an* ***exact potential function*** *$\phi : \mathcal{A} \rightarrow \mathbb{R}$ such that*

$$u_i\left(a_i, \boldsymbol{a}_{-i}\right) - u_i\left(a_i', \boldsymbol{a}_{-i}\right) = \phi\left(a_i, \boldsymbol{a}_{-i}\right) - \phi\left(a_i', \boldsymbol{a}_{-i}\right) \tag{11}$$

for every $i \in \mathcal{N}$, $a_i, a_i' \in \mathcal{A}_i$, and $\boldsymbol{a}_{-i} \in \boldsymbol{\mathcal{A}}_{-i}$.

In other words, a change of the utility function caused by an arbitrary unilateral action change of a user is the same as that in the exact potential function [31].

Theorem 1: The selection game $\mathcal{G}$ is an exact potential function, which has at least one pure strategy NE point.

Proof: For a channel selection profile $\boldsymbol{a} = (a_n, \boldsymbol{a}_{-\boldsymbol{n}})$, without loss of generality, we assume $a_n = m$. Thus, we have

$$u_n(a_n, \boldsymbol{a}_{-\boldsymbol{n}}) = \sum_{k=1}^{K} p_{mk} f\left(N^w_m, N^g_m\right) r_{mk}\,, \tag{12}$$

where r_{mk} is the transmission rate of UE n associated with channel m and p_{mk} is the corresponding transmission rate probability. Motivated by the potential function in [32] and [33], we construct the potential function as

$$\Phi\left(a_n, \boldsymbol{a}_{-\boldsymbol{n}}\right) = \sum_{m=1}^{M} \sum_{s=1}^{N_m(a_n, \boldsymbol{a}_{-\boldsymbol{n}})} \sum_{k=1}^{K} p_{mk} f\left(N^w_m, s\right) r_{mk}.$$

For illustration convenience, we define

$$\phi^k\left(a_n, \boldsymbol{a}_{-\boldsymbol{n}}\right) = \sum_{m=1}^{M} \sum_{s=1}^{N_m(a_n, \boldsymbol{a}_{-\boldsymbol{n}})} p_{mk} f\left(N^w_m, s\right) r_{mk}.$$

Thus, we have

$$\Phi\left(a_n, \boldsymbol{a}_{-\boldsymbol{n}}\right) = \sum_{k=1}^{K} \phi^k\left(a_n, \boldsymbol{a}_{-\boldsymbol{n}}\right). \tag{13}$$

Now, suppose that an arbitrary player n unilaterally changes its channel selection from a_n to a_n', the corresponding change of player n's utility function can be expressed ($a_n = m$ and $a_n' = m'$)

$$u_n\left(a_n, \boldsymbol{a}_{-\boldsymbol{n}}\right) - u_n\left(a_n', \boldsymbol{a}_{-\boldsymbol{n}}\right)$$

$$\begin{aligned} = & \sum_{k=1}^{K} p_{mk} f\left(N_m^w, N_m^g\left(a_n, \boldsymbol{a_{-n}}\right)\right) r_{mnk} \\ & - \sum_{k=1}^{K} p_{m'k} f\left(N_{m'}^w, N_{m'}^g\left(a_n', \boldsymbol{a_{-n}}\right)\right) r_{m'k}. \end{aligned} \tag{14}$$

The change of $\phi^k\left(a_n, \boldsymbol{a_{-n}}\right)$ is presented on the bottom of this page. Since player n'channel selection change only affects the users associated with channel m and m'. We have

$$\begin{aligned} N_m\left(a_n', \boldsymbol{a_{-n}}\right) &= N_m\left(a_n, \boldsymbol{a_{-n}}\right) - 1; \\ N_{m'}\left(a_n', \boldsymbol{a_{-n}}\right) &= N_{m'}\left(a_n, \boldsymbol{a_{-n}}\right) + 1. \end{aligned}$$

Therefore, the change of $\phi^k\left(a_n, \boldsymbol{a_{-n}}\right)$ can be further rewritten as (15), shown at the bottom of the page. By combining (13) and (16), shown at the bottom of the page, we have the following equation:

$$u_n\left(a_n, \boldsymbol{a_{-n}}\right) - u_n\left(a_n', \boldsymbol{a_{-n}}\right) = \Phi\left(a_n, \boldsymbol{a_{-n}}\right) - \Phi\left(a_n', \boldsymbol{a_{-n}}\right),$$

which always holds for all $n \in \mathcal{N}$, $a_n \in \mathcal{A}_n$, $a_n' \in \mathcal{A}_n$ and $\boldsymbol{a_{-n}} \in \boldsymbol{\mathcal{A}_{-n}}$. It is proved that the channel selection game in the dynamic environment is an exact potential game with potential function Φ, according to Definition 2. As shown in [34], a potential game has at least one pure strategy NE point. As a result, Theorem 1 is proved.

■

IV. DYNAMIC DISTRIBUTED LEARNING ALGORITHM FOR ACHIEVING NE IN UNKNOWN ENVIRONMENT

The formulated dynamic channel selection game is an exact potential game, as stated in Theorem 1, and a large number of algorithms can be employed to achieve the NE, e.g. spital adaptive play [29], best response [34] and no-regret learning [35]. However, these algorithms need to know the complete information of the other users about their choices and rewards at each iteration. Furthermore, the environment needs to be stationary, which means that those algorithms are not suitable for the dynamic selection game. Motivated by decentralized multi-armed bandit (D-MAB) technologies, we propose an online learning distributed channel selection algorithm (OLDCSA) to achieve similar performance to the Nash equilibrium point, where the users can automatically learn from their individual action-reward experience and update their selections. We first extend the UCB for selecting the K-th largest expected reward in Algorithm 1 and present the analysis of the algorithm. Then, the online learning distributed channel selection algorithm (OLDCSA) is shown in Algorithm 2.

Algorithm 1 UCB's Selection of the K-Th Largest Expected Reward

```
// Initialization
for $t = 1$ to $M$ do
    Let $i = t$ and play arm $i$;
    $\hat{\theta}_i(t) = X_i(t)$;
    $n_i(t) = 1$;
end for
// Main Loop
while true do
    $t = t + 1$;
    Play arm $k$ with the $K$-th largest index values in (17)
        $\hat{\theta}_i(t-1) + \sqrt{\frac{2\,ln(t)}{n_i(t-1)}};$   (17)
    $\theta_k(t) = \frac{\theta_k(t-1)*n_k(t-1)+X_k(t)}{n_k(t-1)+1}$;
    $n_k(t) = n_k(t-1) + 1$;
end while
```

A. UCB FOR SELECTING THE K-TH LARGEST EXPECTED REWARD

Consider the classical single player multi-armed bandit problem for M arms. UCB1 can pick the largest arm, while we

$$\begin{aligned} \phi^k\left(a_n, \boldsymbol{a}_{-n}\right) - \phi^k\left(a_n', \boldsymbol{a}_{-n}\right) &= \sum_{m=1}^{M} \sum_{s=1}^{N_m(a_n, \boldsymbol{a}_{-n})} p_{mk} f\left(N_m^w, s\right) r_{mk} - \sum_{m=1}^{M} \sum_{s=1}^{N_m(a_n', a_{-n})} p_{mk} f\left(N_{m'}^w, s\right) r_{m'k} \\ &= \sum_{s=1}^{N_m(a_n, a_{-n})} p_{mk} f\left(N_m^w, s\right) r_{mk} + \sum_{s=1}^{N_{m'}(a_n, a_{-n})} p_{m'k} f\left(N_{m'}^w, s\right) r_{m'k} \\ &\quad - \sum_{s=1}^{N_m(a_n', a_{-n})} p_{mk} f\left(N_m^w, s\right) r_{mk} - \sum_{s=1}^{N_{m'}(a_n', a_{-n})} p_{m'k} f\left(N_{m'}^w, s\right) r_{m'k} \\ &= p_{mk} f\left(N_m^w, N_m\left(a_n, a_{-n}\right)\right) r_{mk} - p_{m'k} f\left(N_{m'}^w, N_{m'}\left(a_n', a_{-n}\right)\right) r_{m'k} \end{aligned} \tag{15}$$

$$\begin{aligned} \Phi\left(a_n, a_{-n}\right) - \Phi\left(a_n', a_{-n}\right) &= \sum_{k=1}^{K} \left(\phi^k\left(a_n, a_{-n}\right) - \phi^k\left(a_n', a_{-n}\right)\right) \\ &= \sum_{k=1}^{K} p_{mk} f\left(N_m^w, N_m\left(a_n, a_{-n}\right)\right) r_{mk} - \sum_{k=1}^{K} p_{m'k} f\left(N_{m'}^w, N_{m'}\left(a_n', a_{-n}\right)\right) r_{m'k} \\ &= u_n\left(a_n, a_{-n}\right) - u_n\left(a_n', a_{-n}\right) \end{aligned} \tag{16}$$

Algorithm 2 Distributed Learning Channel Selection Algorithm for M Channels and N Users

```
// Initialization Step 1
Random give each user a unique number index $l$, which is in range $[1, N]$
for $t = 1$ to $M$ do
    $n = t \ mod \ N$;
    $m = t/N$;
    Let user $l \ (l \leq n)$ select channel $m$ and automatically run LBT mechanism for transmission;
    The gNB broadcasts the information of channel index $m$, the number of users $n$ and successful transmission probability $p_{m,n}$, the normalized utility $u_{m,n}$.
    $\hat{\theta}_{m,n}(t) = u_{m,n}(t)$;
    $c_{m,n}(t) = 1$;
end for
// Main Loop
while true do
    $t = t + 1$;
    users parallel select the channel according to the policy specified in Algorithm 3.
end while
```

extend it for selecting the K-th largest arm, as shown in the Algorithm 1.

At the initialization step, the user plays each arm once to have the initial value of $\hat{\theta}_i$ and n_i. The algorithm requires two $1 \times M$ vectors to store the information after the user plays an arm at each slot.

Now, we present some of the above algorithm which will be useful in our analysis of subsequent algorithms in the following. For illustration convenience, we denote K as the arm with K-th largest expected reward and $T_i(n)$ as the number of times that the user plays arm i $(i \neq K)$ after n time slots.

Theorem 2: Under the selection policy proposed in Algorithm 2, the upper bound of $\mathbb{E}[T_i(n)]$ *is*

$$\frac{8 \ln n}{\Delta_{K,i}^2} + 1 + \frac{\pi^2}{3},$$

where $\Delta_{K,i} = |\theta_K - \theta_i|$, θ_K *is the* K*-th largest expected reward and* θ_i *is the expected reward of arm i.*

Fact 1: Chernoff-Hoeffding bound

Let $X_1, \ldots, X_n$ be random variables with common range $[0, 1]$, such that $\mathbb{E}[X_n | X_1, \ldots, X_{n-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then, for all $a \neq 0$,

$$\mathbb{P}\{S_n \geq n\mu + a\} \leq e^{-2a^2/n},$$
$$\mathbb{P}\{S_n \leq n\mu + a\} \leq e^{-2a^2/n}.$$

Proof: Denote $\overline{\theta}_{i,s_i}$ as the sample mean reward of arm i when it is picked up for s_i times and θ_i is the expected reward of arm i. Then, define $\mathcal{A}_K^*$ as the set of K arms with the K largest expected rewards and $\mathcal{A}_K(t)$ as the set of K arms with the K largest observed expected rewards at time t.

More generally, we denote $I_i(n)$ as the indicator function which is equal to 1 if arm i is selected at time slot n. Let l be an arbitrary positive integer. Then, for arm i which is not the desire arm:

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=M+1}^{n} I_i(t) \\ &\leq l + \sum_{t=M+1}^{n} \{I_i(t), T_i(t-1) \neq l\}. \end{aligned} \tag{18}$$

In the case of $\theta_i < \theta_K$, arm i is picked up at time t and therefore there must exist an arm $j, j \in \mathcal{A}_K^*$ and $j \notin \mathcal{A}_K(t)$. We can thus have the following inequality:

$$\overline{\theta}_{j,T_j(t-1)} + \sqrt{\frac{2 \ln t}{T_j(t-1)}} \leq \overline{\theta}_{i,T_i(t-1)} + \sqrt{\frac{2 \ln t}{T_i(t-1)}}. \tag{19}$$

Let $C_{t,s} = \sqrt{(2 \ln t)/s}$ and we rewrite (19) as:

$$\overline{\theta}_{j,T_j(t-1)} + C_{t,T_j(t-1)} \leq \overline{\theta}_{i,T_i(t-1)} + C_{t,T_i(t-1)}. \tag{20}$$

Then, we have

$$\begin{aligned} T_i(n) &\leq l + \sum_{t=M+1}^{n} \{I_i(t), T_i(t-1) \geq l\} \\ &\leq l + \Bigg\{ \sum_{t=M+1}^{n} \overline{\theta}_{j,T_j(t-1)} + C_{t,T_j(t-1)} \\ &\leq \overline{\theta}_{i,T_i(t-1)} + C_{t,T_i(t-1)}, T_i(t-1) \geq l \Bigg\} \\ &\leq l + \sum_{t=M+1}^{n} \Bigg\{ \min_{0<n_j<t} \overline{\theta}_{j,n_j} + C_{t,n_j} \\ &\leq \max_{l<n_i<t} \overline{\theta}_{i,n_i} + C_{t,n_i} \Bigg\} \\ &\leq l + \sum_{t=1}^{\infty} \sum_{n_j=1}^{t-1} \sum_{n_i=l}^{t-1} \Big\{ \overline{\theta}_{j,n_j} + C_{t,n_j} \leq \overline{\theta}_{i,n_i} + C_{t,n_i} \Big\}. \end{aligned} \tag{21}$$

Now, we observe that $\overline{\theta}_{j,n_j} + C_{t,n_j} \leq \overline{\theta}_{i,n_i} + C_{t,n_i}$. Therefore, at least one of the following events must hold

$$A := \left\{ \overline{\theta}_{j,n_j} \leq \theta_j - C_{t,n_j} \right\}, \tag{22}$$

$$B := \left\{ \overline{\theta}_{i,n_i} \geq \theta_i + C_{t,n_i} \right\}, \tag{23}$$

$$C := \left\{ \theta_j \leq \theta_i + 2C_{t,n_i} \right\}, \tag{24}$$

We bound the probability of events A and B using Fact 1 (Chernoff-Hoeffding bound):

$$\mathbb{P}\left\{ \overline{\theta}_{j,n_j} \leq \theta_j - C_{t,n_j} \right\} \leq e^{-4 \ lnt} = t^{-4} \tag{25}$$

$$\mathbb{P}\left\{ \overline{\theta}_{i,n_i} \geq \theta_i + C_{t,n_i} \right\} \leq e^{-4 \ lnt} = t^{-4} \tag{26}$$

For $l = \lceil (8\ ln\ n)\,/\Delta^2_{K,i} \rceil$, event C is false. In fact,

$$\begin{aligned}\theta_j - \theta_i + 2C_{t,n_i} &\geq \theta_K - \theta_i - 2\sqrt{\frac{2\ ln\ t}{ln\ n_i}}\\ &\geq \theta_K - \theta_i - 2\sqrt{\frac{2\Delta^2_{K,i}\ ln\ t}{8\ ln\ n}}\\ &\geq \theta_K - \theta_i - \Delta_{K,i} = 0. \qquad (27)\end{aligned}$$

for $n_i \geq (8\ ln\ n))/\Delta^2_{K,i}$.
Hence, from bounds (25) and (26), we obtain

$$\begin{aligned}\mathbb{E}\left[T_i(n)\right] \leq \lceil (8\ ln\ n)\,/\Delta^2_{K,i} \rceil + \sum_{t=1}^{\infty}\sum_{n_j=1}^{t-1}\sum_{n_i=\lceil (8\ ln\ n)/\Delta^2_{K,i}\rceil}^{t-1}\\ \Big(\mathbb{P}(A) + \mathbb{P}(B)\Big)\\ \leq (8\ ln\ n)\,/\Delta^2_{K,i} + 1 + \sum_{t=1}^{\infty}\sum_{n_j=1}^{t-1}\sum_{n_i=1}^{t-1} t^{-4}\\ \leq (8\ ln\ n)\,/\Delta^2_{K,i} + 1 + \frac{\pi^2}{3}. \qquad (28)\end{aligned}$$

In the case of $\theta_i > \theta_K$, we denote $\mathcal{A}^*_{K-1}$ as the set of arms whose expected rewards are over θK. Similarly to the case of $\theta_i < \theta_K$, when Arm i is picked up at time t, there must exist an arm $\Phi \in \mathcal{A}_{K-1}(t)$, which is not in the $\mathcal{A}^*_{K-1}$. Thus, we can thus have the following inequality:

$$\overline{\theta}_{i,T_i(t-1)} + C_{t,T_i(t-1)} \leq \overline{\theta}_{\Phi,T_\Phi(t-1)} + C_{t,T_\Phi(t-1)}. \qquad (29)$$

According to $\theta_i \leq \theta_K$, we have

$$\begin{aligned}T_i(n) &= 1 + \sum_{t=M+1}^{n} I_i(t)\\ &\leq l + \sum_{t=M+1}^{n} \{I_i(t), T_i(t-1) \geq l\}\\ &\leq l + \sum_{t=1}^{\infty}\sum_{n_i=1}^{t-1}\sum_{n_\Phi=l}^{t-1} \left\{\overline{\theta}_{i,n_i} + C_{t,n_i} \leq \overline{\theta}_{\Phi,n_\Phi} + C_{t,n_\Phi}\right\}.\end{aligned} \qquad (30)$$

Note that the event $\left\{\overline{\theta}_{i,n_i} + C_{t,n_i} \leq \overline{\theta}_{\Phi,n_\Phi} + C_{t,n_\Phi}\right\}$ implies at least one of the events must occur:

$$A := \left\{\overline{\theta}_{i,n_i} \leq \theta_i - C_{t,n_i}\right\}, \qquad (31)$$

$$B := \left\{\overline{\theta}_{\Phi,n_\Phi} \geq \theta_\Phi + C_{t,n_\Phi}\right\}, \qquad (32)$$

$$C := \left\{\theta_i \leq \theta_\Phi + 2C_{t,n_\Phi}\right\}, \qquad (33)$$

Now, using the Chernoff-Hoeffding bound, we obtain

$$\begin{aligned}\mathbb{P}\left\{\overline{\theta}_{i,n_i} \leq \theta_i - C_{t,n_i}\right\} \leq e^{-4\ lnt} = t^{-4},\\ \mathbb{P}\left\{\overline{\theta}_{\Phi,n_\Phi} \geq \theta_\Phi + C_{t,n_\Phi}\right\} \leq e^{-4\ lnt} = t^{-4}.\end{aligned}$$

Also, for $l = \lceil (8\ ln\ n)\,/\Delta^2_{K,i} \rceil$, the event C cannot happen. In fact, for $n_\Phi \geq (8\ ln\ n))/\Delta^2_{K,i}$,

$$\begin{aligned}\theta_i - \theta_\Phi - 2C_{t,n_\Phi} &\geq \theta_i - \theta_K - 2\sqrt{\frac{2\ ln\ t}{ln\ n_\Phi}}\\ &\geq \theta_i - \theta_K - 2\sqrt{\frac{2\Delta^2_{K,i}\ ln\ t}{8\ ln\ n}}\\ &\geq \theta_i - \theta_k - \Delta_{K,i} = 0. \qquad (34)\end{aligned}$$

By taking expectation on both sides, we obtain

$$\begin{aligned}&\mathbb{E}\left[T_i(n)\right]\\ &\leq \lceil (8\ ln\ n)\,/\Delta^2_{K,i} \rceil + \sum_{t=1}^{\infty}\sum_{n_i=1}^{t-1}\sum_{n_\Phi=\lceil \frac{(8\ ln\ n)}{\Delta^2_{K,i}}\rceil}^{t-1} \Bigg(\\ &\qquad \mathbb{P}\left\{\overline{\theta}_{i,n_i} \leq \theta_i - C_{t,n_i}\right\} + \mathbb{P}\left\{\overline{\theta}_{\Phi,n_\Phi} \geq \theta_\Phi + C_{t,n_\Phi}\right\}\Bigg)\\ &\leq \frac{(8\ ln\ n)}{\Delta^2_{K,i}} + 1 + \sum_{t=1}^{\infty}\sum_{n_i=1}^{t-1}\sum_{n_\Phi=1}^{t-1} t^{-4}\\ &\leq \frac{(8\ ln\ n)}{\Delta^2_{K,i}} + 1 + \frac{\pi^2}{3}. \qquad (35)\end{aligned}$$

Hence, in both cases of $\theta_i < \theta_K$ and $\theta_i > \theta_K$, we have

$$\mathbb{E}\left[T_i(n)\right] \leq \frac{(8\ ln\ n)}{\Delta^2_{K,i}} + 1 + \frac{\pi^2}{3}. \qquad (36)$$

The performance under selection policy $\pi = \{\pi(t)\}_{t=1}^{\infty}$ is measured by the regret $R^\pi_n(\Theta)$, defined as absolute difference between the expected total reward under the policy and the ideal scenario that Θ is known to the player (thus the Kth largest arm is played at each time). Thus, the regret can be expressed as:

$$R^\pi_n(\Theta) = \left|n\theta_K - \mathbb{E}_\pi\left[\sum_{t=1}^{n} S_{\pi(t)}(t)\right]\right|, \qquad (37)$$

where $S_{\pi(t)}(t)$ is the random reward obtained at time t under action $\pi(t)$, and $\mathbb{E}_\pi[.]$ denotes the expectation with respect to policy π.

Corollary 1: The expected regret of the policy UCB for selecting the K-th largest expected reward after any n plays is at most

$$8\sum_{i:\theta_i\neq\theta_K}\frac{ln\ n}{\Delta_{K,i}} + \left(1 + \frac{\pi^2}{3}\right)\sum_{i:\theta_i\neq\theta_K}\Delta_{K,i}. \qquad (38)$$

Proof:

$$\begin{aligned}R^\pi_n(\Theta) &= \left|n\theta_K - \mathbb{E}_\pi\left[\sum_{t=1}^{n} S_{\pi(t)}(t)\right]\right|\\ &\leq \sum_{t=1}^{n}\left|\theta_K - \mathbb{E}_\pi\left[S_{\pi(t)}(t)\right]\right|\\ &= \sum_{i:\theta_i\neq\theta_K}\Delta_{K,i}\mathbb{E}_\pi\left[T_i(n)\right]\\ &\leq 8\sum_{i:\theta_i\neq\theta_K}\frac{ln\ n}{\Delta_{K,i}} + \left(1 + \frac{\pi^2}{3}\right)\sum_{i:\theta_i\neq\theta_K}\Delta_{K,i} \qquad (39)\end{aligned}$$

Algorithm 3

```
user index $l$ select the channel $m$ with the $l$-th largest index values in (40), get the number of users $k$ on channel $m$ and normalized utility $u_{m,k}(t)$.
    $\hat{\theta}^l_{m,n}(t-1) + \sqrt{\frac{2\,ln\,(t)}{c^l_{m,n}(t-1)}};$    (40)
for $n = 1$ to $N$ do
    $\theta^l_{m,n}(t) = \frac{\theta^l_{m,n}(t-1)*c^l_{m,n}(t-1)+u_{m,k}(t)*p_{m,n}/p_{m,k}}{c^l_{m,n}(t-1)+1};$
    $c^l_{m,n}(t) = c^l_{m,n}(t-1) + 1;$
end for
```

B. DISTRIBUTED LEARNING CHANNEL SELECTION ALGORITHM

Based on the UCB policy, we propose an OLDCSA, as shown in Algorithm 2. In the proposed algorithm, we assume that the players can know the number of users who select the same channel, which can obtain by listening the channel. Every user runs the OLDCSA to select the channel and exchange the information at the access phase, which can be seen as the cost of the algorithm.

It is noted that the value of $\hat{\theta}^l_{m,n}$ supports in [0, 1] under the constraint of the UCB algorithm. Thus, we denote $\hat{\theta}^l_{m,n} = \hat{r}^l_{m,n}$, which is defined as the normalized transmission rate. We sort $\hat{\theta}^l_{m,n}$ in ascending order as $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$. As depicted in the algorithm, the users store two $1 \times MN$ vectors, $(\hat{\theta}^l_{m,n})_{1*MN}$ and $(c^l_{m,n})_{1*MN}$. If a user plays the K-th arm, it will select the corresponding channel $m \left(\hat{\theta}^l_{m,n} == \hat{\theta}^l_K\right)$.

To clearly evaluate system performance under the policy π, we define the regret loss by time n, as given by

$$Rl^{\pi}_n(\Theta) = n\theta_K - \mathbb{E}_{\pi}\left[\sum_{t=1}^{n} S_{\pi(t)}(t)\right]. \tag{41}$$

Theorem 3: The expected regret loss under the OLDCSA policy specified in Algorithm 3 is at most

$$\sum_{n=1}^{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{MN}\mathbb{I}(i \neq j)\Big(8\,\frac{ln\ t}{\Delta^2_{i,j}} + 1 + \frac{\pi^2}{3}\Big)\theta_n\right) \tag{42}$$

Proof: For user n, the regret loss is upper bound by

$$\begin{aligned} Rl^{\pi}_t(\Theta, n) &\leq \Big(\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}[T_{i,j}(t)]\Big)\theta_n \\ &\leq \left(\sum_{i=1}^{N}\sum_{j\neq i}\left(8\,\frac{ln\ t}{\Delta^2_{i,j}} + 1 + \frac{\pi^2}{3}\right)\right)\theta_n \end{aligned} \tag{43}$$

Thus, the upper bound for the regret loss is given by

$$\begin{aligned} Rl^{\pi}_t(\Theta) &= \sum_{n=1}^{N} R^{\pi}_t(\Theta, n) \\ &\leq \sum_{n=1}^{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{MN}\mathbb{I}(i \neq j)\Big(8\,\frac{ln\ t}{\Delta^2_{i,j}} + 1 + \frac{\pi^2}{3}\Big)\theta_n\right). \end{aligned} \tag{44}$$

From (44), the upper bound of the regret loss under Algorithm 2 grows as $O\left(MN^3 ln\,t\right)$.

Although we analyze and propose the access approach for the basic coexistence mechanism, however, the proposed algorithm does not rely on the specific access mechanism and specified model of the environment. The users can automatically learn the channel condition and make decisions according to its historical knowledge and experience. Also, the proposed algorithm can be adapted to wireless access point selection of a Wi-Fi network.

V. SIMULATION RESULTS AND DISCUSSIONS

In this section, the performance of the proposed algorithms is evaluated by numerical simulations. With the help of adaptive modulation and coding (ACM), the channel transmission rate can be classified into several states according to the received average signal-to-noise ratio (SNR). The PHY layer modes with different coding and modulation schemes in [36] are applied in the simulation study, in which the channel rate set is {6, 12, 18, 36, 54} Mbps. According to [37], the state classification and corresponding probabilities are jointly determined by the average received SNR γ and the target packet error rate p_e. We adopt Rayleigh fading channel model in the simulation. Taking $\gamma = 5$ dB and $p_e = 10^{-3}$ as an example, the state probabilities are given by $\pi = \{0.3376, 0.2348, 0.2517, 0.1757, 0.002\}$.

We deploy a number of WAPs randomly in the coverage area of an NR-U BS. The WAPs operate on the different orthogonal channels over the 5 GHz band and the NR-U users share the same unlicensed bands with the Wi-Fi users. The NR-U users and Wi-Fi users are randomly distributed in the coverage of the NR-U BS and WAPs, respectively. In our simulations, the number of Wi-Fi users per WAP is randomly and uniformly distributed between 1 and 10. The simulation parameters of the NR-U users are shown in Table 1 and the Wi-Fi parameters are chosen based on [27]. At the beginning, we consider the effects of user selections on transmission efficiency and the performance of UCB-K algorithm. Then, we show the results obtained by OLDCSA algorithm, examining the performance of the proposed algorithms.

TABLE 1. Value of the simulation parameters.

Parameters	Value
Carrier frequency	5 GHz
Sampling rate (f_s)	20 MHz
Useful symbol duration (T_U)	3.2 µs
Guard interval duration (T_g)	0.8 µs
Total symbol duration (T_{Total})	4.0 µs
NR-U transmission time duration (T^L_s)	5.0 ms
Number of data subcarriers (N_D)	48
Number of pilot subcarriers (N_p)	48
Subcarrier spacing (Δ_f)	0.3125 MHz
Bandwidth of each unlicensed channel	16.875 MHz
Channel rate set of NR-U user	$\{6, 12, 18, 36, 54\}$

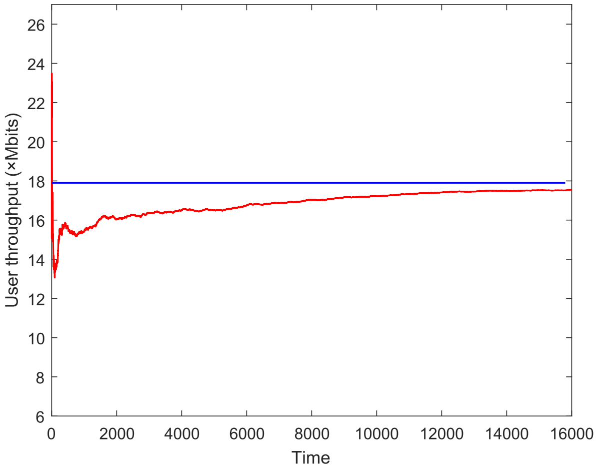


FIGURE 3. User throughput vs time slots under the UCB-K policy.

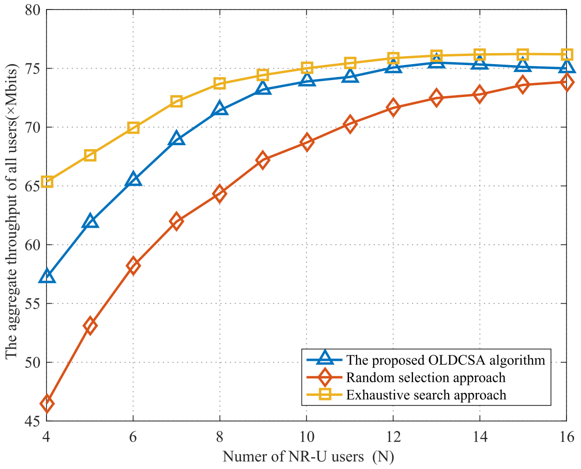


FIGURE 5. The comparison results of different channel selection schemes for aggregate throughput in the coexistence system.

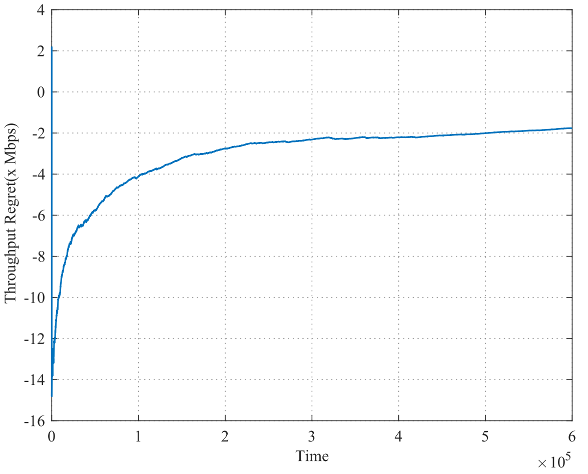


FIGURE 4. The system regret vs time slots under the OLDCSA algorithm.

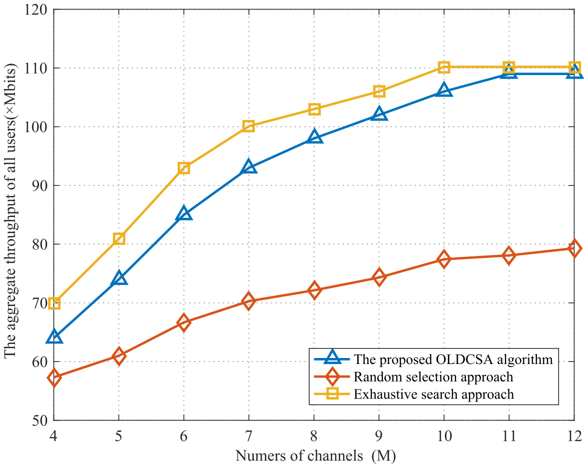


FIGURE 6. The sum data rate of NR-U for different unlicensed channels.

For an arbitrarily chosen user, we study the performance of algorithm UCB-K and show the simulation results in Fig. 3. As expected, the user throughput of UCB-K algorithm converges to the K-th largest expected reward. Specifically, there are $M = 4$ channels with the average received SNR being 5 dB, 6 dB, 7 dB and 8 dB.

In Fig. 5, we evaluate the throughput of the proposed OLDCSA algorithm with an increasing number of NUEs and compare with some other access approaches. For comparison, we independently simulate 1000 trials and take the average results for the random access approaches. It is noted that the the aggregate capacity increases with the number of users when the number of users is small. However, the growth is negligible when the number is large. The reason is that the efficiency grows with the increasing number of NUEs for each unlicensed channel. The access efficiency of all NUEs first grows, then remains almost unchanged, and even degrades due to collisions. It is also noted that when the number of NUEs becomes large, the throughput gap between the random selection approach and the exhaustive algorithm decreases. The reasons are: 1) when the number is small, the users are spread widely over different channels with the proposed multi-agent learning algorithm or exhaustive approach. On the contrary, there is a certain probability that some channels are congested, while the other channels are not selected by any user with the random access approach. 2) when the user number is large enough, the efficiency of the users at every unlicensed channel almost remains unchanged, even degrades in some channels. Thus, the performance gap between the access approaches decreases. In addition, no information exchange in the coexistence system results in the capacity gap between the proposed OLDCSA algorithm and exhaustive approach. Also, the expected system regret loss between the proposed selection policy and the NE points is shown in Fig. 4 when the number of users is $N = 6$.

Fig. 6 shows the sum-data rate of all NR-U users for different numbers of unlicensed channels when the user number is $N = 6$. As shown in Fig. 6, the sum-data rate of the NR-U network increases with an increasing number of available unlicensed channels. When $M > N$, the growth slows down

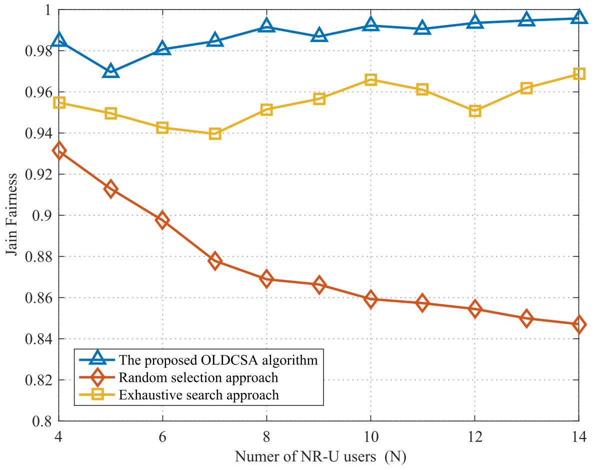


FIGURE 7. Fairness among NR-U users with different selection schemes.

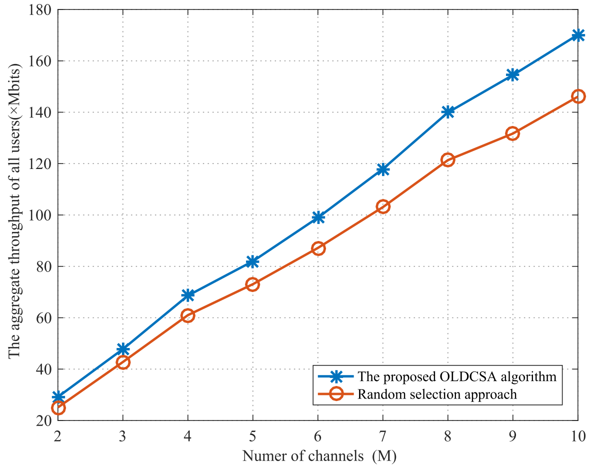


FIGURE 8. The comparison results between the proposed OLDCSA algorithm and random approach with fixed $N/M = 2$.

and the gap increases between the random selection policy and the proposed OLDCSA algorithm, as the number of channels M increases.

Besides the system throughput, fairness is another important indicator in wireless communication systems. Typically, we measure fairness by using Jain's fairness index (JFI) [38]. The JFI is defined by

$$J\left(r_1, \ldots, r_N\right) = \frac{\left(\sum_{n=1}^{N} r_n\right)^2}{N \sum_{n=1}^{N} r_n^2}, \tag{45}$$

where $r_n \leq 0, \forall n \in \mathcal{N}$, denotes the expected throughput obtained by the user n and $J\left(r_1, \ldots, r_N\right) \in [0, 1]$. A higher JFI implies that the system is more fair.

The JFI of the three channel selection schemes are shown in Fig. 7. It is shown that the proposed OLDCSA algorithm provides the best fairness and the random selection approach achieves a poor performance as it overlooks the difference among channel environment. Moreover, the NR-U users are spread over different channels under the random channel selection scheme and achieve good fairness when the number of users is small.

In light of the tendency of dense network in the future, the proposed learning algorithm is evaluated in Fig. 8 for dense networks with a fixed ratio of the number of users to the number of channels $N/M = 2$. The average received SNR of the channels are randomly selected from [5 dB, 6 dB, 7 dB, 8 dB, 9 dB]. The aggregate throughput of the network increases nearly linearly with the number of channels under both strategies.

VI. CONCLUSION

The traffic offloading of NR-U in the unlicensed spectrum is a promising enhancement to meet the requirements for future systems. How to address the coexistence between different access multiple radio access technologies and increase the network capacity on unlicensed bands have been the key challenges and issues. In this framework, we investigated the dynamic frequency selection problem for the WLAN and NR-U heterogeneous networks with a time-varying and unknown environment. We formulated the spectrum access problem as a non-cooperative game and proved that it is a potential game. OLDCSA based on multi-armed bandit learning was proposed to solve the problem, which allows the users to automatically learn online from their individual action-reward history and update their selections. The proposed algorithm does not rely on the specific access mechanism and specified model of the environment, and can be adapted to the unknown dynamic environment. Finally, the simulations verified the performance of the proposed algorithm.

REFERENCES

[1] Cisco. (Mar. 2020). *Cisco Annual Internet Report (2018–2023) White Paper*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] A. Mukherjee, J.-F. Cheng, S. Falahati, H. Koorapaty, D. H. Kang, R. Karaki, L. Falconetti, and D. Larsson, "Licensed-assisted access LTE: Coexistence with IEEE 802.11 and the evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50–57, Jun. 2016.

[3] X. Lu, V. Petrov, D. Moltchanov, S. Andreev, T. Mahmoodi, and M. Dohler, "5G-U: Conceptualizing integrated utilization of licensed and unlicensed spectrum for future IoT," *IEEE Commun. Mag.*, vol. 57, no. 7, pp. 92–98, Jul. 2019.

[4] Y. Yuan, Y. Zhao, B. Zong, and S. Parolari, "Potential key technologies for 6G mobile communications," *Sci. China Inf. Sci.*, vol. 63, no. 8, pp. 1–19, Aug. 2020.

[5] *Study On Licensed-Assisted Access To Unlicensed Spectrum (Release 13)*, document TR 36.889 V13.0.0, 3GPP, Jun. 2015.

[6] *Evolved Universal Terrestrial Radio Access (E-Utra); Physical Layer Procedures (Release 14)*, document TS 36.213 V14.0.0, 3GPP, Sep. 2016.

[7] *Study on NR-Based Access To Unlicensed Spectrum (Release 16)*, document TR 38.889 V1.0.0, 3GPP, Nov. 2018.

[8] Y. Gao, X. Chu, and J. Zhang, "Performance analysis of LAA and WiFi coexistence in unlicensed spectrum based on Markov chain," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[9] J. Tan, S. Xiao, S. Han, Y.-C. Liang, and V. C. M. Leung, "QoS-aware user association and resource allocation in LAA-LTE/WiFi coexistence systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2415–2430, Apr. 2019.

[10] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "LBT-based adaptive channel access for LTE-U systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6585–6597, Oct. 2016.

[11] R. Yin, Y. Zhang, F. Dong, A. Wang, and C. Yuen, "Energy efficiency optimization in LTE-U based small cell networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1963–1967, Feb. 2019.

[12] Q. Chen, G. Yu, A. Maaref, G. Y. Li, and A. Huang, "Rethinking mobile data offloading for LTE in unlicensed spectrum," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4987–5000, Jul. 2016.

[13] Q. Zhang, Q. Wang, Z. Feng, and T. Yang, "Design and performance analysis of a fairness-based license-assisted access and resource scheduling scheme," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2968–2980, Nov. 2016.

[14] A. Galanopoulos, F. Foukalas, and T. A. Tsiftsis, "Efficient coexistence of LTE with WiFi in the licensed and unlicensed spectrum aggregation," *IEEE Trans. Cognit. Commun. Netw.*, vol. 2, no. 2, pp. 129–140, Jun. 2016.

[15] Q. Cui, Y. Gu, W. Ni, and R. P. Liu, "Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: From theory to application," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1754–1767, Aug. 2017.

[16] Q. Chen, G. Yu, R. Yin, A. Maaref, G. Y. Li, and A. Huang, "Energy efficiency optimization in licensed-assisted access," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 723–734, Apr. 2016.

[17] V. Sathya, S. M. Kala, M. I. Rochman, M. Ghosh, and S. Roy, "Standardization advances for cellular and Wi-Fi coexistence in the unlicensed 5 and 6 GHz bands," *GetMobile, Mobile Comput. Commun.*, vol. 24, no. 1, pp. 5–15, Aug. 2020.

[18] A. M. Baswade, V. Sathya, B. R. Tamma, and A. Franklin, "Unlicensed carrier selection and user offloading in dense LTE-U networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[19] A. K. Bairagi, N. H. Tran, W. Saad, Z. Han, and C. S. Hong, "A game-theoretic approach for fair coexistence between LTE-U and Wi-Fi systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 442–455, Jan. 2019.

[20] A. K. Bairagi, S. F. Abedin, N. H. Tran, D. Niyato, and C. S. Hong, "QoE-enabled unlicensed spectrum sharing in 5G: A game-theoretic approach," *IEEE Access*, vol. 6, pp. 50538–50554, 2018.

[21] Y. Su, X. Du, L. Huang, Z. Gao, and M. Guizani, "LTE-U and Wi-Fi coexistence algorithm based on Q-learning in multi-channel," *IEEE Access*, vol. 6, pp. 13644–13652, 2018.

[22] M. Alsenwi, I. Yaqoob, S. R. Pandey, Y. K. Tun, A. K. Bairagi, L.-W. Kim, and C. S. Hong, "Towards coexistence of cellular and WiFi networks in unlicensed spectrum: A neural networks based approach," *IEEE Access*, vol. 7, pp. 110023–110034, 2019.

[23] Z. Guan and T. Melodia, "CU-LTE: Spectrally-efficient and fair coexistence between LTE and Wi-Fi in unlicensed bands," in *Proc. IEEE INFOCOM-35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[24] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4674–4689, Jul. 2018.

[25] Q. Cui, Z. Gong, W. Ni, Y. Hou, X. Chen, X. Tao, and P. Zhang, "Stochastic online learning for mobile edge computing: Learning from changes," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 63–69, Mar. 2019.

[26] J. Jiang, Y. Li, L. Chen, J. Du, and C. Li, "Multitask deep learning-based multiuser hybrid beamforming for mm-wave orthogonal frequency division multiple access systems," *Sci. China Inf. Sci.*, vol. 63, no. 8, Aug. 2020.

[27] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[28] J. Elias, F. Martignon, A. Capone, and E. Altman, "Competitive interference-aware spectrum access in cognitive radio networks," in *Proc. 8th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw.*, May/Jun. 2010, pp. 85–90.

[29] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 2, pp. 180–194, Apr. 2012.

[30] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[31] K. Yamamoto, "A comprehensive survey of potential game approaches to wireless networks," *IEICE Trans. Commun.*, vol. 98, no. 9, pp. 1804–1823, 2015.

[32] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380–1391, Apr. 2012.

[33] B. Vöcking and R. Aachen, "Congestion games: Optimization in competition," in *Proc. ACiD*, 2006, pp. 9–20.

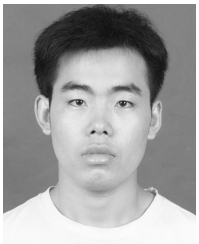
[34] D. Monderer and L. S. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, no. 1, pp. 124–143, 1996.

[35] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," *Mobile Netw. Appl.*, vol. 11, no. 6, pp. 779–797, Dec. 2006.

[36] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 172–180, May 2002.

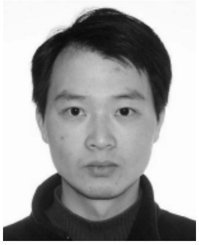
[37] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.

[38] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," 1998, *arXiv:cs/9809099*. [Online]. Available: https://arxiv.org/abs/cs/9809099



YANPENG SHI received the B.S. degree in communications engineering from Xidian University, China, in 2018. He is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include stochastic optimization theory, reinforcement learning, and their applications to 5G heterogeneous wireless networks.

QIMEI CUI (Senior Member, IEEE) received the B.E. and M.S. degrees in electronic engineering from Hunan University, Changsha, China, in 2000 and 2003, respectively, and the Ph.D. degree in information and communications engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2006. She has been a Full Professor with the School of Information and Communication Engineering, BUPT, since 2014. She was a Guest Professor with the Department of Electronic Engineering, University of Notre Dame, IN, USA, in 2016. Her main research interests include spectral efficiency or energy efficiency-based transmission theory, and networking technology for 4G/5G broadband wireless communications and green communications. She serves as a Technical Program Committee Member for several international conferences, such as IEEE ICC, IEEE WCNC, IEEE PIMRC, IEEE ICCC, IEEE ISCIT 2012, and IEEE WCSP 2013. She was a recipient of the Honorable Mention Demo Award at the ACMMobiCom 2009, the Best Paper Award at the IEEE ISCIT 2012 and the IEEE WCNC 2014, and the Young Scientist Award at the URSI GASS 2014. She serves as the Guest Editor for the *EURASIP Journal on Wireless Communications and Networking*, *International Journal of Distributed Sensor Networks*, and *Journal of Computer Networks and Communications*.



WEI NI (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. He was a Postdoctoral Research Fellow with Shanghai Jiaotong University, from 2005 to 2008, the Deputy Project Manager of the Bell Labs R&I Center, Alcatel/Alcatel-Lucent, from 2005 to 2008, and a Senior Researcher at Devices Research and Development, Nokia, from 2008 to 2009. He is currently the Senior Scientist, the Team Leader, and the Project Leader of the Commonwealth Scientific and Industrial Research Organization, Australia. He also holds adjunct positions at the University of New South Wales, Macquarie University, and the University of Technology Sydney. His research interests include stochastic optimization, game theory, graph theory, and their applications to network and security. He has served as the Track Chair for the VTC-Spring 2017, the Track Co-Chair for the IEEE VTC-Spring 2016, and the Publication Chair for the BodyNet 2015. He also served as the Student Travel Grant Chair for WPMC 2014, a Program Committee Member for CHINACOM 2014, and a TPC Member for IEEE ICC 2014, ICCC 2015, EICE 2014, and WCNC 2010. He has been serving as an Editor for the *Journal of Engineering* (Hindawi) since 2012 and a Secretary for the IEEE NSW VTS Chapter since 2015.

ZESONG FEI (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the Beijing Institute of Technology (BIT), in 2004. He is currently a Professor with the Research Institute of Communication Technology, BIT, where he is also involved in the design of the next-generation high-speed wireless communication. He has published more than 50 articles in the IEEE journals. His research interests include wireless communications and multimedia signal processing. He is also a Senior Member of the Chinese Institute of Electronics and the China Institute of Communications. He is also the Chief Investigator of the National Natural Science Foundation of China. He serves as the Lead Guest Editor for *Wireless Communications and Mobile Computing* and *China Communications*, Special Issue on Error Control Coding.