Abstract—Dynamic spectrum access (DSA) has been regarded as a viable solution to the spectrum shortage problem. To find idle spectrum, partial spectrum sensing could be employed by selecting a suitable sensing window (SW). Since the SW selection determines how many available bands to access, the transmission performance after the access could be used to guide the SW selection. Hence, a sophisticated joint design on spectrum sensing and access is necessary, which, however, is a challenging task when considering the dynamic nature of spectrum environment, and also the mutual impact among different secondary users (SUs). In this paper, we propose a joint partial spectrum sensing and power allocation (PA) scheme to facilitate SUs to make the best decisions on SW and PA to maximize the network throughput with reduced mutual interference. Considering the environmental dynamics and spectrum uncertainty, we develop a viable solution based on hierarchical multi-agent deep reinforcement learning (HMADRL). Our solution enables mutual design with two stages: making each SU learn the best SW and PA strategies autonomously while adapting to the dynamic environment. By using both simulated spectrum data and real spectrum data measured by SAM60-BX, we have demonstrated the effectiveness of our proposed scheme.

Index Terms—Dynamic spectrum access (DSA), partial spectrum sensing, power allocation, hierarchical deep reinforcement learning, multi-agent.

I. INTRODUCTION

The Internet of Things (IoT) era will see billions of IoT devices connected to telecommunication networks, resulting in a dramatic increase in wireless data traffic [2], which makes spectrum an extremely valuable resource [3]. However, under the current static spectrum allocation policy, in which licensed spectrum bands are exclusively owned by certain incumbents [4], the spectrum utilization is very low as demonstrated in many spectrum measurement campaigns [5], [6]. To deal with this paradox, dynamic spectrum access (DSA) is regarded as one of the key candidate technologies to resolve this problem and has attracted intensive research interests [7]–[13]. In DSA, incumbent users, i.e., primary users (PUs), are the owners of the licensed spectrum bands and allow secondary users (SUs) to opportunistically access their bands if they do not use the bands. It can effectively mine spectrum holes and significantly improve the utilization of the limited and precious spectrum resources, which meets the stringent requirements in 6G for high traffic and massive access. To enable DSA, two key issues need to be addressed, namely, spectrum sensing and spectrum access.

• Spectrum Sensing. In general, spectrum sensing is a time-energy-consuming process. Hence, partial spectrum sensing has been considered as an effective way to find idle bands due to the hardware limitation [14]. For each SU, it needs to determine a suitable sensing window (SW) to find as many idle bands as possible. Unfortunately, PUs’ activities are usually uncertain and dynamic, and how to quickly and accurately select a suitable SW is not an easy task [15]. Furthermore, considering the multi-user scenario, it is necessary to prevent SUs from selecting overlapping SWs to avoid strong mutual interference among them. Hence, the performance of spectrum access needs to be considered when making the SW decision.

• Spectrum Access. After acquiring the sensing result under certain SW, each SU needs to make the spectrum access decision, i.e., perform power allocation (PA) on the idle bands. To maximize the throughput of SUs, interference management is the key [16], which calls for an effective PA decision to reduce the overall mutual interference effectively.

As a result, the sensing strategy could directly impact the final access performance, which can also be used to evaluate how good a sensing strategy is. Hence, a joint design on both spectrum sensing and access is crucially important. Thus, we should address two main problems: 1) SW Determination: as aforementioned, the selection of SW could directly affect the final access performance, which determines the available bands for PA. However, selecting overlapping SWs may cause serious mutual interference. Thus, SW determination should consider both idle band discovery and final access performance. 2) PA Determination: After selecting a SW, each SU needs to perform PA on the idle bands captured within SW. To maximize the
throughput, it is necessary to make an appropriate PA decision to manage the interference. Thus, the selected SW determines the final performance of PA, which could in turn be employed to evaluate and guide the SW determination. In other words, a sophisticated joint design is needed to make SUs accurately capture the idle bands with non-overlapping SWs and make an effective PA decision with mutual interference management to maximize the throughput.

Unfortunately, most existing works on DSA studied spectrum sensing and spectrum access separately, where the close correlation between them is not considered in-depth. Based on the traditional optimization approaches, many works focused on either sensing strategy to improve the accuracy of finding idle bands [17]–[19] or access decision to optimize the performance of DSA networks [20]–[22]. Although such optimization-based centralized control methods are easy to implement, they may not be well suited for dynamic and uncertain spectrum environments [23]. It is much better to facilitate each SU to make decisions efficiently in a decentralized manner, where mutual interference among different SUs should be well managed.

With the advent of the 6G and beyond and the increasing demands from IoT devices, AI-based method, i.e., the intelligent spectrum sensing and access, is expected to become an effective enabler to support more dynamic and complex spectrum management in 6G systems. Some works have utilized deep reinforcement learning (DRL) to make DSA decisions under dynamic and uncertain environment recently. They mainly focused on either partial spectrum sensing design to effectively identify spectrum holes [24]–[27], or band selection and power control to improve the access throughput with low interference level [28]–[33]. However, we notice that spectrum sensing and spectrum access are two closely related issues for DSA. Since spectrum sensing strategy determines how many available bands to access, transmission performance after performing access strategy could be used to guide sensing strategy. For example, sensing and accessing the same bands will cause strong interference among SUs, resulting in a significant decrease in throughput. Such mutual impact and close correlation between two issues should be carefully taken into consideration in order to achieve a sophisticated joint optimal design, which, however, are not comprehensively considered in most existing works. A few works do have jointly considered two issues [34]–[36], which took spectrum sensing and access as two sequential stages and designed sensing and access strategies simultaneously based on DRL. However, they mainly focused on single-user scenarios without taking mutual impact among different SUs into account.

In this paper, by considering the close correlation between spectrum sensing and access, we propose a Joint Partial spectrum Sensing and Power Allocation (JPSPA) scheme to help multiple SUs make the best SW and PA decisions to achieve high throughput. To deal with the environmental dynamics and uncertainty with mutual impact among SUs, we develop a viable solution based on hierarchical multi-agent deep reinforcement learning (HMADRL). Different from most existing works solely relying on multi-agent deep reinforcement learning (MADRL) to determine all the decisions simultaneously, the proposed HMADRL approach can achieve effective mutual adaptation and guidance between sensing and access stages, and improve two strategies by well considering the mutual impact between them. Each SU can learn the statistical characteristics of the spectrum environment, as well as other SUs’ decisions, and thus make the optimal SW and PA decisions autonomously based on its local partial observation. In our preliminary work [1], we have proposed a MADRL approach to facilitate DSA in a dynamic and uncertain spectrum environment, which, however, did not consider the close correlation between spectrum sensing and access in depth, and SW and PA are selected simultaneously without such a hierarchical framework design.

The main contributions of this paper can be summarized as follows:

- Unlike most existing works that separately investigate spectrum sensing and spectrum access, we take the correlation between them into account and jointly address them to improve the transmission performance of DSA. By considering the mutual impact between two stages, we propose a JPSPA scheme, which helps each SU make the optimal SW and PA decisions autonomously, so that SUs can identify as many idle bands as possible with non-overlapping SWs while maximizing their throughput.
- By considering the uncertainty and dynamics of the spectrum environment, we develop a hierarchical multi-agent deep reinforcement learning (HMADRL) algorithm for the JPSPA scheme. The reward functions for two learning processes are jointly designed, where the final access performance is considered during the sensing stage to evaluate and guide the SW selection. Such a hierarchical learning framework could effectively cascade two learning processes, and thus improve SW and PA strategies by taking their mutual impact into account.
- The proposed HMADRL approach can be implemented in a centralized training and decentralized execution manner, where each SU can make adaptive decisions based on its local observation with reduced mutual interference. Moreover, considering that the sensing result determines which bands are available to access, to make the PA action feasible, we develop an action adjustment (AA) mechanism for the HMADRL algorithm to ensure that SUs only perform PA on the idle bands within SW. Real spectrum data measured by SAM-60BX has been adopted to show the effectiveness of the proposed scheme.

The rest of this paper is organized as follows. Related works are reviewed in Section II. The system model is presented in Section III. In Section IV, we introduce the joint partial spectrum sensing and power allocation scheme. Then, the proposed solution based on hierarchical multi-agent deep reinforcement learning is presented in Section V. In Section VI, the simulation results and discussions are presented, followed by the conclusions in Section VII.

II. RELATED WORKS

Due to the great potential of DSA to improve the temporal and spatial spectrum utilization, many works have been devoted to DSA networks [17]–[22], [24]–[36]. Based on the traditional
optimization approaches, many papers focused on either sensing strategy to improve the accuracy of identifying idle bands [17]–[19] or access decision to optimize the performance of DSA networks [20]–[22]. Towards the spectrum sensing design, in [17], aiming at optimizing the detection performance in an efficient way, Zhang et al. proposed a cooperative spectrum sensing (CSS) algorithm while satisfying a given error bound. To make the spectrum sensing more accurate and reliable, in [18], Edward et al. proposed a CSS scheme to make an efficient decision fusion rule at the control center. Similarly, in [19], Liu et al. developed an optimal CSS strategy to determine the final decision threshold to maximize the throughput of secondary networks. Regarding the spectrum access design, in [20], Zhou et al. proposed a probabilistic approach to optimize the channel and power allocation, which can maximize the overall utility and support quality-of-service requirements for SUs. To improve the energy efficiency in DSA networks, in [21], Ren et al. proposed a joint power allocation and channel access scheme. In [22], Lin et al. developed a secure and distributed power control algorithm, which can achieve optimal power control to maximize the sum rates of SUs. Although such centralized control methods are easy to implement, they may not be well suited for the dynamic and uncertain spectrum environments.

Recently, some papers developed the DRL approaches to deal with the dynamic and uncertain environments of DSA networks. They mainly focused on either partial spectrum sensing design to effectively capture the spectrum holes [24]–[27], or band selection and power control to improve the access throughput with low interference level [28]–[33]. As for the partial spectrum sensing design, some works treated it as a partially observable Markov decision process (POMDP) problem. To reduce the probability of collisions with PUs, [24] and [25] developed a deep Q-network based approach and an actor-critic DRL framework, respectively, to enable SU to maximize the number of successful transmissions. In [26], Gao et al. designed a CSS algorithm to efficiently find spectrum holes with reduced overheads. In [27], Chang et al. proposed a novel framework based on a deep echo state Q-network to facilitate SUs to effectively learn the sensing criterion with limited training data. As for the band selection and power control design, in [28], Xu et al. developed a DRL based approach to learn a channel access strategy with a low collision rate but a high channel utilization rate under predetermined partial sensing modes. In [29], Chang et al. designed a DRL based solution to help each SU make spectrum access decision in a distributed manner. Considering the privacy protection and reduction of overhead, they further developed a novel framework called Fed-MADRL to combine federated learning (FL) and MADRL in [30] to enable SUs to learn a collaborative spectrum access policy. To manage the interference among PUs and SUs, in [31], Song et al. developed a deep echo state Q-network based framework to enable SUs to perform efficient power control in a distributed manner. Similarly, in [32], Zhang et al. propose an asynchronous advantage actor critic (A3C)-based power control approach to help SUs adjust its power effectively to meet the quality of service requirements for PUs and SUs. Aiming to maximize the weighted sum rate of a device-to-device (D2D) network, Tan et al. proposed a distributed DRL-based scheme in [33] to enable D2D pairs to autonomously optimize channel selection and transmit power.

Although these works can enable SUs to well address either sensing or access problems under dynamic and uncertain spectrum environment based on DRL approaches, most of them studied two issues separately. Since the spectrum sensing strategy determines how many available bands to access, the transmission performance after access could be used to guide the sensing decision. Spectrum sensing and spectrum access are two closely related issues, which should be jointly considered. A few recent works have jointly taken both issues into account [34]–[36]. In [34], focusing on spectrum sensing, channel probing and power control problem, Wu et al. proposed a DRL based solution to maximize the throughput of SU. Similarly, in [35], Wang et al. developed a POMDP aided network association scheme for the joint communication and radar system to maximize the network throughput while minimizing mutual interference between two systems. In [36], Bokobza et al. developed a novel deep Q-learning algorithm to learn the optimal sensing and access policy simultaneously, aiming at maximizing the long-term throughput of SU. However, they only focused on the single-user scenarios without considering mutual impact among different SUs, and the mutual guidance and close relationship between sensing window and power allocation strategies are not considered in depth. Focusing on such a research gap in existing works, in this paper, we jointly investigate both partial spectrum sensing and access power allocation for multi-user networks. Considering the dynamic and uncertain spectrum environment, and also the mutual interference among SUs, we propose a HMA DRL based solution to learn the optimal sensing and access strategies adaptively.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>The set of PUs</td>
</tr>
<tr>
<td>$N$</td>
<td>The set of SUs</td>
</tr>
<tr>
<td>$B_m$</td>
<td>Bandwidth of the band $m$</td>
</tr>
<tr>
<td>$K$</td>
<td>The size of SW</td>
</tr>
<tr>
<td>$\Phi_{m,t}$</td>
<td>The bands covered by the SW of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$X(\Phi_{m,t})$</td>
<td>The sensed idle bands under SW of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$P_{e,t}$</td>
<td>Transmit power on the idle bands of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$P_{m,n,t}$</td>
<td>Transmit power on the band $m$ of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$P_{max}$</td>
<td>Max transmit power of each SU</td>
</tr>
<tr>
<td>$o_{n,t}$</td>
<td>The sensing result of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$U_{e,t}$</td>
<td>The utility of the sensing stage for SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$\nu_{e,n,t}$</td>
<td>The rate loss of SU $e$ caused by SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$R_{e,n,t}$</td>
<td>The transmission rate of SU $e$ at time slot $t$</td>
</tr>
<tr>
<td>$R_{e,n,t}$</td>
<td>The transmission rate of SU $e$ without the interference from the transmitter $n$</td>
</tr>
<tr>
<td>$U_{a,n,t}$</td>
<td>The utility of the access stage for SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$U_{u,n,t}$</td>
<td>Overall utility function for the whole DSA process</td>
</tr>
<tr>
<td>$S_{e,n,t}$</td>
<td>The state in the sensing stage of SU $n$ at time slot $t$</td>
</tr>
<tr>
<td>$p_{n,t}$</td>
<td>The proportion of power allocated to each SU</td>
</tr>
<tr>
<td>$P_{n,t}$</td>
<td>The proportion of power allocated to each band within SW</td>
</tr>
<tr>
<td>$\tilde{P}_{n,t}$</td>
<td>The adjusted proportion of power allocated to each band within SW based on the sensing result</td>
</tr>
<tr>
<td>$\tilde{p}_{n,t}$</td>
<td>The normalized proportion of power allocated to each band within SW</td>
</tr>
</tbody>
</table>
III. SYSTEM MODEL

A. Network Model

As shown in Fig. 1, we consider a DSA network with one BS, \( M = \{m|m = 1, 2, ..., M\} \) PUs and \( N = \{n|n = 1, 2, ..., N\} \) SUs, in which each PU transmits data on one band [31], corresponding to \( M \) bands available for SUs to opportunistically access. We denote \( B_m \) as the bandwidth of band \( m \). We consider a system operating in a time-slotted way with each time slot divided into two stages, i.e., sensing stage and access stage [35]. As for the sensing stage, considering that spectrum sensing is usually time-energy-consuming, we assume that SU \( n \) can only choose \( K (K < M) \) adjacent bands, i.e., a SW with size \( K \), to sense from all the \( M \) bands at time slot \( t \), where the covered bands are denoted as \( \Phi_{n,t} \). For each SU, the selection of its SW is not fixed, but a decision variable that needs to be determined. After implementing partial spectrum sensing, the state of the bands \( \Phi_{n,t} \) can be obtained, and the sensed idle bands are represented as \( X(\Phi_{n,t}) \). Based on the sensing result, SU \( n \) will allocate power on the idle bands, which is denoted as \( p_{n,t} = \{p_{n,t}^m|\forall m \in X(\Phi_{n,t})\} \). The power allocation on band \( m \) satisfies \( p_{n,t}^m \in [0, P_{max}] \) and \( P_{max} \) is the maximum transmit power at each SU.

B. Communication Model

We apply the widely used Rician channel model to evaluate the channel gain [29]. For the link between the transmitter of SU \( j \) and the receiver of SU \( n \) on band \( m \), it can be expressed as

\[
 h_{j,n,t}^m = \sqrt{\frac{\kappa}{\kappa + 1}} \xi e^{i\delta} + \sqrt{\frac{1}{\kappa + 1}} w_{t}^m ,
\]

where \( \kappa \) is the \( K \)-factor indicating the signal power ratio received from the line-of-sight (LoS) path and scattered paths. \( \delta \) is the phase of arrival signals on LoS path, which follows uniform distribution \( \delta \sim U(0, 1) \). \( w_{t}^m \sim \mathcal{CN}(0, \xi^2) \) denotes a circularly symmetric complex Gaussian random variable with variance \( \xi^2 = 10^{-PL(d_j,n,f_c)/10} \). As for each link, we use WINNER II channel model [29], [37] to evaluate the path loss by

\[
 PL(d, f_c) = PTL + A \cdot \log_{10}(d[m]) + C \cdot \log_{10} \left( \frac{f_c[\text{GHz}]}{5} \right) ,
\]

where \( d \) represents the Euclidean distance between two transceivers and \( f_c \) is the central frequency of the wireless bands. \( PTL \) is the path loss at a reference distance. \( A \) is the path loss exponent and \( C \) is the path loss frequency dependent parameter.

Thus, the signal to interference plus noise ratio (SINR) on band \( m \) at the receiver of SU \( n \) at time slot \( t \) can be expressed as

\[
 \text{SINR}_{n,t}^m = \frac{p_{n,t}^m \cdot h_{n,t}^m}{\sum_{j \in \mathcal{N}, j \neq n} p_{j,t}^m \cdot |h_{j,t}^m|^2 + \sigma_n^2} ,
\]

where \( \sigma_n^2 \) is the noise at the receiver. Correspondingly, the
transmission rate of SU $n$ can be described by

$$R_{n,t} = \sum_{m \in X(\Phi_{n,t})} B_m \cdot \log_2 (1 + \text{SINR}_{n,t}^m).$$

(4)

IV. A JOINT PARTIAL SPECTRUM SENSING AND POWER ALLOCATION SCHEME

To find as many idle bands as possible and maximize the throughput of SUs, we propose a joint scheme to determine SW decisions $\{\Phi_{n,t}\}_{n=1}^N$ in the sensing stage and PA decisions $\{p_{n,t}\}_{n=1}^N$ in the access stage.

A. SW Determination in the Sensing Stage

At the beginning of time slot $t$, since only idle bands are available for access, SU $n$ should determine an appropriate SW $\Phi_{n,t}$ to find idle ones. The sensing result is represented as

$$o_{n,t} = \{o_{n,t}^m \mid \forall m \in \mathcal{M}\},$$

(5)

where $o_{n,t}^m \in \{0, 1, -1\}$ indicates that the band $m$ is idle, not selected, and occupied at time slot $t$, respectively [25]. Since the goal of the sensing stage is to capture as many idle bands as possible based on partial spectrum sensing, we define the utility of the sensing stage for SU $n$ as

$$U^{S}_{n,t} = \sum_{m \in \mathcal{M}} l_n (o_{n,t}^m),$$

(6)

where $l_n (o_{n,t}^m) = 1$ if $o_{n,t}^m = 1$, otherwise, $l_n (o_{n,t}^m) = 0$.

Subject to the size of SW, we take the long-term utility of the sensing stage as the objective and express the SW determination problem as

$$\textbf{P1} : \max_{\Phi} \left\{ \sum_{t=1}^{\infty} \gamma^t \cdot \sum_{n \in \mathcal{N}} U^{S}_{n,t} \right\},$$

(7a)

s.t. $\sum_{m \in \mathcal{M}} o_{n,t}^m = K, \forall n, \forall t,$

(7b)

where $\gamma$ is a discount factor.

B. PA Determination in the Access Stage

For SU $n$, after getting the sensing result $o_{n,t}$ based on SW $\Phi_{n,t}$, it should make a proper PA decision and allocate power to idle bands $X(\Phi_{n,t})$ accordingly. The goal of the access stage is to maximize the throughput of SUs, where the interference management between them is crucial. To achieve that, we define the rate loss as follows.

**Definition 1**: The rate loss of SU $e$ caused by SU $n$ is defined as

$$\nu_{e \backslash n,t} = R_{e \backslash n,t} - R_{e,t},$$

(8)

where $R_{e,t}$ is the transmission rate of SU $e$, and $R_{e \backslash n,t}$ is that without the interference from the transmitter $n$ expressed as

$$R_{e \backslash n,t} = \sum_{m \in X(\Phi_{e,t})} B_m \cdot \log_2 \left( 1 + \frac{p_e^m \cdot |h_{e,c,t}^m|^2}{\sum_{j \in \mathcal{N}, j \neq n} p_j^m \cdot |h_{j,c,t}^m|^2 + \sigma_n^2} \right).$$

(9)

As for SU $n$, we take the difference between its transmission rate and the total rate loss of others caused by it as its utility described by

$$U^{A}_{n,t} = \sum_{m \in X(\Phi_{n,t})} B_m \cdot \log_2 \left( 1 + \text{SINR}_{n,t}^m \right) - \sum_{e \in \mathcal{N}, e \neq n} \nu_{e \backslash n,t}.$$  

(10)

Subject to the maximum transmit power, we take the long-term utility of the access stage as the objective similar to $\textbf{P1}$ and express the PA determination problem as

$$\textbf{P2} : \max_{\mathbf{p}} \left\{ \sum_{t=1}^{\infty} \gamma^t \cdot \sum_{n \in \mathcal{N}} U^{A}_{n,t} \right\},$$

(11a)

s.t. $\sum_{m \in X(\Phi_{n,t})} p_{n,t}^m \leq P_{max}, \forall n, \forall t.$

(11b)

**Remark 1**: To achieve $\textbf{P2}$ based on partial spectrum sensing, for each SU $n$, it needs to determine SW $\Phi_{n,t}$ and perform PA $p_{n,t}$ on the idle bands $X(\Phi_{n,t})$. Note that SW $\Phi_{n,t}$ would determine how many bands are captured to access, and a bad SW in $\textbf{P1}$ may not effectively discover idle bands, resulting in poor PA performance in $\textbf{P2}$. Furthermore, considering the mutual interference among SUs, it is necessary to avoid the overlapping of SWs of different SUs, where the final access performance in $\textbf{P2}$ should be fed back to guide the SW determination in $\textbf{P1}$. Hence, a sophisticated joint design for both two stages is crucial.

C. Joint Design for SW and PA Determinations

Considering the close correlation between two stages, we define an overall utility function for the whole DSA process as

$$U_{n,t} = w_1 \cdot U^{S}_{n,t} + w_2 \cdot U^{A}_{n,t},$$

(12)

where $w_1$ and $w_2$ are the weighting factors of the sensing stage and access stage, respectively. To develop SW and PA strategies sequentially and effectively, we take the long-term overall utility as the objective. This problem is typically hard to solve since solving the long-term utility maximization requires the statistical characteristics of the spectrum environment, which, however, are difficult to acquire in practice, especially under an uncertain and dynamic environment [38]. Furthermore, two stages are closely coupled and affect each other as aforementioned. Therefore, we propose a hierarchical framework to enable better mutual guidance between two stages. To deal with the uncertainty and dynamics, we develop a HMADRL based solution for the JPSPA scheme to enable each SU to select SW and PA autonomously and adaptively based on its partial observation.

V. A SOLUTION BASED ON HIERARCHICAL MULTI-AGENT DEEP REINFORCEMENT LEARNING

To present our solution based on DRL, we first present the basic knowledge of reinforcement learning. Then, we give the details of our proposed solution as shown in Fig. 2, where sensing strategy $\pi_S$ is learned based on Deep Q Network (DQN) algorithm in Tier I and access strategy $\pi_A$ is obtained based on
Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm with an action adjustment (AA) mechanism, called AA-MADDPG, in Tier II. Finally, we describe the whole process implemented in a centralized training and decentralized execution manner.

A. Basic Knowledge of Reinforcement Learning

Reinforcement learning (RL) [39] is an algorithm in which an agent learns in a trial and error manner, mapping from the state \( s_t \) to an action \( a_t \) according to a certain strategy \( \pi^*_a(s_t) \). Then, the agent will get reward \( r_t \) after executing the action. The agent will continuously improve its strategy based on the result of interaction with the environment. It aims to obtain the optimal strategy \( \pi^*_a \) that maximizes the expected future discounted cumulative reward expressed as

\[
Q_{\pi^*_a}(s_t, a_t) = \mathbb{E}_{\pi^*_a} \left[ \sum_{\lambda=0}^{\infty} \gamma^\lambda \cdot r_{t+\lambda+1} \mid s_t, a_t \right],
\]

which is called Q value.

Q-Learning [39] is a classical reinforcement learning algorithm that learns a look-up table, i.e. Q-table, as its 'brain', to guide its choice. The Q-table will be updated by

\[
Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[ r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right],
\]

where \( \alpha \in (0, 1) \) is the learning rate. To achieve the trade-off between exploration and exploitation, an agent can choose an action based on \( \epsilon \)-greedy policy expressed as

\[
a_t = \begin{cases} 
\arg\max_a Q(s_t, a), & \text{with probability } 1 - \epsilon, \\
\text{Random action}, & \text{with probability } \epsilon,
\end{cases}
\]

where \( \epsilon \in [0, 1] \) is the exploration rate.

Since the Q-Learning relies on a look-up table to evaluate all the state-action pairs, it is difficult to deal with the problems with large state and action spaces, which is called the curse of dimensionality [23]. The success of deep learning proves that deep neural networks (DNNs) can be used to achieve a good mapping of a Q value table. Using neural network mapping instead of table lookup can efficiently offer an optimal solution in complex scenarios, leading to what is now called deep reinforcement learning (DRL).

As for our work, considering the multi-band, multi-user scenario, the state and action spaces are very large. Thus, we develop a multi-agent DRL approach to learn the sensing strategy \( \pi_S \) and access strategy \( \pi_A \) in two tiers. Furthermore, by properly designing the reward function, two tiers can be associated to achieve a sophisticated joint design between them.

B. Tier-I: SW Determination Based on DQN

We first present the reinforcement learning framework of the sensing stage as follows.

1) State: Since historical sensing results reflect the statistical characteristics of the spectrum environment to some extent [25], we define those as the state of SU \( n \) at time slot \( t \) as

\[
S_{n,t} = \{o_{n,t-\omega}\}_{\omega=1}^{\Omega},
\]

where \( \Omega \) is the history length, and \( o_{n,t-\omega} \) is the band state observed by SU \( n \) at time slot \( t - \omega \).

2) Action: We define SW \( \Phi_{n,t} \) as the action of sensing stage, which indicates the covered bands. Then, the sensing strategy of SU \( n \) can be described by

\[
\pi^*_S : S_{n,t} \rightarrow \Phi_{n,t}.
\]

3) Reward: As aforementioned, the goal of the sensing stage is to capture as many idle bands as possible as shown in P1. Furthermore, since the power can only be allocated to idle bands within SW, the network performance after spectrum access should be fed back to evaluate and guide the SW selection. Hence, we define the overall utility \( U_{n,t} \) as the reward of the sensing stage as in (12).
Algorithm 1 SW determination based on DQN

1. Initialize: $\alpha$, $L_1$, $\gamma$, $\epsilon$, $F$, $T_s$, $T_e$, $\{D_n\}_{n=1}^N$, $\{\theta_n\}_{n=1}^N$, $\{\theta_n\}_{n=1}^N$, Train=True.
2. for $t = 1, 2, \ldots$ do
3. for $n \in \mathcal{N}$ do
4. if $t \leq \Omega$ then
5. Randomly choose action $\Phi_{n,t}$ and update the state $S_{n,t}$.
6. end if
7. if $t > \Omega$ then
8. Convert the state $S_{n,t}$ into a vector and input to the DNN. Choose action $\Phi_{n,t}$ according to the $\epsilon$-greedy policy.
9. end if
10. end for
11. if $t \leq T_e$ then
12. Linearly decrease $\epsilon$ from 0.9 to 0.01.
13. end if
14. for $n \in \mathcal{N}$ do
15. Get reward $U_{n,t}$ and next state $S_{n,t+1}$ after performing partial spectrum sensing and access based on SW and PA, respectively.
16. Store $\{S_{n,t}, \Phi_{n,t}, U_{n,t}, S_{n,t+1}\}$ into memory $D_n$ for SU $n$.
17. if $t \geq T_s$ and Train then
18. Randomly sample $L_1$ experiences from the memory $D_n$ to train the main network by minimizing the loss function as (19).
19. if $t \mod F = 0$ then
20. Update the parameters $\hat{\theta}_n$ of the target network through:
21. $\hat{\theta}_n \leftarrow \theta_n$
22. end if
23. end if
24. end for
25. end for
26. Algorithm terminates

As aforementioned, the state space of sensing stage is very large. Hence, we apply the deep Q-network (DQN) algorithm [24] to determine the SW. In DQN algorithm, two DNNs with the same structure are used to learn the decision strategy, namely, main network $Q^S$ with parameter $\theta$ and target network $\bar{Q}^S$ with parameter $\hat{\theta}$. For each SU $n$, the main network can map from the state to the Q-value of all actions. The target network is used to obtain the target value in the sensing stage represented as

$$y_{n,t}^S = U_{n,t} + \gamma \cdot \max_{\Phi_n} \bar{Q}_n^S(S_{n,t+1}, \Phi_n; \hat{\theta}_n), \quad (18)$$

which is utilized to update parameters of the main network by minimizing the loss function expressed as

$$L(\theta_n) = \mathbb{E} \left[ \left( y_{n,t}^S - Q_n^S(S_{n,t}, \Phi_n; \theta_n) \right)^2 \right]. \quad (19)$$

The process of SW determination based on DQN is summarized in Algorithm 1. Since the state of sensing stage is the historical sensing results of the past $\Omega$ time slots, each SU $n$ will randomly choose the SW action $\Phi_{n,t}$, observe the selected bands to obtain the sensing result $o_{n,t}$ and update the state $S_{n,t+1}$ in the first $\Omega$ time slots. After $\Omega$ time slots, SU $n$ will choose SW according to $\epsilon$-greedy policy formulated as in (15), and the exploration rate $\epsilon$ will decrease linearly as time goes by during the exploration time $T_e$ [40]. After performing PA on idle bands $X(\Phi_{n,t})$ in Tier-II, the overall utility $U_{n,t}$ and the next state $S_{n,t+1}$ can be obtained. The tuple $\{S_{n,t}, \Phi_{n,t}, U_{n,t}, S_{n,t+1}\}$ will be stored into the memory $D_n$ for network training. Specifically, after the start of training time, $T_s$, DNNs will be trained by randomly sampling a mini-batch with size $L_1$ from the memory $D_n$, and the parameters of the main network will be updated by minimizing the loss function formulated as in (19). In addition, the target network will be updated by copying the parameters from the main network every $F$ time slots, where $F$ is the update frequency.

C. Tier-II: PA Determination based on AA-MADDPG

The reinforcement learning framework of the access stage is as follows.

1) State: After sensing stage, the sensing result $o_{n,t}$ can be obtained based on $\pi_n^S$. We define $o_{n,t}$ as the state of SU $n$ at time slot $t$ in the access stage [24].

2) Action: After getting the sensing result $o_{n,t}$, each SU should make a proper PA decision on the idle bands within SW. Considering the fixed size of SW and the stability of DRL [41], we take the proportion of power allocated to each band within SW as its action described by

$$\bar{p}_{n,t} = \left\{ \frac{p_{n,t,m}}{P_{\text{max}}} \mid \forall m \in \Phi_{n,t} \right\}. \quad (20)$$

Note that the size of SW is fixed, and thus the dimension of the action is fixed as well. However, since the spectrum environment is dynamic, the number of idle bands within the SW is uncertain. Hence, to perform PA only on idle bands, the PA action needs to be further adjusted based on the sensing result. The details will be presented later.

3) Reward: As for the access stage, considering the throughput and interference issues [33], [42], we take the utility of access stage $U_{n,t}$ as its reward function as in (10).

Considering the continuous action space and mutual interference among SUs, we develop the MADDPG algorithm [43] for the access stage to determine the PA action. In MADDPG algorithm, there are four DNNs, namely, Actor Main Network (AMN) with parameters $\theta_\mu$, Actor Target Network (ATN) with parameters $\hat{\theta}_\mu$, Critic Main Network (CMN) with parameters $\theta_\pi$, and Critic Target Network (CTN) with parameters $\hat{\theta}_\pi$.

a) Actor: For each SU $n$, its AMN is used to determine the action $\bar{p}_{n,t} = \pi_{\theta_\mu}(o_{n,t}) + \mathcal{N}(0, \sigma_\mu^2)$, where $\mathcal{N}$ is the exploration noise and the variance $\sigma_\mu^2$ will decrease as time goes by. The parameters of AMN can be updated based on the policy gradient expressed as

$$\nabla_{\theta_\mu} J(\theta_\mu) = \mathbb{E} \left[ \nabla_{\pi_{\theta_\mu}} Q_n(o_t, \bar{p}_t; \theta_\pi) \right] \nabla_{\theta_\mu} \pi_{\theta_\mu}(o_{n,t}) \Big|_{o_{t}=\{o_{n,t}\}_{n=1}^N, \bar{p}_t=\{\bar{p}_{n,t}\}_{n=1}^N}, \quad (21)$$

where $o_t = \{o_{n,t}\}_{n=1}^N$ and $\bar{p}_t = \{\bar{p}_{n,t}\}_{n=1}^N$ are the global state and action, respectively. ATN is used to train the Critic, which
Algorithm 2 Action Adjustment Mechanism for MADDPG

1: Input: $o_{n,t}$, $p_{n,t}$
2: Output: $\bar{p}_{n,t}$
3: for $n \in N$ do
4:     Input the state of the access stage $o_{n,t}$ to the actor and output an action $\pi_{\theta_\mu}^D(o_{n,t})$.
5:     Input $o_{n,t}$ and the action $\tilde{p}_{n,t} = \pi_{\theta_\mu}(o_{n,t}) + \mathcal{L}(0, \sigma_\mu^2)$ to the action adjuster, where $\mathcal{L}$ is the exploration noise.
6:     Adjust the proportion of power allocated to the occupied bands to 0 and those of the other bands are the same as $p_{n,t}$. Obtain the adjusted vector $p_{n,t} = \{\tilde{p}_{n,t}^m | \forall m \in \Phi_{n,t}\}$.
7:     Normalize the adjusted vector $p_{n,t}$ as in (26) and obtain $\bar{p}_{n,t}$.
8: end for
9: Algorithm terminates

will be introduced later. The parameters of ATN will be updated in a soft manner every $T_a$ time slots according to

$$\theta_\mu^D \leftarrow \tau_\mu \theta_\mu^D + (1 - \tau_\mu)\theta_\mu^\mu,$$  \hspace{1cm} (22)

where the forgetting factor of ATN satisfies $0 < \tau_\mu \ll 1$.

b) Critic: For each SU $n$, it has two critic networks, namely, CMN $Q^A_n$ and CTN $Q^A_n$, and the critic networks can guide the actor networks to choose proper actions. Specifically, the output of its CMN is the Q-value $Q^A_n(o_{n,t}, \check{p}_t; \theta_\mu^D)$ of the global state-action pair, and the parameters of CMN are updated by minimizing the loss function expressed as

$$L(\theta_\mu^D) = \mathbb{E}[{(y_{n,t} - Q^A_n(o_{n,t}, \check{p}_t; \theta_\mu^D))^2}].$$  \hspace{1cm} (23)

$y_{n,t}^A$ is the target value in the access stage expressed as

$$y_{n,t}^A = U_{n,t}^A + \gamma \cdot \tilde{Q}_n^A(o_{n,t+1}, \tilde{p}_{n,t+1}; \theta_\mu^D),$$  \hspace{1cm} (24)

where $\tilde{Q}_n^A(o_{n,t}, \check{p}_{n,t}; \theta_\mu^D)$ is the output of CTN. $\check{p}_{n,t+1}$ is the output of ATN, which can be expressed as $\check{p}_{n,t+1} = \{\pi_{\theta_\mu}(o_{n,t+1})\}_{n=1}^N$. The parameters of CTN will be updated every $T_c$ time slots according to

$$\theta_\varphi^D \leftarrow \tau_\varphi \theta_\varphi^D + (1 - \tau_\varphi)\theta_\varphi^\varphi,$$  \hspace{1cm} (25)

where the forgetting factor of CTN satisfies $0 < \tau_\varphi \ll 1$.

Note that only idle bands within SW are available to perform PA, and the output of the Actor network is the proportion of power allocated to each band within SW. To ensure that each SU only allocates power to idle bands, we design an action adjustment (AA) mechanism for MADDPG algorithm to adjust the action output by the Actor network, called AA-MADDPG here. Specifically, for each SU $n$, it can get the sensing result $o_{n,t}$ after the sensing stage. After inputting the sensing result to the ANN, the action $\tilde{p}_{n,t}$ can be obtained. First, input $o_{n,t}$ and $\tilde{p}_{n,t}$ to AA, the proportion of power allocated to the occupied bands is set to 0, and that to other bands is the same as $p_{n,t}$. We denote the adjusted vector as $p_{n,t} = \{\tilde{p}_{n,t}^m | \forall m \in \Phi_{n,t}\}$. Then, to ensure that the sum of the power allocated to the idle bands is $P_{\text{max}}$, AA would adjust the action $\tilde{p}_{n,t}$ to $\bar{p}_{n,t}$ by normalizing its proportions. In particular, the final output of

Algorithm 3 PA determination based on AA-MADDPG

1: Initialize: $D_2$, $L_2$, $\gamma$, $\alpha^\mu$, $\alpha^\varphi$, $\sigma_\mu^2$, $\tau_\varphi$, $T_a$, $T_c$, $\{\theta_\mu^D_{m=1}^N\}$, $\{\theta_\mu^\mu_{m=1}^N\}$, $\{\theta_\varphi^D_{p=1}^N\}$, $\{\theta_\varphi^\varphi_{p=1}^N\}$, Train = True.
2: for $t = 1, 2, 3, \ldots$ do
3:     $\sigma_\mu^2 \leftarrow \max\{\sigma_\mu^2 \times 0.9999, 0.005\}$
4:     for $n \in N$ do
5:         Based on the sensing result $o_{n,t}$, SU $n$ selects action $p_{n,t} = \pi_{\theta_\mu}(o_{n,t}) + \mathcal{L}(0, \sigma_\mu^2)$.
6:         Adjust the action based on AA presented in Algorithm 2, and obtain the adjusted PA action $\bar{p}_{n,t} = \mathcal{F}(o_{n,t}, p_{n,t})$.
7:         Perform PA based on $\bar{p}_{n,t}$.
8: end for
9: for $n \in N$ do
10:     Receive the reward $U_{n,t}^A$ and obtain $o_{n,t+1}$ after the next sensing stage.
11: end for
12: Store experience $\{o_{n,t}, \bar{p}_{n,t}, U_{n,t}^A, o_{n,t+1}\}_{n=1}^N$ in the memory $D_2$ for the training of the actor and critic networks.
13: if $t \geq T_a$ and Train then
14:     Randomly sample a mini-batch with $L_2$ size from the memory $D_2$ for networks’ training.
15:     Update the parameters of AMN based on the policy gradient formulated as in (21).
16:     Update the parameters of CMN by minimizing the loss function formulated as in (23).
17:     if $t \mod T_a = 0$ then
18:         Update the parameters of ATN according to (22).
19:     end if
20:     if $t \mod T_c = 0$ then
21:         Update the parameters of CTN according to (25).
22:     end if
23:     end if
24: end for

AA can be expressed as

$$\bar{p}_{n,t} = \left\{\frac{\hat{p}_{n,t}^m}{\sum_{m \in \Phi_{n,t}} \hat{p}_{n,t}^m} | \forall m \in \Phi_{n,t}\right\}. \hspace{1cm} (26)$$

We denote the adjusted PA action as $\bar{p}_{n,t} = \mathcal{F}(o_{n,t}, \tilde{p}_{n,t})$, and the process of AA is summarized in Algorithm 2.

After performing PA based on $\bar{p}_{n,t}$, the transmission rate $R_{n,t}$ can be obtained. Then, each SU $n$ uploads its obtained parameters, i.e., $\{o_{n,t}, \bar{p}_{n,t}, R_{n,t}, h_{n,t}\}$, to the central server, which will calculate the utility $U_{n,t}^A$ according to (10). Based on the uploaded parameters from SUs, each experience tuple $\{o_{n,t}, \bar{p}_{n,t}, U_{n,t}^A, o_{n,t+1}\}$ will be stored in the memory $D_2$ for the training of Actor and Critic networks. Specifically, when the number of experience tuple is larger than $T_a$, AMN and CMN will be trained by randomly sampling a mini-batch experience tuples with size $L_2$ from the memory $D_2$. The central server will calculate the mini-batch gradient and update the networks based on stochastic gradient descent (SGD) method, which can be achieved by minimizing the loss function [44], [45]. The parameters of ATN and CTN will be updated in a soft manner based on (22) and (25), respectively. The process of
PA determination is summarized in Algorithm 3\textsuperscript{1}.

Remark 2: In the original MADDPG algorithm, the action stored in the memory is the action output by AMN plus the exploration noise to realize the exploration of the action space. In our scenario, the action stored in the memory is $\hat{p}_{n,t}$, and not the adjusted action $\hat{p}_{n,t}$. As aforementioned, due to the fixed size of SW and the uncertainty of the available bands, it is necessary to design an AA to adjust the action based on the sensing result. Taking the sensing result $o_{n,t}$ and the PA action $\hat{p}_{n,t}$ as input to such an AA, it will uniquely determine a feasible action executed in the environment, so that each SU can perform PA only on the idle bands within its SW. Fortunately, based on the DRL based approach, each SU is able to learn such one-to-one mapping and obtain the feasible PA decision to improve the throughput of SUs. Furthermore, through such a hierarchical framework with AA, the network performance after spectrum access could also be fed back to guide the SW determination.

D. Implementation Based on Centralized Training and Decentralized Execution

The proposed HMADRL based solution can be implemented in a centralized training and decentralized execution manner, and each SU can make the optimal SW and PA decisions autonomously and adaptively to improve its throughput based on its partial observation. As shown in Fig. 3, each SU $n$ is equipped with one DQN’s main network and one AMN with AA locally for SW and PA determining, and all the deep neural networks (DNNs) will be deployed at the central server, e.g., BS, for training. The main computational tasks come from the training of DNNs. Since the training process can be performed by the central server and the distributed SUs only need to carry out SW and PA inference, the computational complexity can be handled just as in [33]. The detailed computational complexity analysis will be presented in Section V-E. Next, the whole implementation process is presented briefly as follows.

- **Tier-I:** At time slot $t$, each SU $n$ first obtains the current state of sensing stage $S_{n,t}$, which is the historical sensing results. Then, it determines a SW $\Phi_{n,t}$ through the local DQN’s main network. After performing partial spectrum sensing, the observation $o_{n,t}$ and the utility of sensing stage $U_{n,t}^S$ can be obtained. The observation $o_{n,t}$ will play two roles in **Tier II:** (a) Determine a PA action based on the actor network by taking $o_{n,t}$ as the input; (b) Adjust the determined PA action according to $o_{n,t}$ based on AA mechanism.

- **Tier-II:** Based on $o_{n,t}$ from Tier I, each SU $n$ obtains a PA action from AMN, namely, $\hat{p}_{n,t} = \pi_{\theta_{n,t}}(o_{n,t}) + \mathcal{L}(0, \sigma_{n,t}^2)$. Based on $o_{n,t}$ and $p_{n,t}$, AA adjusts $p_{n,t}$ to a feasible one $\hat{p}_{n,t}$ according to Algorithm 2, ensuring that each SU can only perform PA on idle bands within SW.

After performing SW and PA in two tiers, each SU $n$ will upload its obtained observation parameters to the central server, and there is no need to exchange information among SUs. Then, the central server will update all the DNNs by randomly sampling mini-batch data from the corresponding memory. SUs only need to download the parameters of DQN’s main network and AMN from the central server periodically to update their local networks to make SW and PA decisions.

E. Computational Complexity Analysis

The computational complexity of the proposed HMADRL approach mainly comes from two parts, namely, inference process and training process, which depends on the architecture of the deep neural networks (DNNs) [46], [47]. Next, we analyze the computational complexity for the fully connected DNNs. For the inference process, we denote $N_{\text{max}}^\text{neu}$ as the number of neurons for the widest layer in the DNN with $W$ layers. According to [47], the complexity can be expressed as

$$O(\cdot) = O\left(W(N_{\text{max}}^\text{neu})^2\right).$$

(27)

For the training process, the computational complexity depends on both forward and backward propagation in the DNNs. For the fully connected DNNs with $W$ layers, we denote the number of neurons in layer $w$ as $N_{\text{neu}}^w$. According to [47], the computational complexity of the forward propagation can be calculated by

$$O(\cdot) = O\left(W(N_{\text{neu}}^w)^W\sum_{w=2}^W N_{\text{neu}}^w N_{\text{neu}}^{w-1} N_{\text{neu}}^{w-2} + \sum_{w=1}^W N_{\text{neu}}^w\right),$$

(28)

where $\sum_{w=1}^W N_{\text{neu}}^w$ represents the number of activation functions employed in a DNN, and the rest of the expression is the number of multiplications performed. The computational com-
The computational complexity of the backward propagation can be calculated as

$$O(T_{ \text{bwd}}) = O\left( N_{\text{neu}}N_1^1 + 2 \sum_{w=2}^{W} N_{\text{neu}}N_{w-1}^1N_{w-2}^2 + W(W-1) \right), \quad (29)$$

where $O\left( \sum_{w=2}^{W} N_{w}^1N_{w-1}^1N_{w-2}^2 + W(W-1) \right)$ is the amount of computation required for the gradient operations during the backward propagation. Thus, the computational complexity of the training process can be represented as

$$O(T_{\text{train}}) = O\left( (T_{\text{bwd}} + T_{\text{bwd}}) \right). \quad (30)$$

Since the proposed HMADRL based solution is implemented in a centralized training and decentralized execution manner, each SU is equipped with a DQN’s main network and an AMN with AA and all the DNNs, i.e., DQN, Actor, and Critic networks, are trained at the central server. The computational complexity of a SU can be expressed as

$$O(T_{\text{local}}) = O\left( 2T_{\text{inference}} \right). \quad (31)$$

The computational complexity of a central server can be represented by

$$O(T_{\text{central}}) = O\left( 3N \cdot T_{\text{train}} \right). \quad (32)$$

Thus, the overall computational complexity of the proposed HMADRL can be expressed as

$$O_{\text{HMADRL}} = \varpi_1O\left(T_{\text{central}}\right) + \varpi_2O\left(T_{\text{local}}\right), \quad (33)$$

where $\varpi_1$ and $\varpi_2$ represent the number of iterations taking to converge in the training and inference process, respectively.

VI. SIMULATION RESULTS AND DISCUSSIONS

A. Simulation Setup

As shown in Fig. 4, we consider a network consisting of $N = 3$ SUs in a $200 \times 200$ m$^2$ area. The locations of different SUs’ transmitters are randomly generated in the area. For the receivers, to verify the effectiveness of the proposed scheme on the interference management among SUs, we consider the worst case scenario that the receiver of each SU is close to someone’s transmitter. Specifically, we set the receiver of each SU located 40-80m away from the transmitter of some other SU. We consider $M = 30$ bands available for SUs to access. Among these bands, we set 8 bands with high probability to be idle, which are called potential bands here, and each SU can select $K = 3$ adjacent bands to sense. The idle transition probabilities $P_{\text{t}}(0 \rightarrow 1, 1 \rightarrow 0)$ of these potential bands are randomly set within the range $[0, 1]$, and those of other bands are set within the range $[0, 0.2]$ [29]. We generate 50000 simulated data following the transition probabilities as the spectrum environment to evaluate our proposed solution, 100 of which are shown in Fig. 5. Other simulation parameters are presented in Table II.

B. Results and Discussions Based on Simulated Spectrum Data

We compare our proposed scheme with the following five policies.

- Sense Only (SOL): Each SU learns SW strategy based on the DQN algorithm, and chooses PA randomly in the access stage.
- Access Only (AOL): Each SU learns PA strategy based on the AA-MADDPG algorithm, and chooses SW randomly in the sensing stage.

![Fig. 5. Simulated spectrum data.](image-url)

**TABLE II**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max transmit power $P_{\text{max}}$</td>
<td>60dbm [31]</td>
</tr>
<tr>
<td>Noise $n^2$</td>
<td>-114486 [32]</td>
</tr>
<tr>
<td>Bandwidth of the band $B_{\text{max}}$</td>
<td>2MHz [31]</td>
</tr>
<tr>
<td>Path loss exponent $\alpha$</td>
<td>22.7</td>
</tr>
<tr>
<td>Path loss frequency dependent parameter $C$</td>
<td>20</td>
</tr>
<tr>
<td>$K$-factor $c$</td>
<td>8</td>
</tr>
<tr>
<td>Path loss at a reference distance $P_{\text{ref}}$</td>
<td>41</td>
</tr>
<tr>
<td>Carrier frequency of wireless bands $f_c$</td>
<td>5GHz [29]</td>
</tr>
<tr>
<td>Weighting factor of sensing stage $w_1$</td>
<td>18</td>
</tr>
<tr>
<td>Weighting factor of access stage $w_2$</td>
<td>100</td>
</tr>
<tr>
<td>Start of training time $T_{\text{start}}$</td>
<td>1,000</td>
</tr>
<tr>
<td>Exploration time $T_{\text{expl}}$</td>
<td>35,000</td>
</tr>
<tr>
<td>Exploration rate $\epsilon$</td>
<td>0.9 → 0.01 Linearly</td>
</tr>
<tr>
<td>History length $\Omega$</td>
<td>16 [25]</td>
</tr>
<tr>
<td>Discount rate $\gamma$</td>
<td>0.9</td>
</tr>
<tr>
<td>Learning rate of DQN $\alpha$</td>
<td>0.01</td>
</tr>
<tr>
<td>Number of neurons in DQN</td>
<td>[128,256,128] [40]</td>
</tr>
<tr>
<td>Activation function of DQN</td>
<td>ReLu,ReLu,ReLu</td>
</tr>
<tr>
<td>Update frequency of the target network $F$</td>
<td>200</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
<tr>
<td>Batch size $L_0$</td>
<td>256</td>
</tr>
<tr>
<td>Learning rate of AMN $\alpha^*$</td>
<td>0.0001</td>
</tr>
<tr>
<td>Learning rate of CMN $\alpha^*$</td>
<td>0.0005</td>
</tr>
<tr>
<td>Variance of the exploration noise $\sigma_0^2$</td>
<td>2</td>
</tr>
<tr>
<td>Forgetting factor of ATN $\gamma_t^*$</td>
<td>0.01</td>
</tr>
<tr>
<td>Forgetting factor of CTN $\gamma_t^*$</td>
<td>0.01</td>
</tr>
<tr>
<td>Number of neurons in actor</td>
<td>[64,64]</td>
</tr>
<tr>
<td>Number of neurons in critic</td>
<td>[64,64]</td>
</tr>
<tr>
<td>Activation function of actor</td>
<td>ReLu,Segmoid</td>
</tr>
<tr>
<td>Activation function of critic</td>
<td>ReLu,ReLu</td>
</tr>
<tr>
<td>Batch size $L_2$</td>
<td>256</td>
</tr>
<tr>
<td>Updating cycle of actor $T_{\text{a}}$</td>
<td>1</td>
</tr>
<tr>
<td>Updating cycle of critic $T_{\text{c}}$</td>
<td>1</td>
</tr>
</tbody>
</table>
Hierarchical Single-Agent DRL (HSADRL): Each SU independently learns SW strategy based on the DQN algorithm and PA strategy based on the DDPG algorithm in [41].

- All Random (ARN): Each SU chooses SW and performs PA randomly at each time slot.
- Ideal Case: Assuming that the statistical characteristics of all bands are known by SUs, and each SU directly chooses different potential bands to sense and access the idle ones without overlapping with each other.

First, to demonstrate the performance of the sensing stage, we define the capture rate to represent the proportion of the idle bands captured by SUs as

$$\text{capture rate} = \frac{\#SW_{idle}}{\min\{K \times N, \#Total_{idle}\}},$$  \hspace{1cm} \text{(34)}$$

where $\#SW_{idle}$ is the total number of idle bands that SUs capture. $\#Total_{idle}$ is the total number of idle bands in $M$ ones. Fig. 6 shows the average capture rate, average throughput, average interference, and average total reward under different policies, respectively. The average means the results are the average of the previous 199 time slots [33], and the interference in Fig. 6(c) is the sum of the interference at each SU. The results in Fig. 6 show that our solution can achieve a higher capture rate and throughput, and also effectively control the interference without affecting PUs, which demonstrates that

Fig. 6. Performance under different schemes.

Fig. 7. Average capture rate under different spectrum data.
our solution can efficiently learn the dynamics of the spectrum environment and improve the performance. Our solution also outperforms SOL and AOL, which are two schemes that only consider SW or PA separately. This indicates that the joint design of SW and PA is beneficial for SUs to optimize their performance. By jointly designing SW and PA, our solution can learn the SW and PA simultaneously and improve each other's strategies more efficiently. Compared with HSADRL, which is another joint design scheme based on DDPG, our solution has a faster convergence speed and a higher performance. This indicates that our solution can better capture the dynamics of the spectrum environment and other SUs' decisions, and learn more accurate and efficient SW and PA strategies to better manage the interference among them.

Next, we investigate the effect of different statistical characteristics of the spectrum data on the average capture rate, as shown in Fig. 7. In Fig. 7(a), we increase the number of bands $M$, while keeping the number of potential bands constant at eight. A larger $M$ represents a more complex spectrum environment, which poses a challenge for SUs to capture idle bands. The result shows that the performance can reach a higher level despite the increased dimensions of the state and action spaces, showing the scalability of our proposed solution. In Fig. 7(b), we change the number of potential bands, while keeping the idle transition probabilities unchanged from the previous setting. The result shows the average of the final 100 time slots, which can represent the final convergence level. We notice that a larger number of potential bands makes the spectrum environment more favorable and easier for the SUs to capture idle bands. Consequently, our proposed solution can converge to a higher level, approaching to the Ideal Case in all scenarios. Additionally, our proposed solution surpasses HSADRL, demonstrating its ability to better exploit the potential bands and avoid collisions among SUs. Similarly, in Fig. 7(c), we alter the idle transition probabilities of potential bands. As the idle transition probabilities of potential bands increase, it becomes easier to capture the idle ones, leading to the increase in the capture rate. Furthermore, although the idle transition probabilities are relatively low, our proposed solution can still maintain higher capture rate than HSADRL, which further shows the effectiveness and robustness of our proposed solution.

Finally, to evaluate the adaptive nature of the DRL based solution, we change the spectrum environment at time slot 25000 and show the average capture rate, average total reward, and average throughput in Fig. 8. Specifically, at time slot 25000, we decrease the idle transition probabilities of the eight original potential bands from $[0, 0.8, 1]$ to $[0, 0.2]$, and also select eight from 22 non-potential bands and increase their idle transition probabilities of the selected bands from $[0, 0.2]$ to $[0, 0.8, 1]$. The results show that the performance increases steadily during the first 25000 time slots. However, at time slot 25000, the spectrum environment changes abruptly, making the learned SW and PA strategies unsuitable for the new environment and causing a sharp drop in the performance. Nevertheless, the performance recovers quickly and converges to a relatively high level, demonstrating the adaptation of our proposed solution. In other words, our proposed solution can enable each SU to capture the statistical characteristics of the spectrum environment and adjust its strategies autonomously and adaptively.

C. Results and Discussions Based on Real Spectrum Data

To further evaluate the scalability and effectiveness of our proposed scheme and make the experiment more practical, we use SAM-60BX to collect real spectrum data for the experiment. The spectrum measurement is carried at 17:00 on March 25, 2023 in the Haishan Building of Dalian University of Technology (38°52′49″N, 121°31′41″E). The measurement covers 50 bands, ranging from 2130MHz to 2230MHz. The sample interval of SAM-60BX is set to 10ms. The measurement lasts for 500 seconds, and we obtain 50000 samples for each band, which are further processed to 0 (idle) or 1 (busy) by setting the power threshold $P_{th} = -110$dBm. Fig. 9(a) shows the measurement results.

Based on the collected real spectrum data, we conduct experiments in four different scenarios, i.e., 5 Users-50 Bands, 5 Users-50 Bands, 3 Users-30 Bands, and 3 Users-50 Bands. 30 bands is the real spectrum data selected from 50 bands, ranging from 2130MHz to 2190MHz. The network topologies of 3 Users and 5 Users are shown in Fig. 4 and Fig. 9(b), respectively.

Fig. 10 shows the average total throughput, interference, and capture rate under different numbers of users and bands. The
results indicate that our proposed scheme can maintain a high level of throughput and capture rate, and keep the interference level low even when the number of users and bands increases. This suggests that our scheme can adapt to different network scenarios and scale well with the network size. Moreover, we observe that the 50 bands scenario reaches a higher level of performance earlier than the 30 bands scenario. This is because more bands can provide SUs with more access opportunities, which means they can find idle bands more easily to improve the throughput and capture rate of SUs.

VII. CONCLUSION

In this paper, we have proposed a joint partial spectrum sensing and power allocation scheme to help each SU choose the best sensing window and power proportions to achieve high throughput. By considering the dynamics and uncertainty of the spectrum environment, we have developed a viable solution based on hierarchical multi-agent deep reinforcement learning. Through the hierarchical framework, the mutual guidance between spectrum sensing and access can be well achieved. Furthermore, the multi-agent deep reinforcement learning algorithm enables each SU to learn the statistical characteristics of the spectrum environment and adjust its sensing window and power proportions autonomously and adaptively based on its local observation. Simulation results based on simulated spectrum data and real spectrum data measured by SAM-60BX are presented to show the effectiveness of the joint design and our proposed solution. It enables SUs to achieve relatively high capture rate and throughput while well managing the mutual interference.

REFERENCES


Xuanheng Li (S’13-M’18) received the Ph.D. degree in communication and information system in 2018, from the Dalian University of Technology, Dalian, China. From 2015 to 2017, he was with the Department of Electrical and Computer Engineering at the University of Florida as a visiting scholar. He is currently an Associate Professor at the Dalian University of Technology. He was a recipient of the Best Paper Award at the IEEE Globecom 2015. His current research interests include intelligent communications, wireless sensing, UAV-assisted networks, mobile edge computing and caching.

Haichuan Ding received the B.Eng. and M.S. degrees in electrical engineering from the Beijing Institute of Technology (BIT), Beijing, China, in 2011 and 2014, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2018. From 2012 to 2014, he was with the Department of Electrical and Computer Engineering, the University of Macau, as a Visiting Student. He is currently a professor with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China. His current research interests include covert communications and space-air-ground integrated networks.

Yulong Zhang (S’22) received the B.S. degree from the Dalian University of Technology, China, where he is currently pursuing the graduate degree with the School of Information and Communication Engineering. His current research interests include cognitive communications and wireless sensing.

Yuguang Fang (S’92-M’97-SM’99-F’08) received an MS from Qufu Normal University, China in 1987, a PhD from Case Western Reserve University in 1994, and a PhD from Boston University in 1997. He joined the Department of Electrical and Computer Engineering at University of Florida in 2000 as an assistant professor, then was promoted to associate professor in 2003, full professor in 2005, and distinguished professor in 2019, respectively. Since 2022, he has been the Chair Professor of Internet of Things with Department of Computer Science at City University of Hong Kong. Dr. Fang received many awards including the US NSF CAREER Award, US ONR Young Investigator Award, 2018 IEEE Vehicular Technology Outstanding Service Award, IEEE Communications Society AHSN Technical Achievement Award (2019), CISTC Technical Recognition Award (2015), and WTC Recognition Award (2014), the Best Paper Award from IEEE ICNP (2006), and 2010-2011 UF Doctoral Dissertation Advisor/Mentoring Award. He was the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2013-2017) and IEEE Wireless Communications (2009-2012) and has served on several editorial boards of premier journals. He also served as the Technical Program Co-Chair of IEEE INFOCOM2014. He is a fellow of ACM, IEEE, and AAAS.