Spectrum Allocation for Covert Communications in Cellular-Enabled UAV Networks: A Deep Reinforcement Learning Approach

Xinzhe Pi[1,*], Bin Yang[2,3]
[1]School of Systems Information Science, Future University Hakodate, Hakodate, Hokkaido, 041-8655, Japan
[2]School of Computer and Information Engineering, Chuzhou University, Chuzhou, 239000, China
[3]MOSAIC Lab (www.mosaic-lab.org), Espoo 02150, Finland
[*]Corresponding author

This paper investigates the covert communications via spectrum allocations in a cellular-enabled unmanned aerial vehicle (UAV) network consisting of a base station (BS), UAVs, ground users (GUs), and a warden, where warden attempts to detect the transmission from a target GU to a UAV receiver. We formulate the spectrum allocation as an optimization problem with the constraints of covertness performance requirement and the qualities of service (QoS) of cellular communications. This is a nonlinear and nonconvex problem, which is generally challenging to be solved. Thus, we propose a deep reinforcement learning (DRL) approach to solve it. Under such an approach, we first model the multi-agent DRL environment in such networks. Then we define the state, action, reward and interaction mechanism of the DRL environment. Finally, a DRL algorithm is presented for learning the optimal policy of spectrum allocation.

***Index Terms*—Covert communication, cellular-enabled UAV network, spectrum allocation, deep reinforcement learning, multi-agent reinforcement learning.**

I. Introduction

Cellular-enabled UAV network, which enables UAVs to resue spectrum resource of cellular networks for communications, has been considered as an effective way to improve the UAV network performance [1]. UAVs have appealing features of high mobility, low cost and flexible deployment and are widely used in many scenarios, such as smart traffic control, surveillance, delivery services and so on. To achieve variable and complicated application scenes, UAVs need to interact with other devices by radio, and traditional UAV communication is point-to-point communication with an unlicensed spectrum. To overcome this defect, we can use the cellular spectrum resource to realize long-distance communications in the UAV network with the assistant of cellular networks. Cellular-enabled UAV networks have been envisioned as a critical component in the sixth generation (6G) wireless networks.

Unfortunately, the wireless channel characteristics of broadcast and openness pose unprecedented security and privacy threats on transmitting sensitive information, especially financial and military data in the presence of adversaries. To protect the information transmission security, the most commonly used security methods rely on upper-layer cryptographic techniques, which require high computational complexity and may not be suitable for cellular-enabled UAV networks due to large energy consumption. Meanwhile, these techniques may also be infeasible with the appearance of powerful computing devices. Covert communication, which is a promising technique to hide the existence of wireless transmission, can provide strong security protection for cellular-enabled UAV networks.

Manuscript received December 16, 2021; revised August 23, 2022. Corresponding author: Xinzhe Pi (email:pxinzhe@163.com).

Different from encrypted communication which protects the data of communications, the purpose of covert communication is to hide the communication signal. The adversaries need to figure out whether the communication occurs, other than crack the encrypted information. After the bench-marking work [2] proposed a square root law with Gaussian noise channels for covert communication, extensive works have been dedicated to the studies of the covert performances in various scenarios. Some works focus on the performance of single-hop covert communication such as [3] and [4]. On the other hand, researchers analyze the covertness and covert rate of multi-hop covert communications in [5] and [6]. Meanwhile, some other works pay attention to the covert performance under UAV assisted networks such as [7], [8] and [9]. The works above put forward some schemes for covert performance improvement like jamming and relaying, but they all focus on simple scenarios with a pair of source and destination nodes with/without relay nodes.

We notice that all the works above take traditional methods in performance analysis, such as probability theory, mathematical statistics and mathematical optimization methods. However, these methods can not be applied to the dynamic and changing environment. Besides, the derivations of such methods are very complex in scenarios with a large number of communicators and links. Remarkably, the capability of machine learning (ML) to learn from training data and unveil hidden patterns has driven the recent trend of using ML for UAV networks, especially the DRL methods. In fact, the DRL algorithms have been widely used for communication and networking analysis, especially for resource allocation problems [10]. Much research studies the performance enhancement problems by finding the optimal resource allocation scheme in the UAV network, such as [11], [12] and [13].

Although we can improve the UAV network performance by

reusing the spectrum resource of cellular networks, spectrum reuse can cause co-channel interference (CCI), which deteriorates the original communication quality in cellular networks. Besides, interference and noise also have a great influence on the performance of covert communication. Therefore, the spectrum allocation is of great importance to enhance the covert performance of UAV networks. To address this issue, we design a multi-agent DRL framework for the spectrum allocation optimization problem. The main contributions of this paper are summarized as follows:

- In this paper, we consider a cellular-enabled UAV network consisting of a base station (BS), UAVs, GUs, and a warden. UAVs and GUs are communicating with BS in cellular links, while a GU is sending covert messages to a UAV in the covert link. Due to the scarcity of spectrum resources, different links reuse the same spectrum resource blocks (RBs) for communications. Our goal is to optimize the spectrum allocation for maximizing the covert rate subject to the covert requirement and the QoS of cellular links. Therefore we further design a multi-agent DRL model to address this problem.
- Based on the model, we propose a multi-agent DRL algorithm to learn the optimal spectrum allocation. *To my best knowledge, this is the first work using multi-agent DRL algorithms to solve the spectrum allocation problem in such network scenarios.* Compared to traditional optimization methods like mathematical analysis, DRL methods can be applied to such a complicated network and learn the hidden patterns in the environment.
- We program and implement our proposed DRL framework based on PyTorch. Simulation results show the impact of some key system parameters on the covert rate performance of such network.

The rest of this paper is arranged as follows. Section II introduces some related works about covert communication performance analysis in detail. In Section III, we present what covert communication is, show the network scenario and formulate our optimization problem. In section IV, the multi-agent DRL framework is proposed to solve the spectrum allocation problem. Section V presents the simulation results and corresponding analysis. At last, Section VI gives the conclusion of this paper.

II. Related Works

Past related works about covert communication performance analysis can be divided into two categories: one-hop covert communication and multi-hop covert communication. The former means the covert signal is transmitted directly from the transmitter to the receiver, while the latter means the covert signal passes through other nodes (e.g., relays) before it reaches the receiver. As for one-hop situations, Jiang et al. [3] studied the covert performance enhancement under the device-to-device (D2D) underlaying cellular network. They used nonorthogonal multiple access (NOMA), an emerging technique used for throughput optimization problems such as [14] and [15], to improve covert throughput. In [4], Zhou et al. applied successive convex approximation techniques to solve the UAV trajectory and transmission power optimization problem. They proved that the method performed well in the problem and improved the covert performance between the UAV and the receiver.

Regarding the multi-hop scenarios, Wang et al. [5] designed a covert link with multiple relay nodes, which supporting long-distance covert communications. The physical layer security (PLS) was also considered combining with covert performance. They derived the closed-form expressions of the optimal transmit power, transmission rates and secrecy rates with a fixed number of hops. In [6], Azadeh Sheikholeslami et al. studied the covert performance and PLS in the middle-scale network. The covert rate and delay time was contrasted with a different number of relays. They also made performance comparisons between a shared key and independent keys used in relays.

Due to the extensive application of UAVs, scenarios of covert communication under UAV networks are becoming increasingly popular, like [4]. Yan et al. [7] researched the joint optimization problem of transmission power and UAV height when a UAV is sending the covert signal to a ground user. In [8], Yan et al. also proposed two heuristic approaches analyzing the covert performance under the same scenario. In that paper, they subdivided the scenario into six different situations and made a further analysis and comparison. In [9], the UAV transmitter is sending confidential data to multiple ground users and trying to hide the communication with the location uncertainty of warden. Jiang et al. proposed a block coordinate descent (BCD) method based iterative algorithm to optimize the UAV trajectory and transmission power for covert performance improvement.

Notice that none of the works above takes ML methods, we investigate and find there are seldom few papers to analyze the performance of covert communication using ML. Liao et al. [16] research a power allocation problem for a cooperative cognitive covert communication system, where the relay secondary transmitter covertly sends private information under the supervision of the primary transmitter. The secondary transmitter uses the forwarding signal sent to the primary receiver to hide the covert signal. Authors use a deep learning (DL) method called generative adversarial network (GAN) [17] to find the optimal allocation scheme for covert performance optimization. Under the proposed framework, the generator adaptively generates the power allocation solution for covert communication, while the discriminator judges whether to transmit the covert signal or not. In [18], Kim et al. consider a clever eavesdropper using a DL classifier to help the detection. They show that signals with different modulation types can effectively hide the covert messages, but they do not use ML methods for the covert performance enhancement at Alice.

Compared to all works mentioned above, we propose a DRL-based framework for the spectrum allocation optimization problem in the cellular-enabled UAV network, which is a novel idea that has never been considered in any past work. The DRL framework will be illustrated in Section IV.

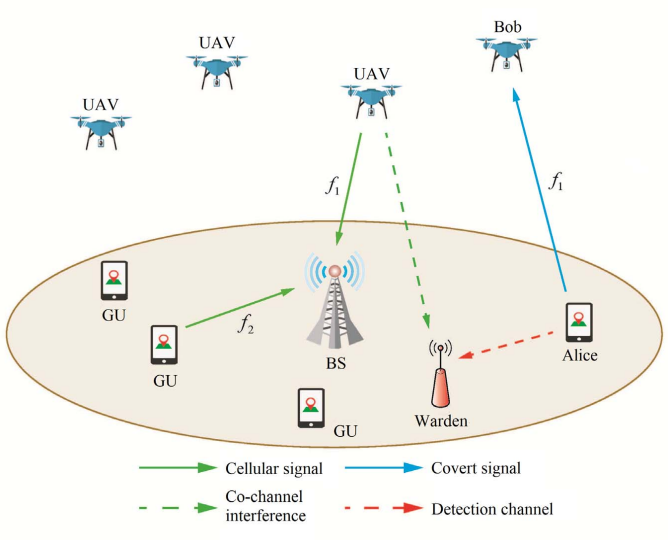

Fig. 1: Network model.

III. System Model

In this section, we firstly introduce the network model and channel model in our work. Then the detection mechanism at warden is illustrated. Finally, we give the formulation of the optimization problem in this paper.

A. Network Model

As illustrated in Fig. 1, we consider a cellular-enabled UAV network consisting of a warden, M UAVs and N GUs within the coverage of BS. More specifically, each GU is a point uniformly and randomly distributed in the circle with the BS as the center and a fixed radius on the ground, and each UAV is a random point uniformly distributed in the intersection space of the airspace within a certain altitude range and the sphere with the BS as the center and a fixed radius. The value of radius and altitude range are listed in Table I. There is a fixed GU (Alice) sending covert signals to a UAV (Bob), while warden tries to detect the transmission from Alice to Bob. Alice can only have one covert link and Bob may change in each time step. Considering a limited number K of orthogonal spectrum RBs, each UAV/GU can reuse an RB to communicate with BS in cellular links, and the covert link also reuses one RB.

Due to the frequency reuse, the co-channel interference (CCI) will be generated between the links using the same RB simultaneously. It will result in the performance degradation of wireless communications. But to some extent, the interference is conducive to enhancing covert communication performance. In Fig. 1, one UAV uses the same RB of frequency f_1 as Alice, thus generating CCI to the warden (also to Bob). And the GU using another RB of frequency f_2 does not generate CCI to warden. A fundamental issue is how to allocate these RBs for achieving the optimal covert performance (e.g., covert rate) in cellular-enabled UAV networks. Therefore, we plan to formulate it as a nonlinear and nonconvex optimization problem, which is challenging to be solved. We will propose a novel theoretical framework to solve this challenging optimization problem based on a DRL algorithm.

B. Channel Model

We consider all channels of cellular and covert links as Rayleigh fading channels. The channel gain between nodes i and j is denoted as $g_{i,j}$, here nodes include the BS, the warden, UAVs and GUs. Similarly, $d_{i,j}$ denotes the distance between nodes i and j. Then the pass loss is expressed as $d_{i,j}^{-\alpha}$, where α is the pass-loss exponent. Thus, the channel gain is an random variable subject to exponential distribution with mean $d_{i,j}^{-\alpha}$. That is, $g_{i,j} \sim E(d_{i,j}^{-\alpha}), \forall i,j \in U, i \neq j$, where U denotes the set of all nodes and α is usually set to 4 in typical urban macrocell environment.

C. Detection at warden

Private and critical information travels in the communication network every second. To keep our privacy secret, encrypted communication is used in most protocols. The plaintext is converted to ciphertext by a password, which means the eavesdropper can not get the content directly. However, sometimes the eavesdropper just cares about whether the communication occurs like military cases, and encrypted communication can not avoid it. To address this problem, the emerging concept of covert communication is proposed. Assuming that Alice sends messages to Bob, and warden is detecting from the radio environment. warden judges whether Alice is transmitting or not based on the radio power received.

For warden, the detection process is hypothesis testing. H_0 denotes that Alice is transmitting covert signals, and H_1 denotes the opposite. The judgment is made by whether the received power P_W is smaller than a threshold τ or not. That is, $P_W \underset{H_0}{\overset{H_1}{\gtrless}} \tau$. warden makes mistakes when he gives wrong judgments, and we use P_{FA} denotes the probability of rejecting H_0 when H_0 is true, and P_{MD} denotes the probability of rejecting H_1 when H_1 is true. Then warden's error probability is $P_{err} = P_{FA} + P_{MD}$. For Alice, her communication is covert if she can make $P_{err} > 1 - \epsilon$ for any $\epsilon > 0$. This is the definition of covertness performance. And covert rate means the data rate of covert communication.

D. Problem Formulation

In this paper, we aim to solve the spectrum allocation optimization problem in the cellular-enabled UAV network. The objective of the spectrum allocation is to maximize the covert rate of covert communication while obeying the covertness performance requirement and guaranteeing the QoS of cellular communications. We use $a_{i,j}$ denoting the index of RB used in the link between node i and j. $a_{i,j} = 0$ if the link is not active. Notice that we allocate RBs to newly created cellular and covert links only, we use $L = \{(i,j)|i \text{ and } j \text{ are both nodes of each new link}\}$ to represent the set of new links in a time step. Therefore the spectrum allocation can be defined as $\mathbf{a} = \{a_{i,j}|(i,j) \in L\}$. Generally, we set the minimal signal to interference plus noise ratio (SINR) threshold as the lowest QoS requirement. For each link, SINR $\xi = \frac{S}{I+N}$, where S is the signal power, I is the interference of the link and N is the noise power (in other

parts of this paper, N denotes the number of GUs). Then the problem is formulated as follows:

$$\begin{aligned} & \arg\max_{\mathbf{a}^* \in \mathcal{A}} C_{co} \\ s.t. \quad & P_{err} > 1-\epsilon, \qquad \forall \epsilon > 0 \\ & \xi_{ce} \geq \xi_{min}, \qquad \text{for each cellular link} \end{aligned}$$

In the above formulas, $C_{co} = W \log(1+\xi_{co})$ denotes the covert rate and W denotes the bandwidth of an RB. $\mathcal{A}$ is the set of all possible spectrum allocations and $\mathbf{a}^*$ is the optimal spectrum allocation of $\mathcal{A}$ with the maximal C_{co}. We use ξ denoting the SINR of a link, thus ξ_{co} and ξ_{ce} denotes the SINR of the covert link and cellular link respectively. ξ_{min} means the minimal SINR threshold for cellular links. As illustrated before, the covertness performance requirement can be presented as $P_{err} > 1-\epsilon$ for any $\epsilon > 0$. In our work, P_{err} is the rate of the number of detection errors and the total number of detections at warden, which are measured by simulation experiments using Monte Carlo methods. With the model learning, P_{err} will meet the covertness requirement.

For the link between nodes i and j, the SINR of the link is calculated by $\xi_{i,j} = \frac{P_i g_{i,j}}{I_{i,j}+\sigma_j^2}$ where P_i is the transmission power of node i, $g_{i,j}$ is the channel gain of current link, $I_{i,j}$ is the interference received by node j and σ_j^2 is a constant representing the received noise power at node j. If the link is a cellular link, $P_i = P_{ce}$, and $P_i = P_{co}$ for a covert link. The interference is the sum of received power at node j of all other signals using the same RB as node i.

IV. Multi-Agent Deep Reinforcement Learning Based Spectrum Allocation Framework

In this section, we list and compare some popular algorithms in DRL and multi-agent reinforcement learning first. Then we give the multi-agent environment model under the network scenario. The proposed DRL framework based on the model is illustrated at last.

A. Deep Reinforcement Learning

Reinforcement learning (RL) is an emerging method in artificial intelligence (AI). The agent (learning entity) will explore and understand what the best action is by taking possible actions and getting the corresponding reward from the environment. Nevertheless, as for problems with high dimensions of state space and action space, the computational complexity of RL becomes unacceptably huge. It's often called "the Curse of Dimensionality". Therefore deep reinforcement learning (DRL) is proposed to dispel the curse. In RL, the agent selects action by comparing the recorded values of all actions under the given state, while the amount of records is huge with high dimension problems. Thus the neuron network (NN) is introduced to replace the value table of actions and states in RL due to its strong presentation ability. In this way, we can just input the state into NNs and get the values of actions, thus avoiding the problems caused by high dimensional space.

Deep Q-Network (DQN) is the first DRL method that integrates DL and RL into the novel DRL framework. It performs well in Atari games with large discrete state space and limited discrete action space. Later, deep deterministic policy gradient (DDPG) is declared to deal with the continuous action space. Notice that although there are many related works, no DRL method has a good and steady performance with high dimensional discrete action space, which is a problem in our work. Therefore we can not simply apply a single-agent DRL algorithm in this paper.

B. Multi-agent Reinforcement Learning

In traditional RL algorithms, there is usually a single agent interacting with the environment. We assume the environment is static, that is, the state transition probabilities with actions are fixed. Thus the optimal policy for the agent will not change while learning, which ensures the convergence and effectiveness of RL algorithms.

However, we often face cooperative or competitive tasks with one or more participants, such as prisoner's dilemma. In such tasks, the maximization of individual interests and collective interests are often contradictory. Besides, one agent's action may change the state of the environment, resulting in a non-static environment for another agent. To solve these problems, multi-agent reinforcement learning (MARL) is proposed. MARL is suitable for the complicated tasks stated above, aiming to make each agent learn with the consideration of other agents' behaviors. By this method, the algorithm can reach the maximal collective interests, not personal ones.

QMIX is a MARL algorithm based on value approximation [19]. In QMIX, each agent use DQN for learning, while there is a mixing network approximating the Q value of all agents' observations and their actions. Thus the algorithm can learn a good policy for all agents. Besides, multi-agent deep deterministic policy gradient (MADDPG) is another famous MARL algorithm that relying on the policy gradient [20]. The algorithm applies the deterministic policy and demonstrates that the multi-agent environment is stable when the actions of all agents are determined. The performance is proved by simulations.

C. Modeling of Multi-Agent Environment

Since the action space of spectrum allocation is discrete, MADDPG does not perform well in our scenario. Gumbel softmax is a choice, but it is not always effective. As for QMIX, it needs Q-values of all agents in each time step and approximates the Q-value of global state and actions. Because the cellular link may not transmit information and thus need no RB, one cellular agent may not choose action and output Q-value, thus makes QMIX unavailable.

Therefore, we propose a multi-agent DRL framework for our problem. In our scenario, each cellular link and the covert link play the role of an agent. We assume that each UAV and GU has only one cellular link communicating with BS, and only the target GU has a covert link communicating with a random UAV in different time steps. Cellular links transmit messages with a preconfigured probability, and the covert link transmits data all the time. Each agent has an evaluation net and a target net for learning, as shown in Fig. 2. In each time step, all agents observe the current state of the

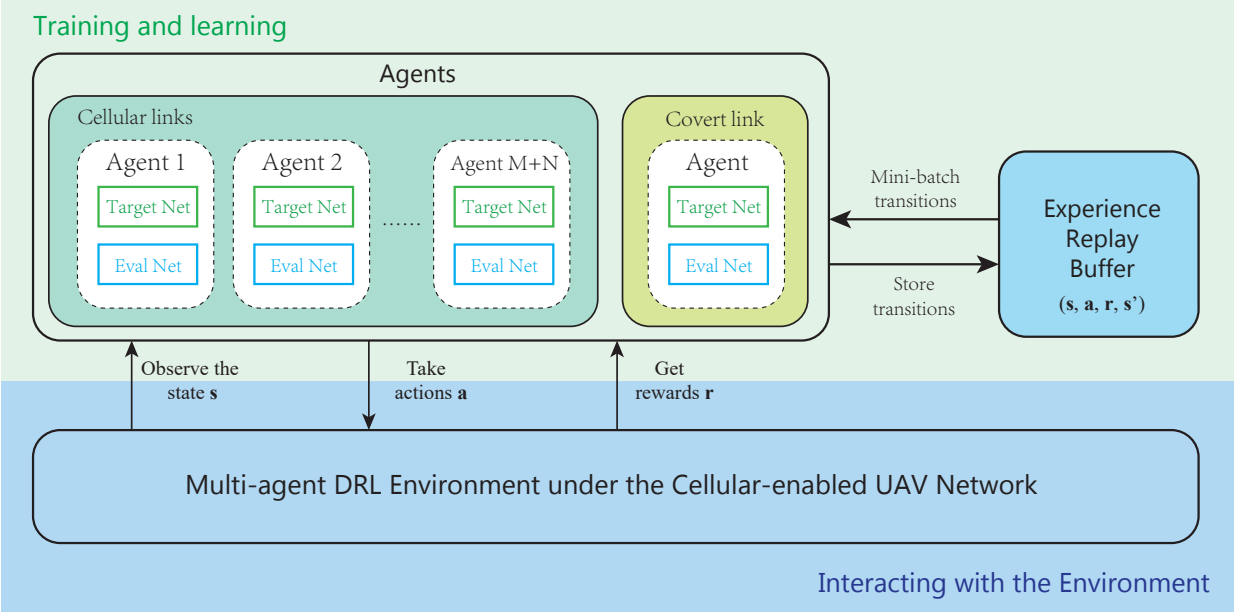

Fig. 2: Multi-agent DRL framework.

environment and take their actions (i.e., spectrum allocation) independently. By interacting with the environment, each agent gets the reward reflecting the value of action and state.

To learn the optimal policy on the system level, we take the used RB and interference of all agents in the last time step as a part of the NN input. Besides, we will give the average reward of all agents to each one after they take actions. Then the transitions of the states, actions, rewards and the following states are stored in the experience replay buffer. After the buffer is full, agents start the learning process in each time step based on the mini-batch of historical transitions. When the loss value of a DQN converges or the maximum number of episodes is reached, the learning process of the agent is terminated.

The definitions of the state, the action and the reward in the DRL framework are given as follows:

State: In each time step, the state of the environment consists of the global state information and the local state information. The global information includes the RB used and the interference received for each agent in a single time step, and the positions of all UAVs and GUs. Since we assume all UAVs and GUs are static, the position of all nodes are fixed during simulations. Location information is known to all nodes before simulations. The information of the RB used and interference received in last time step is known to BS and BS can send it to all nodes before the beginning of a time step. Notice that agents do not know where warden is. The local information for each agent contains the indexes, the equivalent of node ID, of the transmitter and the receiver in a time step. The local information of a node is assumed to be known only to the node itself.

Action: The action of an agent in a time step is the behavior of choosing an RB for its communication to the responding receiver. Since RBs are less than the number of links, spectrum reusing is inevitable. The actions of all agents constitute the spectrum allocation of the algorithm in a certain time slot. The objective of the framework is to give the optimal allocation of all agents to achieve the maximum covert rate with constraints.

Reward: According to the actions (i.e., spectrum allocation) given, we allocate the RBs to all links and then make a certain number of simulations (e.g., a thousand times). For each cellular link, we calculate the SINR in each simulation. The reward of agent is given based on the following rules:

1. If the covert communication is detected by warden, the agent receives a fixed minus reward (e.g., -10).

2. If the covert communication is not detected, but the SINR of the cellular link does not satisfy the threshold requirement, the agent receives a smaller fixed minus reward (e.g., -1).

3. If the covert communication is not detected and the cellular link satisfies the SINR threshold requirement, the reward is the value of covert rate. A higher covert rate means a higher reward.

For the covert link, we calculate the covert rate in each simulation. The reward of agent is given based on the following rules:

1. If the covert communication is detected by warden, the agent receives a fixed minus reward (e.g., -10).

2. If the covert communication is not detected, the reward is the value of covert rate.

D. Proposed DRL Algorithm

The proposed DRL algorithm is illustrated in Algorithm 1.

E. Structure of Deep Neuron Network

Based on the definitions in IV-C , the structure of the deep neuron network (DNN) for each agent is shown in Fig. 3. We use a four-layer DNN structure which has two hidden layers inside. There are $(M+N)l+K$ neurons in the input layer, where M and N are defined in Section III-A. l is the

information dimension of each agent, which includes the state in the last step $\mathbf{s}_{t-1}$, the action in the last step $\mathbf{a}_{t-1}$ and the state in the current step $\mathbf{s}_t$. The output of the DNN is the spectrum allocation for the current agent, and the size of the output layer is the number of RBs K. The dimension of each hidden layer is the square root of the dimension product of the input layer and output layer, which is $\sqrt{((M+N)l+K)K}$ here. All layers are fully connected.

Algorithm 1: proposed DRL algorithm for spectrum allocation

Input: M, N, K
Output: trained DNN models

Generate the locations of UAVs, GUs and warden randomly and evenly;
For cellular links of all agents, create a trained DQN and a target DQN with weights θ and θ' respectively, initialize θ randomly, let $\theta' = \theta$;
For the covert link of Alice, create a trained DQN and a target DQN with weights θ_A and θ'_A, initialize θ_A randomly, let $\theta'_A = \theta_A$;
Let $p = 1$, $t = 1$, initialize other parameters of the environment and algorithm;
for $p < p_{max}$ **do**
    Reset the environment;
    for $t < t_{max}$ **do**
        Generate link connection requests randomly for all links;
        Define U_t as the set of the covert link and the cellular links taking communication in current time step;
        for *each agent in* U_t **do**
            Get state info $\mathbf{s}_t$;
            Input $\mathbf{s}_{t-1}$, $\mathbf{a}_{t-1}$ and $\mathbf{s}_t$ into the trained DQN;
            Get action $\mathbf{a}_t$;
        end
        Calculate the covert performance;
        Get transferred state after taking actions $\mathbf{s}_{t+1}$;
        for *each agent in* U_t **do**
            Get reward $\mathbf{r}_t$;
            Store transition $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{s}_{t+1})$ into the experience buffer;
            Model learning;
        end
        Update the global state information;
    end
    if *covertness is satisfied and the covert rate difference is small enough* **then**
        break;
    end
end

V. Numerical Results

In order to verify the performance of the proposed framework, we implement the framework based on PyTorch. All UAVs and GUs are within the coverage of BS. The channels of all links are assumed to be Rayleigh fading channels, and their

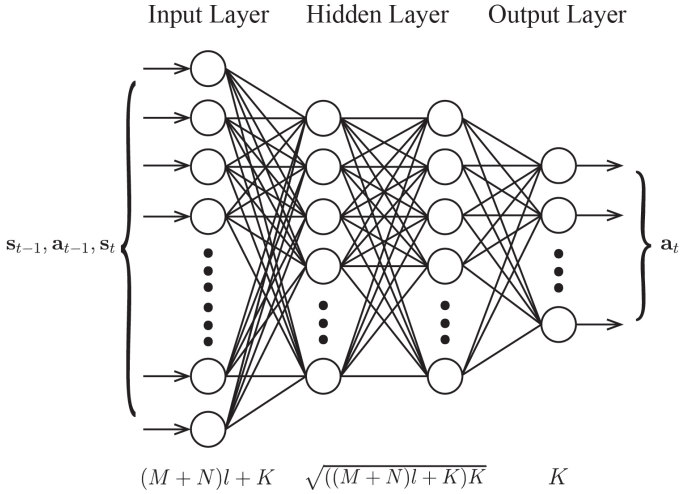

Fig. 3: Structure of the neuron network for each agent.

channel gains are random variables subject to an exponential distribution. All links expire and are released after one time step, thus the occupied RBs are released too. In each time step, the algorithm procedure consists of observing state, choosing actions, interacting with the environment, getting rewards and learning.

As for the detection threshold, warden will try to take the best threshold to make his detection as accurate as possible. Since Alice and Bob do not know the detection mechanisms of warden, we assume that warden knows the value of the optimal threshold all the time, which is the worst situation for Alice. To achieve this assumption, the program computes the received signal power at warden based on the spectrum allocation given by DNNs. Then different thresholds within a valid range will be applied to calculate the corresponding error probabilities. By comparing the probabilities, the optimal threshold with the minimal error probability will be picked and used for warden's detection in the current time step. Only after warden makes his judgment, the framework can calculate the rewards of all agents.

For the DRL algorithm, we configure the exploration probability and the learning rate to 0.1. The discount factor is 0.9. The size of the experience replay buffer for each agent is set to 1000, and the mini-batch size is 32. Notice that in many DL-based works, the learning rate and exploration probability are becoming smaller and smaller with the training process for more effective learning. We have tried this trick in our program, but it seems does not make obvious improvement, thus we do not apply the trick in simulations. As for the communication environment, major parameters are listed in Table I below.

Due to the randomness of channel gains and agents' behaviors, covert rate of a single time step shows an uncertain fluctuation, which makes a difficult numerical analysis. Thus, we analyze the average of maximum covert rate in massive simulations instead after the convergence of the algorithm.

The relation between the maximum covert rate and the number of UAVs M is shown in Fig. 4 where K is fixed and $M = N$. The positions of all roles are immutable during

TABLE I: Communication Parameters

Parameter name	Value
BS coverage radius	300m
UAV height	50∼120m
Carrier frequency	2GHz
RB bandwidth	150KHz
Minimal cellular communication SINR	-5dB
Path-loss exponent	4
Thermal noise density	-174dBm/Hz

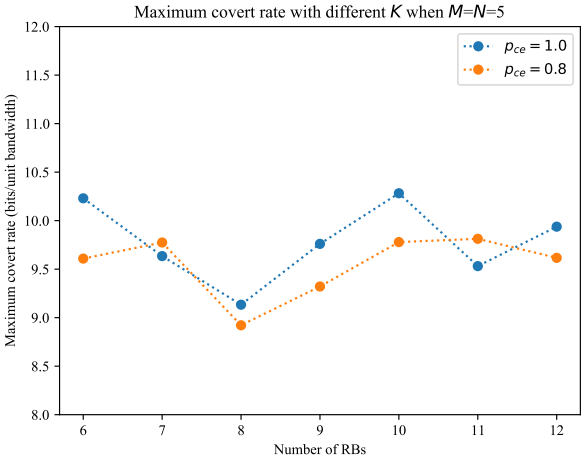

Fig. 4: Maximum covert rate with different M when $M = N, K = 10$.

the learning process. p_{ce} denotes the probability of a cellular link transmitting messages. As illustrated in this figure, we find the covert rate decreases with the increase of M and N when p_{ce} is fixed. This is easy to understand, because with the increase of the number of links, the spectrum resources become more and more scarce, which will increase the CCI of the covert link. According to Shannon formula, SINR decreases as the interference increases, thus the upper limit of covert rate decreases. Besides, we find that covert rate does

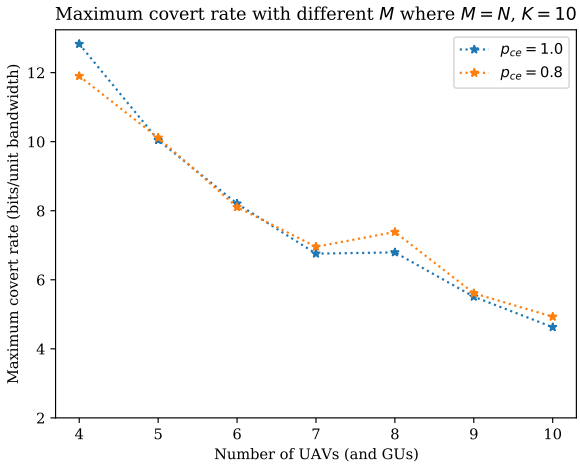

Fig. 5: Maximum covert rate with different K when $M = N = 5$.

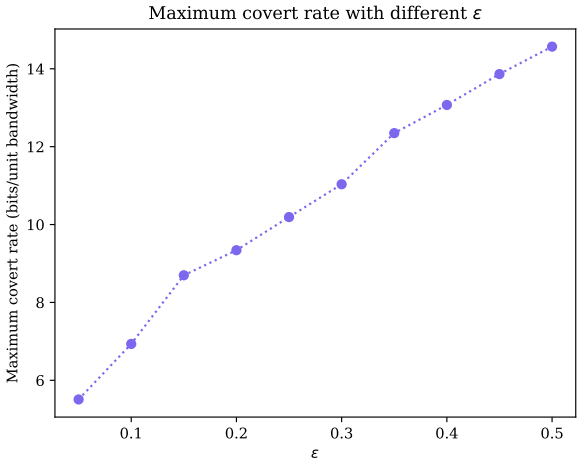

Fig. 6: Maximum covert rate with different ϵ when $M = N = 5$ and $K = 8$.

not show obvious changes and patterns with different values of p_{ce} when M, N and K are the same.

Fig. 5 shows the relation between the maximum covert rate and the number of RBs K when M is fixed and $M = N$. Here we set $M = N = 5$, thus we have 10 cellular links in total and one covert link as the warden's target. We can find that the covert rate shows an irregular fluctuation with the changing parameter K. The deviation of fluctuation is small compared with the absolute value of maximum covert rate. Theoretically, more spectrum resources mean less interference, thus the data rate should increase. We think the reason why the covert rate does not increase is that the interference can not decrease to meet the covertness performance requirement, although the number of RBs increases. Thus the patterns between the covert rate and K is not obvious. Similarly with Fig. 4, the patterns of covert rate and p_{ce} is not clear too in Fig. 5. We will make a further analysis of these patterns in future work.

Fig. 6 shows the value of maximum covert rate with different ϵ after algorithm convergence when $M = N = 5$ and $K = 8$. We can see that maximum covert rate increases while ϵ increases because higher covertness requirement means a lower covert rate. In addition, covert rate rises slower when ϵ becomes bigger.

Loss is an index used to measure whether the model is trained to an acceptable state, and can be used to judge whether the model converges. In Pytorch, we can get the loss of NN model by library functions easily. The mean loss values of an agent corresponding to a cellular link and the agent corresponding to the covert link in all episodes are shown in Fig. 7. Regarding to the cellular agent, the mean loss value does not change obviously in about the first 5 episodes. Then the value decreases with shakes while models learning, and it goes down very quickly at first, then the descent speed becomes slower and slower. In the last 10 episodes, the mean loss value is relatively stable which means the model achieves convergence. The trend of the mean loss value with the model of the covert link is very similar to the model of the cellular

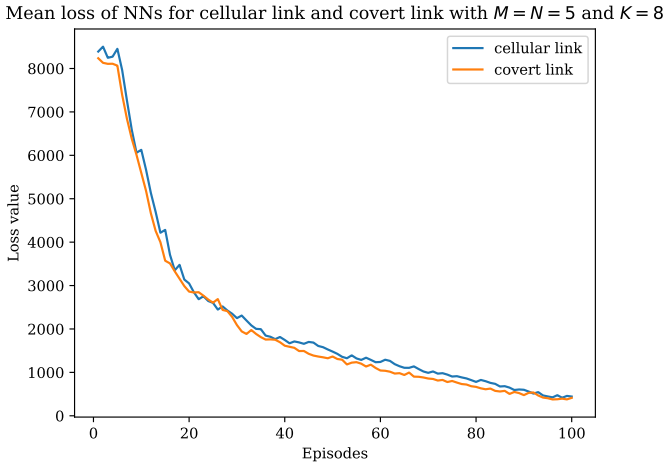

Fig. 7: Mean loss of NNs for cellular link and covert link when $M = N = 5$ and $K = 8$.

link.

Since the spectrum allocation scheme is the objective in this paper, we assume the power of all communicators is fixed. This may be the reason why the framework is not very effective in some cases, because signal power plays an important role in covert communication. The action space of multi-agent spectrum allocation is huge and discrete, which makes the solution space not continuous. Even only a single different RB allocation of one agent can cause an apparent difference on covert performance. This will influence the convergence of DNN and affect the simulation results.

VI. Conclusion and Future Work

In this paper, we explore the spectrum allocation problem of covert performance optimization in the cellular-enabled UAV network. The objective is to give the optimal spectrum allocation with maximum covert rate, while preventing the covert signal from being detected and guaranteeing the QoS of cellular communications. To address this problem, we propose a multi-agent DRL framework to learn the hidden patterns about the spectrum allocation and solve the nonconvex problem.

We program the proposed framework based on Pytorch to verify its performance by experiments. Simulation results show that our framework can effectively solve the problem, but there are some cases in which the framework can not converge quickly and well, and even can not give an effective allocation. This is the problem we need to solve in the future. We will check and find out how to improve the framework performance in the future, such as the environment parameters, DRL algorithms, and so on.

Furthermore, considering the problem stated in the last paragraph in Section V, we plan to take into account the factor of variable power in the following work, which will make the system much more complicated. The DRL algorithm would also become much harder to design. Regardless of these future works, this paper can still provide some ideas and models for other works using ML, especially DRL methods.

References

[1] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

[2] B. A. Bash, D. Goeckel, and D. Towsley, "Square root law for communication with low probability of detection on AWGN channels," in *2012 IEEE International Symposium on Information Theory Proceedings*, 2012, pp. 448–452.

[3] Y. Jiang, L. Wang, H. Zhao, and H.-H. Chen, "Covert Communications in D2D Underlaying Cellular Networks With Power Domain NOMA," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3717–3728, 2020.

[4] X. Zhou, S. Yan, J. Hu, J. Sun, J. Li, and F. Shu, "Joint Optimization of a UAV's Trajectory and Transmit Power for Covert Communications," *IEEE Transactions on Signal Processing*, vol. 67, no. 16, pp. 4276–4290, 2019.

[5] H.-M. Wang, Y. Zhang, X. Zhang, and Z. Li, "Secrecy and Covert Communications Against UAV Surveillance via Multi-Hop Networks," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 389–401, 2020.

[6] A. Sheikholeslami, M. Ghaderi, D. Towsley, B. A. Bash, S. Guha, and D. Goeckel, "Multi-Hop Routing in Covert Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3656–3669, 2018.

[7] S. Yan, S. V. Hanly, and I. B. Collings, "Optimal Transmit Power and Flying Location for UAV Covert Wireless Communications," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3321–3333, 2021.

[8] S. Yan, S. V. Hanly, I. B. Collings, and D. L. Goeckel, "Hiding unmanned aerial vehicles for wireless transmissions by covert communications," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[9] X. Jiang, Z. Yang, N. Zhao, Y. Chen, Z. Ding, and X. Wang, "Resource Allocation and Trajectory Optimization for UAV-Enabled Multi-User Covert Communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1989–1994, 2021.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[11] X. Zhang, Z. Lin, B. Ding, B. Gu, and Y. Han, "Deep Multi-Agent Reinforcement Learning for Resource Allocation in D2D Communication Underlaying Cellular Networks," in *2020 21st Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2020, pp. 55–60.

[12] J. Cui, Y. Liu, and A. Nallanathan, "Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, 2020.

[13] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-Efficient UAV Control for Effective and Fair Communication Coverage: A Deep Reinforcement Learning Approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, 2018.

[14] K. Chi, Z. Chen, K. Zheng, Y.-h. Zhu, and J. Liu, "Energy Provision Minimization in Wireless Powered Communication Networks With Network Throughput Demand: TDMA or NOMA?" *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6401–6414, 2019.

[15] Z. Chen, K. Chi, K. Zheng, Y. Li, and X. Liu, "Common Throughput Maximization in Wireless Powered Communication Networks With Non-Orthogonal Multiple Access," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7692–7706, 2020.

[16] X. Liao, J. Si, J. Shi, Z. Li, and H. Ding, "Generative Adversarial Network Assisted Power Allocation for Cooperative Cognitive Covert Communication System," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1463–1467, 2020.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014, pp. 2672–2680.

[18] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "How to Make 5G Communications "Invisible": Adversarial Machine Learning for Wireless Privacy," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 763–767.

[19] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning," *arXiv preprint arXiv:1803.11485*, 2018.

[20] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," 2020.