Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and Beyond

Susanna Mosleh[†§], Yao Ma[‡], Jacob D. Rezac[‡], and Jason B. Coder[‡]

[†]Associate, Communications Technology Laboratory, National Institute of Standards and Technology, USA
[§]Department of Physics, University of Colorado, Boulder, Colorado, USA
[‡]Communications Technology Laboratory, National Institute of Standards and Technology, USA

***Abstract*—Dynamic spectrum access (DSA) to achieve spectrum sharing in unlicensed bands is a promising approach for meeting the growing demands of forthcoming and deployed wireless networks, such as long-term evolution license-assisted access (LTE-LAA) and IEEE 802.11 Wi-Fi systems. In this paper, we consider a coexistence scenario where multiple LAA and Wi-Fi links compete for spectrum sharing subchannel access. We introduce a reinforcement-learning-based subchannel selection technique which allows access points (APs) and eNBs to select best subchannel distributively considering their medium access control (MAC) channel access protocols along with the physical layer parameters. The performance of this scheme is investigated through simulations, including the convergence property and sum throughput. Numerical results show that the proposed reinforcement-learning scheme converges fast and the sum throughput of the LAA and Wi-Fi systems is reasonably close to the result based on exhaustive search.**

***Index Terms*—5G and beyond, artificial intelligence, coexistence, LTE-LAA, MAC layer, PHY layer, Q-learning, WLAN.**

I. Introduction

Wireless communications are tightly integrated into our daily lives and promise to become more so in the future. Cisco Systems forecasts 3.6 mobile-connected devices per person by 2022 and a 7-fold increase in global mobile data traffic between 2017 and 2022 [1]. A promising approach to accommodating this growth is to transition from exclusively-licensed spectrum to a shared one, which offloads into unlicensed spectrum bands. One difficulty in this approach is balancing new network paradigms with incumbent networks, such as Wi-Fi.

Operating long term evolution (LTE) in unlicensed bands, such as with licensed assisted access (LAA), is a proposed solution to improve spectral-usage efficiency [2]. While such wireless coexistence proposals may have an enormous influence in solving spectrum usage issues, there are a number of challenges for both Wi-Fi and LTE networks to constructively share the spectrum. Assigning the spectrum in an intelligent and adaptive way may satisfy the diverse networks' requirements, build up an effective sharing of spectrum, and help to design future radio technologies (e.g., 5G NR-U [3]).

Innovations in artificial intelligence (AI) and machine learning (ML) are transformative in many industries. Wireless networking is also poised to benefit from these advances. For example, future 5G and beyond mobile devices are expected to access optimal spectral bands using highly-developed spectrum learning and inference. However, future networks will be immensely more intricate owing to complex network typologies and coordination schemes facilitating various end-user applications. Hence, it is computationally difficult for these future mobile devices to optimize key performance indicators (KPIs), such as throughput. Even more difficult to determine are strategies for efficient network management taking spectrum sharing into account. Nonetheless, as we demonstrate below, data-driven and adaptive machine learning algorithms are able to combat some of these difficulties to improve network performance.

Reinforcement learning (RL) is an area of ML and optimization which is well-suited to learning about dynamic and unknown environments [4]–[13]. It is especially suited to solving problems related to coexistence of wireless communications devices, as demonstrated by recent research. These works have been done mostly in the context of cognitive radio networks (CRNs). For example, [4] applied RL to estimate optimal channel access strategies for a single secondary user (SU) who dynamically selects one channel out of some number N channels. In [5], the authors assumed there is no primary user (PU) and applied RL to determine channel allocation for each SU without considering any collision with the PU. In [6], it was shown that the sensing capability, and therefore the transmission ability, could be enhanced for SUs with the proposed multi-agent RL (MARL)-based channel allocation. In [7], a MARL-based power control strategy was introduced to speed up the learning process of energy harvesting in communication systems.

A learning approach that accounts for the coexistence of LTE-U to model the resource allocation problem in LTE-U small stations (SBS), has been studied in [8], [9]. Specifically, to determine fair coexistence between LTE and Wi-Fi in the unlicensed spectrum, a Q-learning algorithm is applied in [8]. In [9] an RL algorithm based on long short-term memory (RL-LSTM) learning architecture was proposed to allocate the resources of LTE-U over the unlicensed spectrum. The most similar prior works ([10]–[13]) also used techniques from adaptive and reinforcement learning to study coexistence problems between Wi-Fi and other radio access technologies

(RATs). An adaptive MAC protocol for wireless sensor networks was proposed in [10], while the authors of [11] proposed a channel-selection method among LTE-U networks given only physical layer parameters. A learning method to select the duty cycle of LTE-U networks was presented in [12], [13].

The main contribution of our study is a novel reinforcement learning-based carrier-selection technique designed to ease the concurrent operation of Wi-Fi and LAA networks in unlicensed bands. Specifically, we use a Q-learning algorithm to select the spectrum in which a transmission node will operate based on self-learning experience. These transmission nodes each learn to maximize total network throughput without communicating with each other. Unlike earlier research on this topic [10]–[13], we take into account both MAC and physical layers and study the coexistence of two different types of network (LAA and Wi-Fi) operating simultaneously. To better reflect reality, we assumed there are different types of networks in which each base station serves several users. We consider the effects of collision between transmitting nodes. Moreover, in order to have a fully distributed strategy, there is neither a centralized controller in the network nor any information explicitly exchanged between different operators. The proposed algorithm enables both networks to constructively share the unlicensed spectrum and increase achieved data rates. It is worth noting that the introduced Q-learning algorithm could be applied to many types of communication systems, but LTE here is used as an example.

The remainder of this paper is organized as follows. Section II describes the system model and assumptions required for our analysis. Section III presents the problem formulation and introduces our proposed intelligent dynamic spectrum access. The impacts of the MAC and physical layer parameters in selecting a channel in a coexistence scenario is also explained in Section III. Simulation results are shown and discussed in Section IV. Finally, in Section V, an overview of the results and some concluding remarks are presented.

II. System Model

We consider a downlink coexistence scenario where two mobile network operators (MNOs) share for their operations the same unlicensed industrial, scientific, and medical (ISM) radio band. We are primarily interested in the operation of cellular base stations in an unlicensed band. However, the LTE base stations may have permission to utilize a licensed band as well. We assume each unlicensed band, indexed by $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$, can be shared between the MNOs in a time sharing fashion. The LAA network consists of n_L eNodeBs, while the Wi-Fi network is composed of n_W APs. The eNodeBs and APs are randomly distributed over a particular area, while LAA user equipment (UEs) and Wi-Fi clients/stations (STAs) are distributed around each eNodeB and AP, respectively, independently and uniformly. The transmission node $i \in \{\mathcal{L}, \mathcal{W}\}$, where $\mathcal{L} \triangleq \{n_\ell | \ell = 1, 2, \ldots, L\}$ and $\mathcal{W} \triangleq \{n_w | w = 1, 2, \ldots, W\}$, serves a set of $|\mathcal{U}_i|$ single antenna UEs/STAs on the unlicensed band, where $\mathcal{U}_i \triangleq \{u_{i,1}, u_{i,2}, \ldots, u_{i,|\mathcal{U}_i|}\}$. We assume that the transmission node i transmits with power p_i and the user association is based on the received power. Moreover, in order to meet the diverse network's requirement, we assume that eNodeBs may belong to different LTE operators with different priority classes (PCs) as introduced by the 3GPP standards committee [14]. Similarly, APs may belong to different Wi-Fi operators with four different access categories (ACs) [15]. We also assume *(i)* both Wi-Fi and LAA are in the saturated traffic condition, *(ii)* there is no hidden node problem[1] in the network (*i.e.,* every transmission node i is able to hear one another), and *(iii)*, the channel knowledge is ideal, so, the only source of packet failure (unsuccessful transmission) is collision.

Medium access in Wi-Fi is based on contention with random back-off. This process is known as carrier sense multiple access with collision avoidance (CSMA/CA) [16], [17]. In order to discover whether the channel is idle or busy, the station that accesses the medium should sense the channel by performing clear channel assessment (CCA). The distributed coordination function (DCF) operation progresses if the channel is found out to be idle. Otherwise, the transmitting station refrains from transmitting data until it senses the channel is available. Similarly, LTE-LAA uses a listen before talk (LBT) channel access mechanism to maintain fair coexistence with the Wi-Fi. Among different LAA-LBT schemes, Cat 4 LBT, which is based on the Wi-Fi CSMA/CA scheme, is well-suited to coexistence [2]. Although LTE and Wi-Fi technologies follow the same channel access procedure, they select different carrier sense mechanisms, different channel sensing threshold levels, and different channel contention parameters, leading to different unlicensed channel access probabilities and thus, different throughput.

In this section, we will briefly derive the normalized network throughput of both systems on each unlicensed band, a quantity we aim to optimize below. Conforming with the analytical model in [18], the probability of packet transmission by a transmitting node i in a randomly-chosen time slot on the k-th unlicensed channel can be written as

$$p_{\mathrm{tr},i}^{(k)} = \frac{2(1-2p_{c,i}^{(k)})}{(1-2p_{c,i}^{(k)})(1+\mathrm{CW}_{\min,i}^{(k)})+p_{c,i}^{(k)}\mathrm{CW}_{\min,i}^{(k)}(1-(2p_{c,i}^{(k)})^{m_i^{(k)}})}, \quad (1)$$

where $\mathrm{CW}_{\min,i}^{(k)}$ and $m_i^{(k)}$ are the minimum contention window size and the maximum back-off stage of the transmitting node i on the unlicensed band k, respectively, and $p_{c,i}^{(k)}$ is the probability of collision experienced by the i-th transmitting node on the k-th unlicensed channel. The probability of collision experienced by n_w AP and n_ℓ eNodeB can be expressed as

$$\begin{aligned} p_{c,n_w}^{(k)} &= 1 - \prod\nolimits_{\acute{w} \neq w} (1 - p_{\mathrm{tr},n_{\acute{w}}}^{(k)}) \prod\nolimits_{\ell} (1 - p_{\mathrm{tr},n_\ell}^{(k)}), \\ p_{c,n_\ell}^{(k)} &= 1 - \prod\nolimits_{w} (1 - p_{\mathrm{tr},n_w}^{(k)}) \prod\nolimits_{\acute{\ell} \neq \ell} (1 - p_{\mathrm{tr},n_\ell}^{(k)}), \end{aligned} \quad (2)$$

respectively, where $\acute{w} = 1, \ldots, W$ and $\acute{\ell} = 1, \ldots, L$.

[1]Here, we assume perfect spectrum sensing in both systems. Therefore, there are neither hidden nodes nor false alarm/miss detection problems in the network. The impact of imperfect sensing is beyond the scope of this paper and investigating the effect of CCA errors is an important topic for future work.

The probability of collision can be split into three parts: the probability of collision due to the collision between the Wi-Fi transmissions, between the LAA transmissions, and between the Wi-Fi and the LAA transmissions, respectively given by

$$p_{c,\mathcal{W}}^{(k)} = (1-p_{tr,\mathcal{L}}^{(k)})\big[p_{tr,\mathcal{W}}^{(k)} - \sum\nolimits_{w} p_{\text{tr},n_w}^{(k)} \prod\nolimits_{\acute{w}\neq w}(1-p_{\text{tr},n_{\acute{w}}}^{(k)})\big],$$
$$p_{c,\mathcal{L}}^{(k)} = (1-p_{tr,\mathcal{W}}^{(k)})\big[p_{tr,\mathcal{L}}^{(k)} - \sum\nolimits_{\ell} p_{\text{tr},n_\ell}^{(k)} \prod\nolimits_{\acute{\ell}\neq \ell}(1-p_{\text{tr},n_{\acute{\ell}}}^{(k)})\big],$$

and $p_{c,\mathcal{W},\mathcal{L}}^{(k)} = p_{tr,\mathcal{L}}^{(k)} \cdot p_{tr,\mathcal{W}}^{(k)}$. Here $p_{tr,\mathcal{L}}^{(k)} = 1 - \prod_{\ell=1}^{L}(1 - p_{\text{tr},n_\ell}^{(k)})$ and $p_{tr,\mathcal{W}}^{(k)} = 1 - \prod_{w=1}^{W}(1-p_{\text{tr},n_w}^{(k)})$ denote the LAA's and Wi-Fi's probability of transmission on the k-th unlicensed channel, respectively. Moreover, the probability of a successful transmission by the i-th transmitter on the k-th unlicensed band can be written as

$$p_{s,n_w}^{(k)} = p_{\text{tr},n_w}^{(k)} \prod\nolimits_{\acute{w}\neq w}(1-p_{\text{tr},\acute{w}}^{(k)}) \prod\nolimits_{\ell}(1-p_{\text{tr},n_\ell}^{(k)}),$$
$$p_{s,n_\ell}^{(k)} = p_{\text{tr},n_\ell}^{(k)} \prod\nolimits_{\acute{\ell}\neq \ell}(1-p_{\text{tr},\acute{\ell}}^{(k)}) \prod\nolimits_{w}(1-p_{\text{tr},n_w}^{(k)}). \quad (3)$$

Hence, the average length of a time slot in the unlicensed channel k can be calculated as

$$T_{\text{avg}}^{(k)} = (1-p_{tr}^{(k)})\mathbb{E}\{T_{\text{idle}}^{(k)}\} + p_{s,\mathcal{W}}^{(k)}\mathbb{E}\{T_{s,\mathcal{W}}^{(k)}\} + p_{s,\mathcal{L}}^{(k)}\mathbb{E}\{T_{s,\mathcal{L}}^{(k)}\}$$
$$+ p_{c,\mathcal{W}}^{(k)}\mathbb{E}\{T_{c,\mathcal{W}}^{(k)}\} + p_{c,\mathcal{L}}^{(k)}\mathbb{E}\{T_{c,\mathcal{L}}^{(k)}\} + p_{c,\mathcal{W},\mathcal{L}}^{(k)}\mathbb{E}\{T_{c,\mathcal{W},\mathcal{L}}^{(k)}\},$$

where $p_{tr}^{(k)}$ is the probability of occupation of the k-th unlicensed channel and can be expressed as

$$p_{tr}^{(k)} = 1 - \prod\nolimits_{w=1}^{W}(1-p_{\text{tr},n_w}^{(k)}) \prod\nolimits_{\ell=1}^{L}(1-p_{\text{tr},n_\ell}^{(k)}),$$

and $p_{s,\mathcal{W}}^{(k)} = \sum_w p_{s,n_w}^{(k)}$ and $p_{s,\mathcal{L}}^{(k)} = \sum_\ell p_{s,n_\ell}^{(k)}$ denote the successful transmission probability of the whole Wi-Fi and LAA networks on the k-th unlicensed band, respectively. Moreover, $T_{s,\mathcal{L}}^{(k)}$, $T_{s,\mathcal{W}}^{(k)}$, $T_{c,\mathcal{W}}^{(k)}$, $T_{c,\mathcal{L}}^{(k)}$, and $T_{c,\mathcal{W},\mathcal{L}}^{(k)}$ indicate the time that the k-th channel is occupied by an LAA successful transmission, a Wi-Fi successful transmission, a collision among the Wi-Fi transmissions, a collision among the LAA transmissions, and a collision between the Wi-Fi and the LAA transmissions, respectively. Considering the basic access scheme into account [18], [19], $T_{s,\mathcal{W}}^{(k)} = T_{P,\mathcal{W}} + T_{\text{SIFS}} + T_{\text{idle}}^{(k)} + T_{\text{ACK}} + T_{\text{DIFS}} + T_{\text{idle}}^{(k)} + (\text{PHY}_{\text{header}} + \text{MAC}_{\text{header}})/R_{\mathcal{W}}^{(k)}$, $T_{c,\mathcal{W}}^{(k)} = T_{P,\mathcal{W}} + (\text{PHY}_{\text{header}} + \text{MAC}_{\text{header}})/R_{\mathcal{W}}^{(k)} + T_{\text{DIFS}} + T_{\text{idle}}^{(k)}$, $T_{s,\mathcal{L}}^{(k)} = T_{P,\mathcal{L}} + T_{\text{SIFS}} + T_{\text{ACK}} + T_{\text{DIFS}} + T_{\text{idle}}^{(k)}$, $T_{c,\mathcal{L}}^{(k)} = T_{s,\mathcal{L}}^{(k)}$, and $T_{c,\mathcal{W},\mathcal{L}}^{(k)} = \max(T_{c,\mathcal{L}}^{(k)}, T_{c,\mathcal{W}}^{(k)})$; where $T_{P,\mathcal{W}}$ ($T_{P,\mathcal{L}}$) indicates the Wi-Fi (LAA) payload duration, $\text{PHY}_{\text{header}}$+$\text{MAC}_{\text{header}}$ presents the packet header, and T_{SIFS}, T_{ACK}, and T_{DIFS} refer to short interframe space (SIFS), acknowledgement signal duration, and distributed interframe space (DIFS), respectively.

Finally, the normalized network throughput of LAA and Wi-Fi systems on the k-th unlicensed band as a function of both MAC and Physical layers parameters and can be expressed as

$$S_{\mathcal{L}}^{(k)} = p_{s,\mathcal{L}}^{(k)} T_{P,\mathcal{L}} R_{\mathcal{L}}^{(k)} / T_{\text{avg},\mathcal{L}}^{(k)}, \quad (4)$$
$$S_{\mathcal{W}}^{(k)} = p_{s,\mathcal{W}}^{(k)} T_{P,\mathcal{W}} R_{\mathcal{W}}^{(k)} / T_{\text{avg},\mathcal{W}}^{(k)},$$

where $T_{\text{avg},\mathcal{W}}^{(k)}$ ($T_{\text{avg},\mathcal{L}}^{(k)}$) denotes the average time duration to achieve a successful transmission in the Wi-Fi (LAA) network and $R_{\mathcal{W}}^{(k)}$ ($R_{\mathcal{L}}^{(k)}$) refers to the Wi-Fi's (LAA's) physical data rate, which can be calculated as follows. Assuming each channel k can be shared between the UEs associated with the eNodeB n_ℓ in a time sharing fashion, the physical data rate of the LAA network on the k-th unlicensed band can be expressed as [20]

$$R_{\mathcal{L}}^{(k)} = \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{i=1}^{|\mathcal{U}_{n_\ell}|} c_{u_{n_\ell,i}}^{(k)} B_k \log_2(1 + \chi_{u_{n_\ell,i}}^{(k)} \cdot \text{SINR}_{u_{n_\ell,i}}^{(k)}),$$

where $0 \leq c_{u_{n_\ell,i}}^{(k)} \leq 1$ denotes the time sharing component of the UE $u_{n_\ell,i}$ on the k-th unlicensed band that fulfills $0 \leq \sum_{n_\ell \in \mathcal{N}_L} \sum_{u_{n_\ell,i} \in \mathcal{U}_{n_\ell}} c_{u_{n_\ell,i}}^{(k)} \leq 1$, B_k is the bandwidth of the k-th channel, and $\chi_{u_{n_\ell,i}}^{(k)}$ is the channel access indicator which is 1 if the eNodeB n_ℓ serves the UE $u_{n_\ell,i}$ on the k-th unlicensed channel, and is 0 otherwise. Moreover, $\text{SINR}_{u_{n_\ell,i}}^{(k)}$ represents the signal-to-interference-plus-noise ratio (SINR) for serving user $u_{n_\ell,i}$ on the k-th unlicensed band. Following the same notation, the physical data rate of the Wi-Fi's network can be written as [20]

$$R_{\mathcal{W}}^{(k)} = \sum\nolimits_{w=1}^{W} \sum\nolimits_{j=1}^{|\mathcal{U}_{n_w}|} B_k \log_2(1 + \chi_{u_{n_w,j}}^{(k)} \cdot \text{SINR}_{u_{n_w,j}}^{(k)}).$$

It is worth noting that in the case where an eNodeB/AP aggregates multiple unlicensed bands, the total throughput may be obtained by summation over all unlicensed channels.

III. Intelligent Dynamic Spectrum Access

If both Wi-Fi and LAA on the unlicensed band select channels appropriately, the spectral efficiency of the whole network will improve. Equation (4) reveals the impact of channel selection by each MNO on network's throughput. Specifically, if transmitter i selects channel k and channel k is not occupied by the other transmission nodes, then the probability of successful transmission by the i-th node in the k-th channel increases, leading to higher throughput. Moreover, if each channel is selected such that interference is avoided (or at least minimized) among the transmission nodes, there will be a higher SINR, leading to a higher physical data rate and throughput. In this section, we employ an algorithm for channel selection that learns from previous experience to improve network throughput. The learning algorithm also makes adaptive use of channel usage information.

We introduce a Q-Learning based dynamic channel selection algorithm here for both MNOs. The aim is to enhance the throughput of both Wi-Fi and LAA networks on the unlicensed band. Q-Learning is a value-based reinforcement learning technique that discovers the best action at each state by maximizing a value function, typically denoted by Q, that returns the expected reward of each action-state pair [21]. Here, "actions" are ways to explore the environment, and "states" describe the environment (we give an example for the coexistence problem below). The goal is to learn how to map environmental conditions into actions that maximize a reward. By interacting with the environment and trying different actions over time, the algorithm learns which action

(in each state) provides the highest reward. Employing the Q-learning algorithm in a coexistence scenario allows each transmission node to learn to select the channel that yields the best throughput based on its interaction with the environment.

To be specific, we consider a set of $\mathcal{L}$ eNodeBs and $\mathcal{W}$ APs as the transmitters which will take actions. Denote the ith transmitter's set of actions by $\mathcal{A}_i = \{a_{i,1}, \ldots, a_{i,|\mathcal{A}_i|}\}$ and states $\mathcal{S}_i = \{s_{i,1}, \ldots, a_{i,|\mathcal{S}_i|}\}$. There is an associated set of Q-values related to each state-action pair. This is often called a Q-table, and it is unknown a priori. In the proposed algorithm, we consider four states for each channel observed by each transmission node: Idle, Successful transmission, Collision, and Contention. In the intelligent dynamic spectrum access of our interests, a collision is detected if both transmitters select the same channel for their operations, both have at least one user to serve, and the energy that they received from each other is less than predefined energy detectors threshold. The observed state is Contention if the initial clear channel assessment returns busy, either immediately or during the DIFS, or if the transmitter wants to contend after a successful transmission. The actions correspond to picking a different transmission band out of $|\mathcal{K}|$ bands in $\mathcal{K}$ and $|\mathcal{K}| = |\mathcal{A}_i|$.

The Q-table determines the best action at each state as judged by the action-state pair which yields highest reward. Particularly, the Q-value at each action-state pair is initialized with an estimate of the future reward. That results from a transmitter selecting a particular action while it is in a certain state. Then, based on interaction and feedback with the environment (e.g., MAC layer parameters, channel propagation model, topology of the Wi-Fi and LAA networks, sensing threshold), the transmitter learns the outcome of that action-state pair. In the proposed algorithm, each transmission node i measures $Q(s_{i,j}, a_{i,k})$, the expected reward achieved by selecting the k-th channel when the i-th transmission node is in the j-th state. If the k-th channel has never been selected by the transmission node i in the past, then $Q(s_{i,j}, a_{i,k})$ value is fixed to an arbitrary value Q_0 at the initialization step.

We denote by $R(s_{i,j}, a_{i,k}) := S_{\mathcal{L}}^{(k)} + S_{\mathcal{W}}^{(k)}$ the reward resulting from environmental interaction. Since our goal is to select the channel that maximizes the spectral efficiency of the whole network, we define this reward as the average throughput acquired by the i-th transmission node when it is in the j-th state and selects the k-th action (channel). The Q-value is then iteratively updated by

$$Q_{t+1}(s_{i,j}, a_{i,k}) \leftarrow (1-\alpha) Q_t(s_{i,j}, a_{i,k}) + \alpha R_{t+1}(s_{i,j}, a_{i,k})$$

where $0 \leq \alpha \leq 1$ is the learning rate and determines how fast the learning happens [21]. As is typically the case in learning algorithms, selecting α requires some care. Indeed, a small value of α results in a long learning process (which could be detrimental in practice), while a large value of α could cause the algorithm to not converge. As soon as the Q-value of the current state-action is updated according to the above equation, the i-th transmission node selects a channel corresponding to its new state.

Note that in order to find the effect of each selected channel, the transmission node needs to explore the environment. This suggests that the transmission nodes not only select the channels that provide the current highest estimated throughput, but also take on new actions to find better channel selection strategies. There are many methods for choosing action-selection strategies to balance this exploration of the environment with reward maximization. Examples are greedy action-selection, ϵ-greedy action-selection, and softmax action-selection [21]. In the greedy action-selection strategy, the transmission node always picks the channel that provides the highest Q-value and never explores any new channels. Alternatively, in the ϵ-greedy action-selection method, the transmission node uniformly picks a channel with probability ϵ. Although this strategy attempts to balance the exploration of the environment with maximizing reward, the chance of selecting the worst-appearing channel is the same as the chance of picking any other channel, due to the uniform selection probability. The softmax action-selection policy, like the above aforementioned action-selection strategies, gives the highest selection probability to the greedy action. In contrast, however, the rest of the actions are weighted according to their estimated values. Under this policy, the probability of picking the k-th unlicensed channel by the i-th transmission node is

$$Pr(i,k) = \exp\left(\frac{Q(s_{i,j}, a_{i,k})}{\tau(i)}\right) / \sum_{m=1}^{K} \exp\left(\frac{Q(s_{i,j}, a_{i,m})}{\tau(i)}\right),$$

where $\tau(i) > 0$ is the so-called temperature. With a high value of $\tau(i)$, the softmax policy is nearly identical to the ϵ-greedy action-selection strategy and so, the probability of selecting the different channels is almost the same. In contrast, selecting $\tau(i)$ with a low value makes a big difference in selection probability of an action that varies in Q-value. As a result, the probability of selecting a channel increases when it has a larger Q-value. Then, as $\tau(i)$ goes to zero, the softmax policy becomes the greedy action-selection strategy. Hence, we select the softmax policy to best balance exploration and reward maximization. Moreover, as a means to decrease the amount of exploration by the transmission node i (when the number of channel selection by this node increases), we consider a cooling function [22] that can be expressed as

$$\tau(i) = \tau_0 / \log_2(1 + n(i)),$$

where τ_0 denotes the initial temperature and $n(i)$ represents the number of channels that have been selected by the i-th transmission nodes.

Remark 1. *Compared to existing approaches, Q-learning is at an advantage. Indeed, the Q-learning approach is model-free and becomes adjusted to the new environment changes quickly. Therefore, there is no need to have the full knowledge of state transition in a dynamic environment. In addition, and in terms of computational complexity, Q-learning has much less complexity compared to the tradition optimization approaches such as branch and bound algorithm. To be specific, Q-learning is computationally limited by the size of the action and state spaces. Q-learning is quadratic in the number of st-*

ates and linear in the number of actions, e.g., $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$.

IV. Simulation Results and Analysis

We evaluate the performance of the proposed algorithm in a coexistence scenario. We simulate a scenario in which 3 eNodeBs compete for the unlicensed channels with 3 APs. All transmitters are randomly distributed over an area of size 120×80 m^2 with minimum distance 40 meters, as shown in Fig. 1. All UEs and Wi-Fi clients are independently and uniformly distributed around each eNodeB and AP, respectively. We consider 5 UEs (Wi-Fi clients) per eNodeB (AP). Each UE (Wi-Fi client) is assigned to the eNodeB (AP) that provides it with the highest received power. The antenna height of the transmission nodes and users are 6 meters and 1.5 meters, respectively. The carrier frequency is 5 GHz and each channel bandwidth is 20 MHz. The path-loss and shadowing between transmission nodes, and between transmission nodes and users are generated following [23] for the indoor scenario. The transmit power at each transmission node is fixed to 23 dBm while the noise figure and the thermal noise level at each user is set to 9 dB and -174 dBm/Hz, respectively. Moreover, we assume the omni-directional antenna pattern with a 0 dBi antenna gain. While Wi-Fi uses the CSMA/CA scheme to access the unlicensed band with CCA-CS and CCA-ED thresholds of -82 dBm and -62 dBm, respectively [15], an LBT with random back-off and variable contention window size with CCA-ED threshold of -72 dBm is considered for the LAA channel access scheme [14]. The MAC layer parameters are given in Table I, and the initial Q-learning parameters are set to $\alpha = 0.1$, $Q_0 = 0.5$, and $\tau_0 = 0.15$.

According to this geometry and propagation model, only some pairs of transmitters can detect each other. We summarize which transmitters can detect each other in matrix $\mathbf{I}_D$ given below, in which the ij-th element of $\mathbf{I}_D$ is nonzero only if transmission node i detects transmission j. In the case of detection, this element is 1. The first three rows (columns) corresponds to AP1 to AP3 while eNodeB1 to eNodeB3 are organized in the last three rows (columns). For the aforementioned scenario the matrix $\mathbf{I}_D$ is given as

$$\mathbf{I}_D = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

Note that in calculating the reward function and updating the Q-table in the proposed algorithm, we need to determine the throughput. For computing the throughput, we need to determine SINR. The $\mathbf{I}_D$ matrix, which tells us which pairs of transmitters can detect each other, helps us to calculate the interference and hence the SINR, and throughput.

We consider the case of 6 different available channels in which all active transmission nodes in the area apply the proposed Q-learning method. Neither Wi-Fi nor LAA has knowledge about the other operators' selected channels. Each node *(i)* measures the expected reward achieved by selecting each channel when it is in a specific state, *(ii)* updates its Q-table, and *(iii)* chooses the channel that maximizes the whole network throughput.

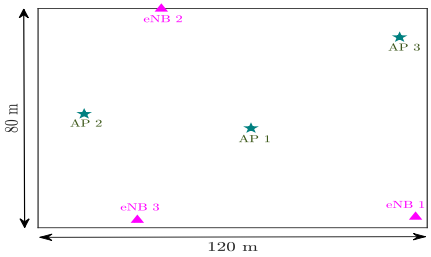

Fig. 1: Simulation layout

TABLE I: MAC Layer Parameters

Parameter	**value**
LAA's packet payload duration	1 ms
Wi-Fi's packet payload duration	1 ms
MAC header	272 bits
PHY header	128 bits
ACK	112 bits + PHY header
SIFS	16 μs
DIFS	34 μs
Idle slot time	9 μs
Wi-Fi contention window size	16
LAA contention window size	16
Wi-Fi maximum backoff stage	6
LAA maximum backoff stage	3

The channel selection probability of each transmission node is shown in Fig. 2. At the initialization step, each transmission node interacts with the environment to explore different channels. It is worth mentioning that if the radio conditions change, the transmitters keep interacting with the environment until they determine which channel best improves the whole network throughput. With a change in radio conditions we will see that a learner must re-learn its policies as well. It is expected that any changes in the radio conditions during the initialization step will increase the initialization step duration. When the initialization step is done, according to the feedback from the environment and past experience, each node selects a channel that maximizes the whole network throughput, with a high probability. Since AP1 detects all other transmission nodes, it selects a fully vacant channel. AP2 and AP3 cannot detect each other, and they select the same channel. eNodeB1 only detects AP1. At first eNodeB1 selects the 4-th channel for its transmission, but when AP1 selects this channel, the eNodeB1 learns that this channel is now occupied, and switches to channel 2. The eNodeB2 only detects AP1 and AP2, so learns that the channels occupied by these two nodes do not maximize its throughput, and therefore selects a different channel. However, this channel is using by eNodeB1 that eNodeB2 does not detect. Finally, eNodeB3 selects the first channel which none of the other nodes picks as the means of maximizing the whole network spectral efficiency.

Simulation time is measured with respect to algorithmic time steps. Each time step is equal to the transmission nodes payload duration, i.e., 1 ms. Taking the MAC layer parameters

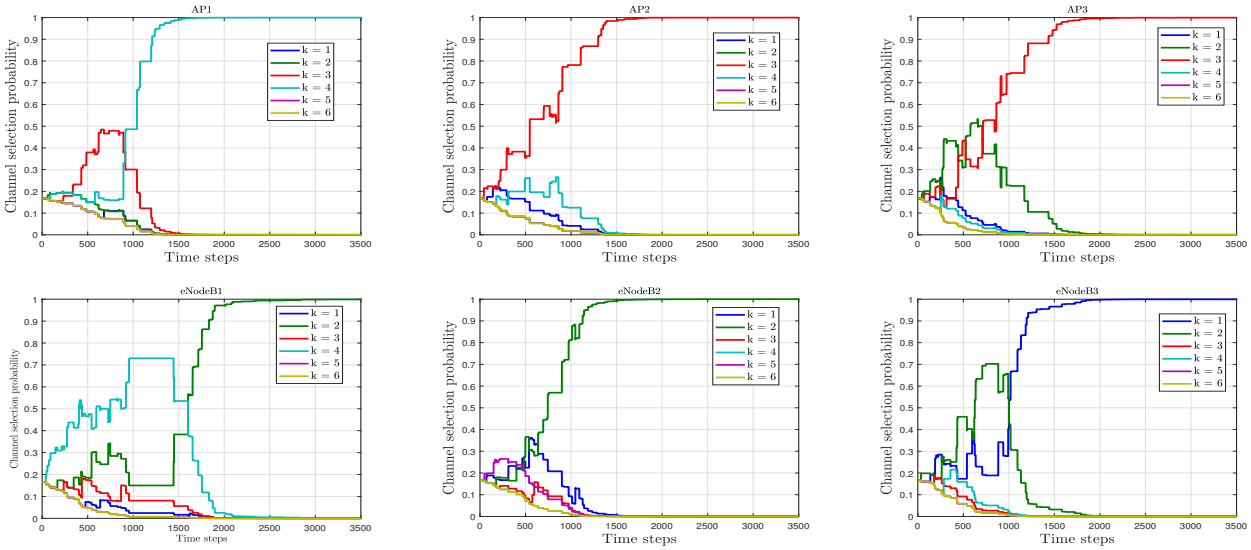

Fig. 2: Channel selection probability of each transmission node versus the time steps.

into account, all the APs and eNodeBs generate activity periods where they transmit data to their users with an average of 69- and 77-time steps. Since, the duration of initial time to learn the best solution is in order of 800-time steps in our simulations, the decisions of selecting a channel for transmitting the data are made after trying 10 initial channel selections.

It is worth mentioning that our simulation did not converge to a simple solution of allocating each transmitter to its own channel, instead favoring a scenario in which the transmitters choose the same channels. The reason for this solution is related to the physical scenario in which we pose the problem, as well as the MAC and physical layer parameters used by the APs and eNodeBs.

In Fig. 3 we consider a challenging case in which there are only 4 channels available. All active transmission nodes apply the proposed Q-learning method. For this scenario, again, we randomly distribute all transmitters and receivers over the area as discussed earlier. According to this geometry and propagation model the new matrix $\mathbf{I}_D$ is given as follows

$$\mathbf{I}_D = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{6}$$

Since there are only 4 available channels, sharing the same spectrum bands in the time domain is inescapable. Fig. 3 represents the channel selection probability of each transmission node. As it is shown in Fig. 3 after an initialization step that all channels are evaluated, AP 2 and eNodeB 2 use the same channel. AP 2 and eNodeB 2 are located far from each other and they are not reciprocally detected during the sensing period. Hence, they can utilize the same channel without sharing it in the time domain. Note that if these two transmitters detected each other and decided to select a same channel, they would share the channel in the time domain obeying the LBT requirements. The same discussion is applied to AP3 and eNodeB 1. AP1 and eNodeB 3 learn to use channel 1 and 4, respectively. This is a good choice since both not only detect each other but also detect AP2 and eNodeB 1 and choose different channels (from those selected by these detected nodes) for their transmissions.

In order to evaluate the usefulness of the proposed algorithm, the performance is measured in terms of the cumulative distribution function (CDF) of the normalized throughput. In Fig. 4 the CDF performance of the normalized throughput of the proposed scheme is compared with that of the optimum channel allocation algorithm. The optimum channel allocation assigns the channels to the transmission nodes in such way that the total network throughput maximizes and, it is found by exhaustive search over all possible channel combinations. Fig. 4 shows that the performance of the proposed Q-learning algorithm is close to the optimum one. Specifically, it is shown that under the proposed algorithm users have high probability to achieve the near-optimum normalized throughput with much less computational complexity.

V. Conclusion

In this paper, we have studied machine-learning-based channel selection in a coexistence scenario of LAA and Wi-Fi networks. We have proposed a reinforcement learning-based spectrum selection algorithm that allows the concurrent operating of Wi-Fi and LTE networks in an intelligent and efficient way. This scheme has taken into account the effects of both MAC and physical layers. We evaluated each channel's probability of selection, convergence property, and sum throughput. Simulation results demonstrate that this scheme converges fast and provides a competitive sum throughput reasonably close to that based on exhaustive search.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022 White Paper," *Cisco Systems*, Feb. 2019.

[2] "Study on licensed-assisted access to unlicensed spectrum (Release 13)," *3GPP TR. 36.889 v13.0.0*, June 2015.

[3] S. Verma and S. Adhikari, "3GPP RAN1 status on LAA and NR Unlicensed," *IEEE 802.11-18/0542r0*, Mar. 2018.

[4] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, Deep reinforcement learning for dynamic multichannel access in wireless networks, IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 2, pp. 257265, Jun. 2018.

[5] O. Naparstek and K. Cohen, Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks, in Proc. GLOBECOM, pp. 17, Dec. 2017.

[6] M. A. Aref, S. K. Jayaweera, and S. Machuzak, Multi-agent reinforcement learning based cognitive anti-jamming, in Proc.IEEE Wireless Commun. Netw. Conf. (WCNC), pp. 16, Mar. 2017.

[7] X. He, H. Dai, and P. Ning, Faster learning and adaptation in security games by exploiting information asymmetry, IEEE Trans. Signal Process., vol. 64, no. 13, pp. 34293443, Jul. 2016

[8] V. Maglogiannis, D. Naudts, A. Shahid, and I. Moerman, A Q-learning scheme for fair coexistence between LTE and Wi-Fi in unlicensed spectrum, IEEE Access, vol. 6, pp. 2727827293, 2018.

[9] U. Challita, L. Dong, and W. Saad, Deep learning for proactive resource allocation in LTE-U networks, in Proc. 23rd Eur. Wireless Conf. Eur. Wireless, pp. 16, May 2017.

[10] Z. Liu, and I. Elhanany, "RL-MAC: a reinforcement learning based MAC protocol for wireless sensor networks," *Int. J. of Sens. Netw. (IJSNET)*, Vol. 1, No. 3/4, 2007.

[11] O. Sallent, J. Prez-Romero, R. Ferrs, and R. Agust, "Learning-based Coexistence for LTE Operation in Unlicensed Bands," *IEEE Int. Conf. on Commun. Wkshp*, June 2015.

[12] N. Rupasinghe, and I. Guvenc, "Reinforcement Learning for Licensed-Assisted Access of LTE in the Unlicensed Spectrum", *IEEE Wireless Commun. and Netw. Conf.*, Track 3: Mobile and Wireless Netw., Mar. 2015.

[13] U. Challita, L. Dong, and W. Saad,"Deep Learning for Proactive Resource Allocation in LTE-U Networks", *23th European Wireless Conf.*, Aug. 2017.

[14] "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, (Release 15)," *3GPP TS 36.213 V15.6.0*, June 2019.

[15] IEEE Standard for Information technology, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pp. 1-23534, 2016.

[16] A. Babaei, J. Andreoli-Fang, Y. Pang, and B. Hamzeh, "On the impact of LTE-U on Wi-Fi performance," *Int. J. of Wireless Inf. Netw.*, vol. 22, issue 4, pp. 336-344, Dec. 2015.

[17] E. Perahia and R. Stacey, "Next Generation Wireless LANs: 802.11n and 802.11ac," Cambridge, U.K.: Cambridge Univ. Press, June 2013.

[18] 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.

[19] Y. Ma, D. G. Kuester, J. Coder, and W. F. Young, "Slot-Jamming Effect and Mitigation Between LTE-LAA and WLAN Systems With Heterogeneous Slot Durations", *IEEE Trans. on Commun.*, vol. 67, no. 6, pp. 4407-4422, June 2019.

[20] S. Mosleh, Y. Ma, J. B. Coder, E. Perrins, and L. Liu, "Enhancing LAA Co-existence Using MIMO Under Imperfect Sensing", accepted for publication in the *IEEE GC Wkshps*, 2019.

[21] R.S. Sutton, A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.

[22] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-6, No. 6, pp. 721-74, Nov. 1984.

[23] "Study on NR-based access to Unlicensed Spectrum; (Release 16)," *3GPP TR 38.889 V16.0.0*, Dec. 2018.

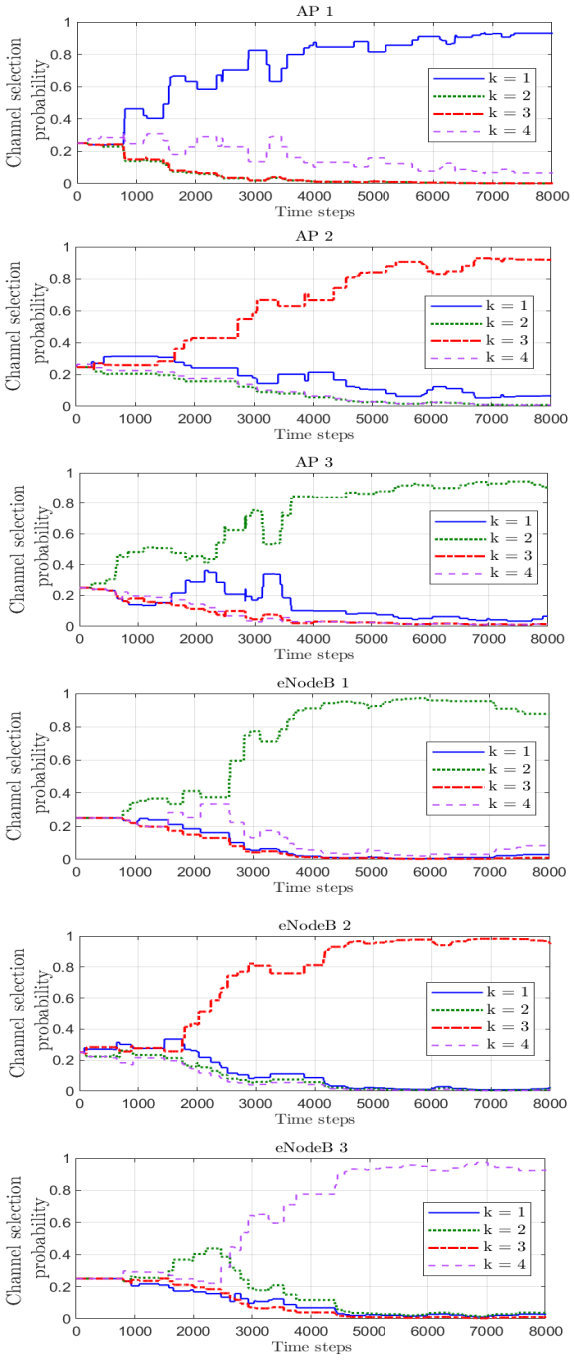

Fig. 3: Channel selection probability of each transmission node versus the time steps.

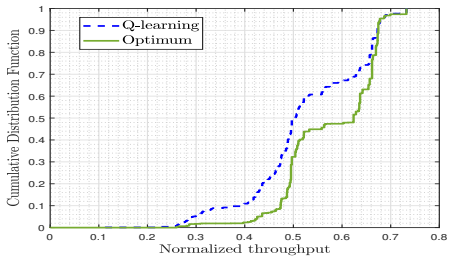

Fig. 4: CDF of achieved normalized throughput.