A Partially Observable Deep Multi-Agent Active Inference Framework for Resource Allocation in 6G and Beyond Wireless Communications Networks

Fuhui Zhou[§], Rui Ding[§], Qihui Wu[§], Derrick Wing Kwan Ng[‡], Kai-Kit Wong[⋆], and Naofal Al-Dhahir[†]
[§]Nanjing University of Aeronautics and Astronautics, China, [‡]The University of New South Wales, Australia,
[⋆]University College London, UK, [†]University of Texas, USA.
Email: *zhoufuhui@ieee.org, rui_ding@nuaa.edu.cn, wuqihui2014@sina.com,*
w.k.ng@unsw.edu.au, kai-kit.wong@ucl.ac.uk, and aldhahir@utdallas.edu

***Abstract*—Resource allocation is of crucial importance in wireless communications. However, it is extremely challenging to design efficient resource allocation schemes for future wireless communication networks since the formulated resource allocation problems are generally non-convex and consist of various coupled variables. Moreover, the dynamic changes of practical wireless communication environment and user service requirements thirst for efficient real-time resource allocation. To tackle these issues, a novel partially observable deep multi-agent active inference (PODMAI) framework is proposed for realizing intelligent resource allocation. A belief based learning method is exploited for updating the policy by minimizing the variational free energy. A decentralized training with a decentralized execution multi-agent strategy is designed to overcome the limitations of the partially observable state information. Exploited the proposed framework, an intelligent spectrum allocation and trajectory optimization scheme is developed for a spectrum sharing unmanned aerial vehicle (UAV) network with dynamic transmission rate requirements as an example. Simulation results demonstrate that our proposed framework can significantly improve the sum transmission rate of the secondary network compared to various benchmark schemes. Moreover, the convergence speed of the proposed PODMAI is significantly improved compared with the conventional reinforcement learning framework. Overall, our proposed framework can enrich the intelligent resource allocation frameworks and pave the way for realizing real-time resource allocation.**

***Index Terms*—Deep active inference, intelligent resource allocation, spectrum sharing, trajectory optimization.**

I. Introduction

RESOURCE allocation design is a fundamental problem in wireless communication networks since it can improve the achievable performance (e.g., spectrum efficiency, energy efficiency, etc.) and efficiently make the best of the precious and limited resources [1]. However, it is extremely challenging to design efficient resource allocation schemes for future wireless communication networks. The reasons are from two aspects. On the one hand, in the sixth-generation (6G) and beyond wireless communication networks, multi-domain resources, such as energy, frequency spectrum, computing, caching, etc., are required to be jointly optimized since sensing, communication, and computing functions are envisioned to be integrated [2], which results in challenging non-convex problems and high-dimensional coupling variables. On the other hand, wireless communication environment and user service requirements vary dynamically over time. In this case, it is imperative to achieve real-time resource allocation [3].

To date existing resource allocation methods for wireless communication networks can be mainly classified into two categories, namely, the optimization theory based methods and the intelligent methods based on machine learning (representative deep reinforcement learning (DRL)). The optimization theory based methods mainly adopt conventional mathematical skills such as alternating optimization and successive convex approximation. For instance, in [4], the authors proposed a successive convex approximation based resource allocation algorithm to jointly optimize the transmit power of the secondary base station and the unmanned aerial vehicle (UAV) trajectory for maximizing the total average secrecy rate of the spectrum sharing network with orthogonal frequency division multiplexing. However, resource allocation problems in future wireless communication networks are generally NP-hard [5] and the solutions obtained by adopting the traditional optimization based methods are not globally optimal. Moreover, the high computational complexity and the poor real-time performance cannot adapt to the highly dynamic wireless communication networks.

Recently, DRL has been widely applied to develop intelligent resource allocation schemes due to its outstanding advantages in efficient and rapid processing of large-scale complicated problems [5]. Specifically, the exploration strategy allows the agent to discover new and potentially better solutions that cannot be identified by the traditional optimization based methods [6]. Moreover, the agent can adapt its policy to the time-varying environment, which is important when dealing with dynamic resource allocation problems [7]. Inspired by these properties, in [8], the deep Q-network (DQN) was utilized to optimize the sub-band allocation and the power level in vehicle-to-vehicle communications. However,

This work was supported by National Key R&D Program of China under Grant 2020YFB1807602, the National Natural Science Foundation of China under Grant 62222107, Grant 62071223, Grant 62031012, and Zhejiang Lab Open Research Project under Grant K2022PD0AB09. The corresponding author is Rui Ding.

the power allocation was based on an ideal assumption that the transmit power can be discretized into three levels. Indeed, the instant quantization errors resulted in significant performance degradation. Therefore, the deep deterministic policy-gradient based algorithm was proposed in order to realize continuous action optimization [9].

Despite the promising results revealed in the literatures, the aforementioned works assumed that the complete environmental observations can be obtained. Unfortunately, it is difficult to be achieved because of the limited sensing ability of sensing devices and the finite communication and computing overhead [5]. Thus, in practice, it is of crucial importance to design resource allocation schemes under the partially observable wireless communication environment. Moreover, DRL based methods rely on the environment feedback for optimizing the long-term expected reward. Therefore, DRL based methods have low convergence speed, which limits their practical applications.

In this paper, in order to tackle the above challenges and overcome the drawback of DRL based resource allocation schemes, a novel partially observable deep multi-agent active inference framework (PODMAI) is proposed for intelligent resource allocation. A belief based learning method is exploited for achieving active inference to simulate how agents perceive and act in the real world by minimizing the free energy. A decentralized training with a decentralized execution multi-agent strategy is proposed to further improve the system performance. In order to demonstrate the efficiency of our proposed framework, an intelligent spectrum allocation and trajectory optimization scheme is proposed for spectrum sharing UAV networks. Simulation results demonstrate that our proposed framework can significantly improve the sum transmission rate of the secondary system while satisfying the dynamic transmission rate requirements of the users. Moreover, a faster convergence speed can be obtained compared with the DRL based benchmark scheme.

The remainder of this paper is organized as follows. The architecture of our proposed deep active inference framework is presented in Section II. Section III presents the exploitation of our framework for designing an intelligent spectrum allocation and trajectory optimization scheme in spectrum sharing UAV networks. Section IV presents the simulation results. Finally, the paper concludes with Section V.

II. Architecture of Our Proposed Deep Active Inference Framework

As shown in Fig. 1, motivated by the free energy theory [10], a deep active inference framework is established for intelligent resource allocation. It contains an experience replay memory and four sub-networks, namely, the value network for the expected free energy (EFE) estimation, the policy network to generate the action policy, the target network for target EFE, and the transition network for the current state prediction. After obtaining the action a_t based on the observed state s_t, the immediate reward r_t and the new state s_{t+1} are returned to the agent. Afterwards, the experience obtained at time slot t

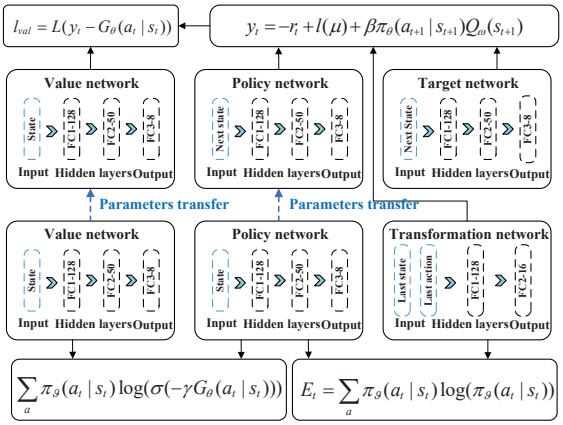


Fig. 1: The architecture of our proposed deep active inference framework for intelligent resource allocation.

and the past state s_{t-1} are stored in the replay memory block. A mini-batch of experiences $(s_{t+1}, s_t, s_{t-1}, a_{t+1}, a_t, r_t)$ with size N_{tr} is randomly selected when the memory block is filled. Compared with the on-policy based methods, the exploitation of mini-batch experiences rather than the current time slot experience is to ensure the independence of the training tuples and avoid the problem caused by the excessive correlation among the tuples.

The active inference based methods rely on the assumption that the agents perceive and act in the environment to minimize the variational free energy, which can be expressed as

$$F_t = \mathbb{E}_{q(s_t)}[\ln(p(o_t|s_t))] + D_{\mathrm{KL}}[q(s_t)||p(s_t|s_{t-1}, a_{t-1})] + D_{\mathrm{KL}}[q(a_t|s_t)||p(a_t|s_t)], \quad (1)$$

where $\mathbb{E}_{q(s_t)}$ is the expectation over the variational density $q(s_t)$ and D_{KL} is the Kullback-Leibler (KL) divergence. The second term is the state prediction error, which is expressed as the KL divergence between the state s_t and the state that is predicted at time step $t-1$. Therefore, the mean squared loss (MSE) is used to replace the KL divergence. Specifically, the value network is used to estimate the EFE, given as $G_\theta(a_t|s_t)$, and θ is the weight parameters of the value network. The target network is used to output the target EFE, given as $Q_\omega(s_{t+1})$, ω is the weight parameters of the target network. According to the output action policy, the target EFE is weighted as $Q_\omega(s_{t+1})\pi_\vartheta(a_{t+1}|s_{t+1})$. Then, the estimated EFE after the policy output $\pi_\vartheta(a_{t+1}|s_{t+1})$ is obtained by bootstrapping, given as

$$y_t = -r_t + l(\mu) + \beta\pi_\vartheta(a_{t+1}|s_{t+1})Q_\omega(s_{t+1}), \quad (2)$$

where ϑ is the weight parameters of the policy network, ω is the weight parameters of the target network and β is the discount rate. $l(\mu)$ is the predicted error of the transition network, given as

$$l(\mu) = \mathcal{L}(A_\mu(s_{t-1}, a_{t-1}), s_t), \quad (3)$$

where $\mathcal{L}$ is the MSE function and μ is the weight parameters of the transition network. The value network loss is used to

show the difference between the predicted EFE and the target EFE, given as

$$l_{val} = \mathcal{L}(y_t, G_\theta(a_t|s_t)). \tag{4}$$

The distribution over actions can be modelled as a precision-weighted Boltzmann distribution over the estimated EFE in the value network, given as $\sigma(-\gamma G_\theta(a_t|s_t))$. σ represents the softmax function. Therefore, the third term in eq. (1) can be expressed as $\sum_a \pi_\vartheta(a_t|s_t) \ln(\sigma(-\gamma G_\theta(a_t|s_t))) + H_t$. Finally, the VFE can be expressed as

$$F_t = \sum_a \pi_\vartheta(a_t|s_t) \ln(\sigma(-\gamma G_\theta(a_t|s_t))) + H_t + l(\mu), \tag{5}$$

where H_t is the entropy, given as

$$H_t = \sum_a \pi_\vartheta(a_t|s_t) \ln(\pi_\vartheta(a_t|s_t)). \tag{6}$$

The value network, target network, transition network, and policy network are jointly updated by minimizing the variational free energy F_t.

Note that the proposed deep active inference framework is tailored for designing resource allocation schemes in wireless communications since it enables the system to actively update the policy and learn the optimal policies for resource allocation under dynamic and partial observations. In order to clearly explain our framework and demonstrate its efficiency, an intelligent joint spectrum allocation and trajectory optimization scheme is proposed in the following section as an example.

III. Intelligent Spectrum Allocation and Trajectory Optimization Enabled by Our Framework

A. System Model and Problem Formulation

As shown in Fig. 2, a dynamic spectrum sharing UAV network is taken as an example, which is a classical scenario in the future wireless communication networks. A practical and emerging case is considered that the user transmission rate requirements are dynamic due to multi-modal communication services, such as audio, video, and haptic transmission [2]. In the network, there are K SUs, J PUs, a primary base station (PBS), and N cognitive UAVs (C-UAVs). In order to serve more SUs and provide better services, a wideband spectrum is divided into M equal sub-carrier frequency bands. Let $k \in \mathcal{K} \triangleq \{1, 2, ..., K\}$, $j \in \mathcal{J} \triangleq \{1, 2, ..., J\}$, and $m \in \mathcal{M} \triangleq \{1, 2, ..., M\}$ denote the set of SUs, PUs, and sub-bands, respectively. To avoid the interference among different PUs, each PU occupies one sub-carrier frequency band.

The three-dimensional Cartesian coordinate system is considered. The horizontal positions of the PBS, the jth PU, the kth SU, and the nth C-UAV are denoted as $\mathbf{w}_b = (x_b, y_b)$, $\mathbf{w}_{p,j} = (x_{p,j}, y_{p,j})$, $\mathbf{w}_{s,k} = (x_{s,k}, y_{s,k})$ and $\mathbf{q}_{c,n} = (x_{c,n}, y_{c,n})$, respectively. Without loss of generality, it is assumed that the UAVs fly at a constant vertical height H_u. The total transmission time interval is within a duration of T, and T is divided into L equal-length time intervals, where each time interval is given by $\delta_t = \frac{T}{L}$. The status of the UAVs

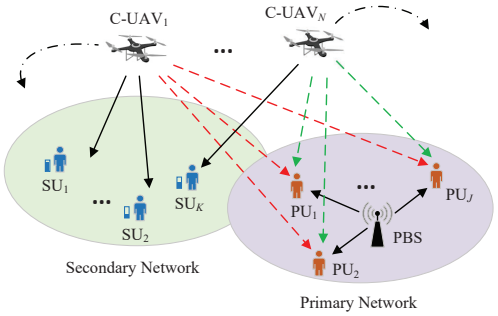


Fig. 2: The spectrum sharing UAV network with dynamic user requirements.

can be regarded as static since δ_t is sufficiently small [13]. The dynamic position of the nth C-UAV can be formulated as

$$x_{c,n}[t+1] = x_{c,n}[t] + v_{c,0} \cos(\phi_{c,n}[t]), \tag{7a}$$
$$y_{c,n}[t+1] = y_{c,n}[t] + v_{c,0} \sin(\phi_{c,n}[t]), \tag{7b}$$

where $\phi_{c,n}[t]$ and $v_{c,0}$ represent the direction of the C-UAV at time step t and the fixed flying speed, respectively. The distance between the nth C-UAV and the kth SU, and that between the PBS and the jth PU are respectively given as

$$d^c_{k,n}[t] = \sqrt{\|\mathbf{q}_{c,n}[t] - \mathbf{w}_{s,k}[t]\|^2 + H_u^2}, \tag{8a}$$
$$d^p_j[t] = \sqrt{\|\mathbf{w}_b[t] - \mathbf{w}_{p,j}[t]\|^2}. \tag{8b}$$

Similar to the works in [10] and [13], the wireless channel between the UAV and the ground users is dominated by the line of sight link. Let β_{ref} represent the channel power gain at the reference distance of 1 meter. Thus, the channel power gain from the nth C-UAV to the kth SU can be expressed as

$$h^{\text{los}}_{k,n}[t] = \beta_{\text{ref}} d^c_{k,n}[t]^{-2}. \tag{9}$$

The channel model between the ground nodes (PBS, PUs, SUs) is different from that of the air-to-ground link. It is required to consider both the distance-dependent path loss with exponent $\varphi \geq 2$ and small-scale Rayleigh fading [11]. Thus, the channel gain from the PBS to the jth PU is given as

$$h^g_j[t] = \beta_{\text{ref}} d^p_j[t]^{-\varphi} \zeta_j, \tag{10}$$

where $d^p_j[t]$ represents the distance between the PBS and the jth PU at time step t. ζ_j is an exponentially distributed random variable with unit mean accounting for the Rayleigh fading.

Considering the sub-band allocation, the binary variable $\rho_{k,n}[m]$ is used to characterize the allocation strategy of SUs. Specifically, the mth sub-band is used by the kth SU from the nth C-UAV when $\rho_{k,n}[m] = 1$, otherwise, $\rho_{k,n}[m] = 0$. Each PU operates on a preassigned orthogonal spectrum channel with a fixed transmit power. The C-UAV aims to learn the best allocation strategy to access the shared spectrum to meet the dynamic transmission rate requirements of SUs.

The SINR between the nth C-UAV and the kth SU in the mth sub-band can be expressed as

$$\gamma^s_{k,n,m} = \frac{\rho_{k,n}[m] P^c_m h^{\text{los}}_{k,n}}{\sigma^2 + \sum_{j=1}^{J} P^b_m h^g_j}, \tag{11}$$

where P^c_m and P^b_m denote the transmit power of the C-UAV and the PBS in the mth sub-band, respectively. σ^2 is the noise power. Similarly, the SINR between the PBS and the jth PU in the mth sub-band can be expressed as

$$\gamma^p_{j,m} = \frac{\rho_j[m] P^b_m h^g_j}{\sigma^2 + \sum_{n=1}^{N}\sum_{k=1}^{K} \rho_{k,n}[m] P^c_m h^{\text{los}}_{k,n}}. \tag{12}$$

Then, the achievable transmission rate of the kth SU and the jth PU can be respectively expressed as

$$R^s_{k,n,m} = B\log_2(1+\gamma^s_{k,n,m}), \tag{13a}$$
$$R^p_{j,m} = B\log_2(1+\gamma^p_{j,m}), \tag{13b}$$

where B denotes the bandwidth for each sub-band.

In this paper, in order to efficiently utilize the spectrum resource, protect the PUs from harmful interference, and guarantee the dynamic transmission rate requirements of SUs, a joint spectrum resource allocation and UAV trajectory optimization problem is formulated as

$$\mathbf{P}_1: \max_{\boldsymbol{\rho},\boldsymbol{\phi}} \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} R^s_{k,n,m} \tag{14a}$$
$$\text{s.t. C1}: \sum_{m=1}^{M} R^p_{j,m} > R^{\min}_j, \forall j \in \mathcal{J}, \tag{14b}$$
$$\text{C2}: \sum_{m=1}^{M}\sum_{n=1}^{N} R^s_{k,n,m} > R^{\text{thr}}_k, \forall k \in \mathcal{K}, \tag{14c}$$
$$\text{C3}: \rho_{k,n}[m] \in \{0,1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{14d}$$
$$\text{C4}: \sum_{n=1}^{N}\sum_{k=1}^{K} \rho_{k,n}[m] \le 1, \forall m \in \mathcal{M}, \tag{14e}$$
$$\text{C5}: |\phi_{c,n}| \le \pi, \forall n \in \mathcal{N}, \tag{14f}$$

where $\boldsymbol{\rho} = \{\rho_{k,n}[m]\}_{m\in\mathcal{M},n\in\mathcal{N},k\in\mathcal{K}}$. $\boldsymbol{\phi} = \{\phi_{c,n}\}_{n\in\mathcal{N}}$. $R^{\min}_j$ is the minimum transmission rate requirement of the jth PU. Constraint C2 indicates the minimum transmission rate requirements of the kth SU, denoted by R^{thr}_k. It should be noted that the services provided for SUs are multi-modal. Therefore, the requirement for the transmission rate R^{thr}_k is dynamically changed over time. Constraints C3 and C4 indicate that a sub-band can only be occupied by one SU in order to avoid the interference among SUs. Constraint C5 means that the direction of C-UAV is within the interval $[-\pi, \pi]$. The problem given by eq. (14) is highly non-convex and difficult to solve. In this paper, in order to tackle this problem, an intelligent spectrum allocation and trajectory optimization scheme is proposed by exploiting our presented PODMAI framework.

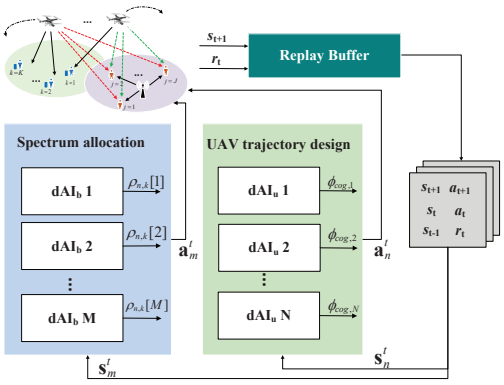


Fig. 3: Our proposed intelligent spectrum allocation and trajectory optimization scheme based on our presented PODMAI framework.

B. Intelligent Spectrum Allocation and Trajectory Optimization in Spectrum Sharing UAV Networks

Since the overall observations of the spectrum sharing UAV network is difficult to be obtained in practice, the formulated optimization problem will be modelled as a multi-agent partially observable problem. Specifically, the spectrum sharing UAV network is regarded as the environment. All the sub-bands and the C-UAVs are regarded as multi-agents. The interaction system between multi-agents is shown in Fig. 3.

State: The optimization objective is achieved by collaborative gaming in the multi-agent system. The state observed by the nth C-UAV agent can be defined as $\mathbf{s}^t_n = \mathbf{q}^{t-1}_{c,n}$. Specifically, the state of the mth sub-band agent at the current time $\mathbf{s}^t_m$ includes the sub-band allocation at the previous time $\mathbf{a}^{t-1}_m$, the location of the C-UAVs at the previous time $\mathbf{q}^{t-1}_c$, the channel information $\mathbf{H}^{t-1}_m$ for the SUs, and the transmission rate in the mth sub-band $\mathbf{R}^{t-1}_m$. Specifically, the state $\mathbf{s}^t_m$ of the mth agent at time slot t is defined as

$$\mathbf{s}^t_m = \{\mathbf{a}^{t-1}_m, \mathbf{H}^{t-1}_m, \mathbf{R}^{t-1}_m, \mathbf{q}^{t-1}_c\}, \tag{15}$$

where $\mathbf{a}^{t-1}_m$ represents the sub-band allocation policy of the mth agent at time step $t-1$. $\mathbf{H}^{t-1}_m = \{h^{\text{los}}_{k,n}, h^g_k\}_{\forall n\in\mathcal{N},\forall k\in\mathcal{K}}$, where $h^{\text{los}}_{k,n}$ and h^g_k denote that the channel gain from the nth C-UAV to the kth SU and that from the PBS to the kth SU at time step $t-1$, respectively. $\mathbf{R}^{t-1}_m = \{R^{t-1}_{k,n,m}\}_{\forall n\in\mathcal{N},k\in\mathcal{K}}$, where $R^{t-1}_{k,n,m}$ is the transmission rate from the nth C-UAV to the kth SU in the mth sub-band.

Action: The action space can be divided into two parts. One part is for the sub-band allocation $\rho_{k,n}[m]$ between the C-UAVs and SUs, and the other part is for the C-UAV flying direction $\phi_{c,n}$. Therefore, the action of the kth sub-band agent at time step t can be defined as

$$\mathbf{a}^t_m = \{\rho_{k,n}[m]\}_{k\in\mathcal{K},n\in\mathcal{N}}, \tag{16}$$

where $\rho_{k,n}[m] = \{0,1\}$ indicates whether the sub-band m is occupied by the communication between the kth SU and the nth C-UAV. The action space dimension is greatly reduced

from 2^M to $N\times K+1$ compared to the use of the SUs as multi-agents. The action of the nth C-UAV agent is $a_{c,n} = \phi_{c,n}$.

Reward: According to the optimization problem formulated in eq. (14), the goal of the reward function for UAV agents is to provide the SUs with high transmission rate requirements without being overly far away from the SUs with low transmission rate requirements. Therefore, the reward function for the nth C-UAV agent is formulated as

$$r_n = -w_1 \sum_{k\in\mathcal{K}_H} d_{k,n} - w_2 \sum_{k\in\mathcal{K}_L} \varsigma_{k,n}, \tag{17}$$

where $\mathcal{K}_H$ and $\mathcal{K}_\mathcal{L} = \mathcal{K}\backslash\mathcal{K}_H$ are the sets of the SUs with high transmission rate requirements and low transmission rate requirements, respectively. $d_{k,n}$ is the distance between the kth SU and the nth C-UAV. $\varsigma_{k,n} \geq 0$ is the penalty item when the distance between the C-UAV and low rate requirements users exceeds the tolerated threshold d_{thr}, which is given as

$$\varsigma_{k,n} = \begin{cases} d_{k,n}, & d_{k,n} > d_{\text{thr}} \\ 0, & 0 \leq d_{k,n} < d_{\text{thr}}. \end{cases} \tag{18}$$

The reward function of the sub-band agents consists of three parts, namely, the sum transmission rate of the secondary system, the interference to the primary system, and the SUs with different transmission rate requirements. Specifically, the reward function is formulated as

$$\begin{aligned} r_m = \alpha_1 \sum_{n=1}^{N}\sum_{k=1}^{K} R^s_{k,n,m} + \alpha_2 \sum_{k\in\mathcal{K}_H} \delta^H_k \\ + \alpha_3 \sum_{k\in\mathcal{K}_L} \delta^L_k + \alpha_4 \sum_{j\in\mathcal{J}} \delta^P_j, \end{aligned} \tag{19}$$

where α_1, α_2, α_3, and α_4 are non-negative constant coefficients. δ^P_j, δ^H_j, and δ^L_j are the penalty items when the transmission rate requirements of the PUs, the high rate requirement SUs and the low rate requirement SUs are not satisfied, given as

$$\delta^P_j = \begin{cases} 0, & R^p_j > R^{\min}_j \\ R^p_j - R^{\min}_j, & 0 \leq R^p_j < R^{\min}_j, \end{cases} \tag{20a}$$

$$\delta^H_k = \begin{cases} 0, & R^s_k > R^H_{\text{thr}} \\ R^s_k - R^H_{\text{thr}}, & 0 \leq R^s_k < R^H_{\text{thr}}, \end{cases} \tag{20b}$$

$$\delta^L_k = \begin{cases} 0, & R^s_k > R^L_{\text{thr}} \\ R^s_k - R^L_{\text{thr}}, & 0 \leq R^s_k < R^L_{\text{thr}}, \end{cases} \tag{20c}$$

where R^H_{thr} and R^L_{thr} are the thresholds for SUs with high transmission rate requirements and low transmission rate requirements, respectively. In order to satisfy all users' dynamic requirements, the penalty term for SUs with low transmission rate requirements is higher than that of high rate requirements, which can avoid the agent from allocating excessive channel resources to users with high requirements to obtain higher sum transmission rate.

Compared with the single agent based method, the dimension of the action space in our proposed framework is reduced significantly. In our considered system with M sub-bands, K SUs, and J UAVs, the action space is $(K \times N)^M$ in the single agent system. However, M sub-bands agents are configured with distributed deep active inference networks in our proposed framework. In this way, the action space can be reduced from $(K \times N)^M$ to $K \times N \times M$. Moreover, collaborative games among multi-agents can compensate for the incomplete observation states of the single agent so as to improve the sum transmission rate.

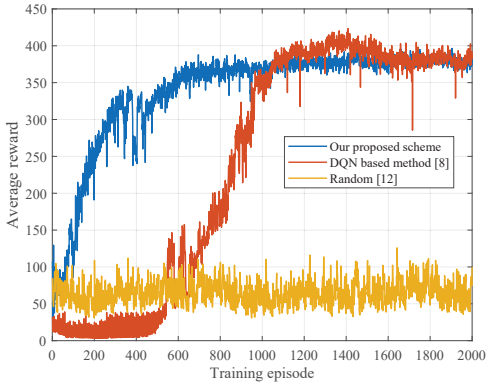


Fig. 4: The sub-band agent convergence performance comparison of our proposed scheme with the benchmark schemes.

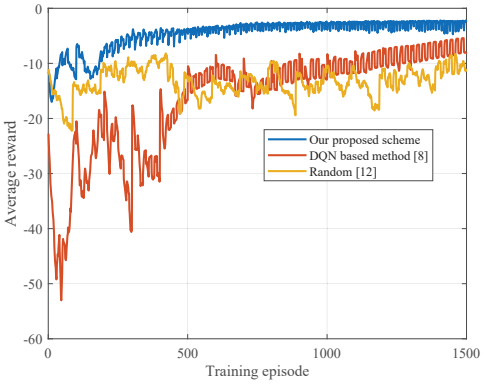


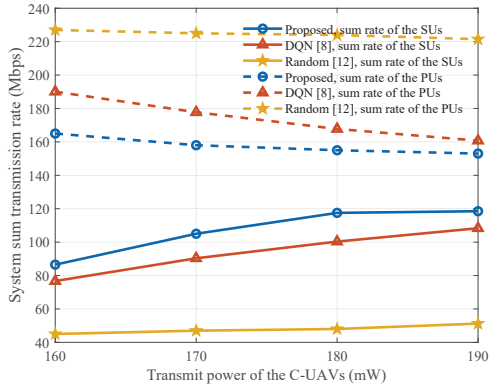
Fig. 5: The C-UAV agent convergence comparison of our proposed scheme with the benchmark schemes.

IV. Simulation Results

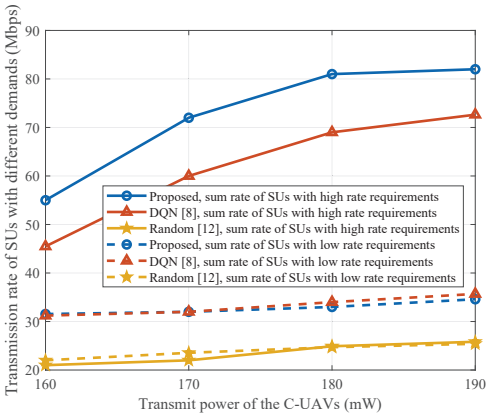
In this section, simulation results are presented to evaluate the performance of our proposed resource allocation scheme. The simulation settings are based on the work in [12]. In the simulation, $\beta_{\text{ref}} = 1 \times 10^{-3}$ represents the channel power gain at the reference distance of 1 meter. A three-dimensional coordinate system is established. SUs and PUs are located randomly in two 500×500 cells, respectively. The UAVs fly at a constant altitude of 100 m. The background noise power is -169 dBm.

Two benchmark schemes are considered for the performance comparison. **DQN**: DQN is used for spectrum allocation policy and UAV trajectory optimization [8], **Random**: spectrum allocation and UAV trajectory are randomly generated [12].

As shown in Fig. 4, the convergence performance of our proposed scheme is presented with respect to the training episodes. It can be seen that the reward of our proposed



(a)



(b)

Fig. 6: System sum transmission rate versus the C-UAV transmit power under different schemes (a) The transmission rate of the primary system and the secondary system. (b) The transmission rate of SUs with different service requirements.

scheme fluctuates in the first 600 episodes and tends to be stable after 1000 episodes. Moreover, compared with the DQN based method, the proposed framework achieves a faster convergence speed by about 400 training episodes due to its capability in performing multiple updates simultaneously. Moreover, it is seen that the proposed PODMAI has a larger reward than the random spectrum allocation and UAV trajectory method. Therefore, it demonstrates the efficiency of our proposed scheme in terms of the achievable rewards.

The comparison of the convergence performance for C-UAV agents is shown in Fig. 5. It can be seen that our proposed scheme converges 600 episodes faster than the DQN based method. It is demonstrated that the belief based learning in our proposed scheme is more practical than the value based learning in DQN. In fact, more efficient exploration is achieved by the proposed one in the early training episodes.

Fig. 6 (a) shows the system sum transmission rate with different transmit power levels. It can be seen that 10 Mbps and 70 Mbps gains of the transmission rate are achieved compared with DQN and random based methods. Moreover, the interference caused to the PUs exceeds the threshold when the transmit power is 180 mW. Therefore, the sum transmission rate gain in the secondary network for these scheme is limited.

The transmission rate of the SUs with different transmission rate requirements is shown in Fig. 6 (b). It can be seen that our proposed scheme achieves about 10 Mbps and 58 Mbps transmission rate gains than the DQN and random based methods, respectively when the transmit power is 180 mW. Moreover, the transmission rate of the SUs with high rate requirements increases significantly with the UAV transmit power while ensuring the performance of the SUs with low requirements. This is because the increasing transmission rate of high requirement users results in a higher instant reward compared to that of the low requirement users.

V. Conclusion

In this paper, a novel PODMAI framework was proposed for designing efficient resource allocation in wireless communications. Moreover, in order to verify the efficiency of our proposed framework, an intelligent spectrum allocation and trajectory optimization scheme was presented by using our proposed framework in a spectrum sharing UAV network. Simulation results demonstrated that our proposed framework obtains the highest sum transmission rate of the secondary system among all the considered benchmark schemes. Moreover, the convergence speed was improved significantly compared with the DRL based benchmark schemes.

References

[1] B. Agarwal, M. Togou, M. Marco, and G. Muntean, "A comprehensive survey on radio resource management in 5G hetnets: current solutions, future trends and open issues," *IEEE Commun. Surv. Tut.*, vol. 24, no. 4, pp. 2495-2534, 2022.

[2] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems" *IEEE Access*, vol. 8, pp. 133995-134030, 2020.

[3] W. Qi, Q. Song, L. Guo and A. Jamalipour, "Energy-efficient resource allocation for UAV-assisted vehicular networks with spectrum sharing" *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7691-7702, Jul. 2022.

[4] Y. Wang, L. Chen, Y. Zhou, X. Liu, F. Zhou, and N. Al-Dhahir, "Resource allocation and trajectory design in UAV-assisted jamming wideband cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, pp. 635-647, Jun. 2021.

[5] Z. Shi, X. Xie, H. Lu, H. Yang, M. Kadoch, and M. Cheriet, "Deep reinforcement learning based spectrum resource management for industrial internet of things," *IEEE Internet Things J.*, vol. 8, no. 5, Mar. 2021.

[6] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, "Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications," *IEEE Trans. Veh. Technol.*, vol. 71, pp. 8810-8824, Aug. 2022.

[7] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surv. Tut.*, vol. 21, no. 4, pp. 3133-3174, 2019.

[8] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163-3173, 2019.

[9] H. Wang, H. Zhang, X. Liu, K. Long and A. Nallanathan,"Joint UAV placement optimization, resource allocation, and computation offloading for THz band: a DRL approach," *IEEE Trans. Wireless Commun.*, 2022.

[10] A. Krayani, A. Alam, L. Marcenaro, A. Nallanathan, and C. Regazzoni, "A novel resource allocation for anti-jamming in cognitive-UAVs: An active inference approach," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2272-2276, Oct. 2022.

[11] C. Fan, C. She, H. Zhang, B. Li, C. Zhao, and D. Niyato, "Learning to optimize user association and spectrum allocation with partial observation in mmWave-enabled UAV networks," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 8, Aug. 2022.

[12] W. Wu, F. Yang, F. Zhou, Q. Wu and R. Q. Hu, "Intelligent resource allocation for IRS-enhanced OFDM communication systems: A hybrid deep reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, 2022.

[13] X. Zhu, Y. Huang, Q. Wu, F. Zhou, X. Ge, and Y. Liu, "Dynamic channel selection and transmission scheduling for cognitive radio networks," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 24429-24443, Dec. 2022.