# Dynamic Role Switching Scheme with Joint Trajectory and Power Control for Multi-UAV Cooperative Secure Communication

Ang Gao, Qinyu Wang, Yansu Hu, Wei Liang and Jiankang Zhang

***Abstract*—Due to the high flexibility and mobility, unmanned aerial vehicles (UAVs) can be deployed as aerial relays touring to serve ground users (GUs), especially when the ground base station is temporally damaged. However, the broadcasting nature of wireless channels makes such communication vulnerable to be wiretapped by malicious eavesdropping users (EUs). Besides the collecting offloading data for legitimate GUs, UAVs are also expected to be friendly jammers, i.e., generating artificial noise (AN) to deteriorate the wiretapping of EUs. With this in mind, a novel role switching scheme (RSS) is proposed in the paper to guarantee the secure communication by the cooperation of multiple UAVs, where each UAV is allowed to switch its role as a collector or a jammer autonomously to explore a wider trajectory space. It's worthy to be noticed that the joint optimization for the trajectory of UAVs and the transmission power of GUs and UAVs with role switching scheme is a non-convex mixed integer non-linear programming (MINLP) problem. Since the relaxation of binary variables will lead the solution dropping into local minimum, a deep reinforcement learning (DRL) combined successive convex approximate (SCA) algorithm is further designed to maximize the achievable secrecy rate (ASR) of GUs. Numerical results illustrate that compared with the role fixed scheme (RFS) and relaxation based SCA approaches, the proposed DRL-SCA algorithm endows UAVs the capacity to fly close enough to target users (both GUs and EUs) with less moving distance which brings better ASR and less energy consumption.**

***Index Terms*—Unmanned Aerial Vehicles, Secure Communication, Role Switching, Deep Reinforcement Learning, Successive Convex Approximate**

## I. Introduction

Due to the mobility, flexibility as well as high probability of line-of-sight (LoS) propagation, unmanned aerial vehicles (UAVs) are envisaged as a promising technology to assist the mobile wireless communication, known as *UAV-assisted communication networks*, where UAVs can be flexibly deployed as aerial relays to collect the offloading data of ground users (GUs). However, the LoS propagation of UAVs is a double-edged sword that the broadcasting nature makes the aerial communications be exposed to malicious eavesdropping users (EUs) and easy to be wiretapped.

Unfortunately, traditional encryption techniques require high computational complexity leading to a large amount of energy consumption which may be not affordable for UAV systems. As an alternative, physical layer security (PLS) is computationally efficient to protect against the potential wiretapping by exploiting the inherent randomness of wireless channels [1], [2]. UAVs are expected to fly close to legitimate users to enhance the communication quality while as far away from eavesdroppers as possible to avoid the wiretapping by trajectory optimization [3], [4], which is known as the '*passive touring scheme*'. Besides, UAVs can also work as friendly jammers to safeguard legitimate users by actively generating artificial noise (AN) to suppress the wiretapping of eavesdroppers [5]–[9], which is known as the '*active jammer scheme*'.

Some existing researches have payed much attention to the subject. A dual-UAV cooperative jamming model for secure communications is proposed where when one UAV collects data from users at uplink [5], [6] or disseminates data at downlink [7], [8], a secondary UAV cooperatively acts as the jammer to disrupt EUs by generating AN synchronously. To further exploit the potential of friendly jammers, the double antennas UAV is studied in [10], [11], which is capable of collecting offloading data and sending jamming signals simultaneously. Subsequently, AN-beamforming and cooperative jamming relying on the location and statistical channel state information (CSI) of eavesdroppers are considered in [12]. The research in [13] aims for the maximization of worst-case downlink secrecy energy efficiency (SEE) in multiple UAVs cooperative communication, where source UAVs cooperatively transmit information to the legitimate users while jamming UAVs are leveraged to send jamming signals to the eavesdroppers.

In general, the literatures above formulate the secure communication as an optimization issue which can be solved by block coordinate descent (BCD) or successive convex approximation (SCA). However, there are still some challenges to be improved:

- Although UAVs are equipped with the capacity to collect the offloading data of GUs and suppress the wiretapping of EUs, they have to be fixed at a specific role during the whole flight. We believe that the role switching of UAVs will increase the flying flexibility and thus be possible to explore a more optimal trajectory.
- No matter pursuing better secrecy rate [5]–[7] or general communication rate [14], [15], the existence of binary variables makes the optimization be a mixed integer non-linear programming (MINLP) problem. Even though it can be solved by the convex approximation based on the binary variables relaxation, the solution may not perform well and drop into local minimum when a large number

The work was supported in part by Shaanxi Keypoint Research and Invention Program under Grant 2021GXLH-01-15, in part by Xi'an Technology Plan under Grant 21RGZN0018, in part by Shaanxi Natural Science Foundation under Grant 2021JM-186, and in part by Taicang Keypoint Science and Technology Plan under Grant TC2019SF03. *(Corresponding author: Yansu Hu, e-mail: huyansu@chd.edu.cn)*

A. Gao, Q. Wang and W. Liang are with the School of Electronics and Information, Northwestern Polytechnical University, China. Y. Hu is with the School of Electronic and Control Engineering, Chang'an University, China. J. Zhang is with the Department of Computing and Informatics, Bournemouth University, UK.

of UAVs and GUs are involved in the system.

Recently, deep reinforcement learning (DRL) has been widely used for the trajectory optimization of UAVs in uncertain environments for their powerful non-linear approximation capability [16]–[19], which makes it easier to solve the model-free and complex problems by training neural networks. In specific, the work in [16], [17] adopts multi-agent deep deterministic policy gradient (MADDPG) to exploit the potential of cooperative friendly jammer UAVs by 3D beamforming in urban environments where exist random non-line-of-slight (NLoS) links caused by the terrain reflection. The bi-directional secrecy communication between UAVs and ground devices is considered in [18], where the ground devices are supposed to be mobile. The joint optimization for trajectory and transmit power of UAVs is formulated as a constrained Markov decision process (CMDP), and then solved by deep deterministic policy gradient (DDPG), which can be regarded as a centralized version of MADDPG. The work in [19] further extends MADDPG to a more challenging scenario that one UAV acts as a smart malicious eavesdropper and intelligently optimizes its trajectory to increase the wiretapping rate. Since the trajectory of the eavesdropper UAV can not be obtained in advance, the secure communication becomes a dynamic uncooperative game and hard to be solved by traditional algorithms.

Despite of the great success in combating the environments' uncertainty, DRL needs to be trained by amount of episodes which is time-consuming. The slow convergence makes it inapplicable in UAV-assisted communication networks with high real-time requirements. In our previous work [20], game theory is introduced to help speeding up the trajectory optimization of UAVs. Although the physical security is out of consideration, it inspires us to combine traditional theories with machine learning to pursue a near closed-form solution with low-complexity. The main contributions of the paper are as follows:

- A challenging scenario is considered for the secure communication where multiple UAVs are dispatched as mobile collectors to gather the offloading data of GUs in the presence of potential eavesdroppers. To achieve a more satisfied secrecy rate, a dynamic role switching scheme (RSS) is creatively designed for UAVs, i.e., each UAV can choose to act as a collector or a jammer by dynamic role switching. Due to the battery or fuel limitation of GUs and UAVs, the role assignment, trajectory as well as transmission power of both GUs and UAVs should be jointly optimized to maximize the achievable secrecy rate (ASR).
- A DRL combined successive convex approximate (SCA) algorithm is proposed to tackle such non-convex MINLP problem. In specific, the binary variables related to the role assignment of UAVs are solved by DRL directly without relaxation, while the trajectory and transmission power are jointly optimized by SCA in sequence to speed up the convergence. The numerical results demonstrate that the proposed DRL-SCA algorithm is more effective in exploring better trajectory compared with the role fixed scheme (RFS) and avoiding dropping into local minimum compared with binary relaxation. As a result, UAVs with more flexibility are able to fly closer to target users with less moving distance, which not only enhances ASR, but also saves the transmission energy for both GUs and UAVs.

The differences between our work and the existing literatures are summarized in TABLE I. The rest of paper is organized as follows. Sec. II describes the mathematical model of the multi-UAV assisted secure communication with RSS. The proposed DRL-SCA algorithm for optimization is detailed in Sec. III, and the convergence and complexity are analyzed in Sec.IV. The numerical results are demonstrated in Sec.V, while the algorithm deployment is discussed and the future work is prospected in Sec.VI. The paper is concluded in Sec.VII.

## II. System Model and Problem Formulation

### A. System Model

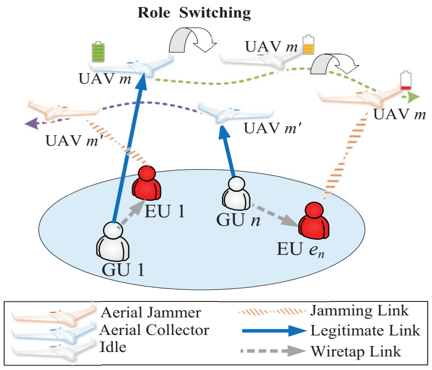


Fig. 1: System model.

Considering a challenging scenario shown as Fig. 1, $M$ UAVs are dispatched as mobile collectors to gather the information from $N$ fixed legitimate GUs in the presence of $E$ non-colluding malicious EUs. Since GUs adopt orthogonal frequency division multiple (OFDM) to transmit the offloading data to UAVs, the channel assignment can be determined via a special pre-assigned command channel maintained by a central controller at the base station or an unique leader UAV. Each UAV is equipped with only one transceiver taking half duplex operation to avoid undesired self-interference. To take full use of the on-board resources, UAVs who are not collecting data for GUs are supposed to generate artificial noise to suppress the wiretapping of EUs, which means each UAV either works as an collector or a jammer by dynamic role switching.

- Collector UAVs tend to fly close to GUs to establish better legitimate channels, while jammer UAVs aim at suppressing eavesdroppers by generating artificial noise.
- GUs have to carefully adapt the transmission power dynamically to guarantee a satisfied offloading rate but not to be wiretapped.
- Since there is no pre-knowledge of channels assignment of GUs, EUs do not know whether the wiretapped data is from GUs or contaminated by artificial noise. Besides,

TABLE I: Comparison with existing literatures

| **Reference** | Active jammer | | Passive touring | Multi-UAV | Optimization objective | DRL | SCA | Notes |
|---|---|---|---|---|---|---|---|---|
| | Power splitting★ | Friendly jamming | | | | | | |
| [3], [4] | | | ✓ | | ASR | | ✓ | ★ Each UAV is equipped with double antennas to transmit confidential messages and AN simultaneously by power splitting.<br>† Secrecy energy efficiency (SEE), considering the energy consumption costed by UAV's propulsion power.<br>‡ Considering the energy consumption costed by both UAV's propulsion power and transmission power. |
| [5], [6] | | ✓ | | Dual-UAV | SEE† | | ✓ | |
| [7], [8] | | ✓ | | Dual-UAV | ASR | | ✓ | |
| [13] | | ✓ | | ✓ | SEE‡ | | ✓ | |
| [10] | ✓ | | | | SEE† | | ✓ | |
| [11] | ✓ | | | | ASR | | ✓ | |
| [16], [17] | | ✓ | | ✓ | ASR | ✓ | | |
| [18] | | | ✓ | ✓ | ASR | ✓ | | |
| Our Work | Friendly Jamming with role switching | | | ✓ | ASR | ✓ | ✓ | |

to guarantee the concealment during wiretapping, EUs should keep silent and not collude with each other [21]. Otherwise, they would be detected by the unwanted leakage of electronic waveforms.

Let $\mathcal{M} = \{1, 2, \ldots, M\}$, $\mathcal{N} = \{1, 2, \ldots, N\}$ and $\mathcal{E} = \{1, 2, \ldots, E\}$ denote the sets of UAVs, GUs and EUs, respectively. Binary matrixes $\mathbf{u}_m[t] = \{u_{m,n}[t]\} \in \mathbb{Z}^{N\times 1}$ and $\mathbf{v}_m[t] = \{v_{m,e}[t]\} \in \mathbb{Z}^{E\times 1}$ are adopted to represent the role of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot, where $u_{m,n}[t] \in \{0,1\}, v_{m,n}[t] \in \{0,1\}$. When $u_{m,n}[t] = 1$, the $m^{\text{th}}$ UAV collects the offloading data for the $n^{\text{th}}$ GU. Similarly, when $v_{m,e}[t] = 1$, the $m^{\text{th}}$ UAV will act as a jammer to suppress the wiretapping of the $e^{\text{th}}$ EU. For each GU, there is at most one UAV to collect its offloading data at a time slot. And for each UAV, it can only serve one GU at most at a time slot. So there are constraints:

$$u_{m,n}[t] \in \{0,1\}, v_{m,n}[t] \in \{0,1\}, \tag{C1}$$

$$0 \leq \sum\nolimits_{m=1}^{M} u_{m,n}[t] \leq 1, \forall n \in \mathcal{N}, \tag{C2}$$

$$0 \leq \sum\nolimits_{n=1}^{N} u_{m,n}[t] \leq 1, \forall m \in \mathcal{M}. \tag{C3}$$

Similarly, each jammer UAV can only suppress one EU at most at a time slot:

$$0 \leq \sum\nolimits_{e=1}^{E} v_{m,e}[t] \leq 1, \forall m \in \mathcal{M}. \tag{C4}$$

For ease of reference, the main notations are summarized in TABLE II.

### B. Movement Model

The finite flight period $T_0$ of UAVs can be equally divided into $T = T_0/\Delta t$ time slots, where $\Delta t$ should be small enough so that the location of UAVs can be regarded as unchanged within a time slot. For mathematical clarity, the 2-dimensional (2D) Cartesian coordinate system is adopted [1], where $\boldsymbol{q}_m[t] = [q_{xm}[t], q_{ym}[t]]^{\mathrm{T}}$ denotes the location of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot.

In general, GUs and EUs move slowly enough compared with UAVs, so their location can be assumed to be static and denoted by $\boldsymbol{q}_n = [q_{xn}, q_{yn}]^{\mathrm{T}}$ and $\boldsymbol{q}_e = [q_{xe}, q_{ye}]^{\mathrm{T}}$, respectively.

The maximum speed of UAVs should be limited by $v^{\max}$, then the maximum moving distance in each time slot is $d^{\max} = v^{\max}\Delta t$. Since UAVs have to return to a pre-fixed docking station for energy refueling, their initial location $\boldsymbol{q}_m^{\mathrm{I}}$ and final location $\boldsymbol{q}_m^{\mathrm{F}}$ should be assigned in advance.

In summary, the Kinematic constraints of UAVs are formulated as following, where $d^{\min}$ is the minimum safe distance among UAVs to avoid collisions, and $\|\cdot\|$ means the Euclid distance.

$$\boldsymbol{q}_m[0] = \boldsymbol{q}_m^{\mathrm{I}} \ , \tag{C5}$$

$$\boldsymbol{q}_m[T] = \boldsymbol{q}_m^{\mathrm{F}} \ , \tag{C6}$$

$$\| \boldsymbol{q}_m[t+1] - \boldsymbol{q}_m[t] \|^2 \leq (d^{\max})^2 \ , \tag{C7}$$

$$\| \boldsymbol{q}_m[t] - \boldsymbol{q}_{m'}[t] \|^2 \geq (d^{\min})^2 \ , \forall m, m' \in \mathcal{M}, m \neq m' \ , \tag{C8}$$

### C. Communication Model

In general, channels between UAVs and GUs are considered to be dominated by LoS links [2] and follow quasi-static block fading, i.e., the channel gain $g_{n,m}[t] = \frac{\beta_0}{\|\boldsymbol{q}_m[t] - \boldsymbol{q}_n\|^2}$ remains unchanged in each time slot, where $\beta_0$ is the channel gain at the reference distance $d_0 = 1$m. Similarly, the jamming channel gain of the $m^{\text{th}}$ UAV to the $e^{\text{th}}$ EU is defined as $g_{m,e}[t] = \frac{\beta_0}{\|\boldsymbol{q}_m[t] - \boldsymbol{q}_e\|^2}$.

Since GUs and EUs are all on the ground, the wiretapping channels can be assumed to be constituted by the distance-dependent path loss with pass-exponent $\alpha > 2$ and the small-scale Rayleigh fading. Therefore, the wiretapping channel gain of the GU-EU pair can be expressed as $g_{n,e} = \frac{\beta_0}{\|\boldsymbol{q}_n - \boldsymbol{q}_e\|^{\alpha}}\zeta$, where $\zeta \sim \mathcal{CN}(0,1)$ is an exponentially distributed random variable with unit mean.

The transmitting power of the $n^{\text{th}}$ GU $p_n[t]$ and the jamming power of the $m^{\text{th}}$ UAV $p_m^{\mathrm{J}}[t]$ are limited by both the average value and peak value, which leads to the constraints:

$$\frac{1}{T}\sum\nolimits_{t=1}^{T} p_n[t] \leq P_{\mathrm{GU}}^{\mathrm{ave}} \ , \tag{C9}$$

$$0 \leq p_n[t] \leq P_{\mathrm{GU}}^{\max}, \forall n, t \ . \tag{C10}$$

$$\frac{1}{T}\sum\nolimits_{t=1}^{T} p_m^{\mathrm{J}}[t] \leq P_{\mathrm{UAV}}^{\mathrm{ave}} \ , \tag{C11}$$

$$0 \leq p_m^{\mathrm{J}}[t] \leq P_{\mathrm{UAV}}^{\max}, \forall m, t \ . \tag{C12}$$

### D. Worst-Case Secrecy Rate

In the $t^{\text{th}}$ time slot , the offloading rate per Hz of the $n^{\text{th}}$ GU to the $m^{\text{th}}$ UAV is:

$$R_{n,m}[t] = \log_2\left(1 + \frac{p_n[t]g_{n,m}[t]}{\sum_{m'=1,m'\neq m}^{M} v_{m',e}[t]p_{m'}^{\mathrm{J}}[t]g_{m',m}[t] + \delta_L^2}\right), \tag{1}$$

[1] The 3D movement of UAVs can be handled in the similar way.

[2] LoS probability of the air-to-ground link can be approximate to 1 when UAVs are above 120 meters [22]–[24].

TABLE II: Main notations

| | | | |
|---|---|---|---|
| $\mathcal{M}$ | UAVs set | $\mathbf{Q}_{\mathrm{C}} = \{\mathbf{q}_{\mathrm{C},m}\}$ | Way points set of all UAVs acting as collectors |
| $\mathcal{N}$ | Legitimate GUs set | $\mathbf{Q}_{\mathrm{J}} = \{\mathbf{q}_{\mathrm{J},m}\}$ | Way points set of all UAVs acting as jammers |
| $\mathcal{E}$ | Potential malicious EUs set | $\mathbf{Q} \in \mathbb{R}^{T\times 2M}$ | Way points matrix of all UAVs |
| $u_{m,n}[t]$ | Service indication of the $m^{\text{th}}$ UAV to $n^{\text{th}}$ GU at the $t^{\text{th}}$ time slot | $p^{\mathrm{J}}_m[t]$ | Jamming power of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot |
| $v_{m,e}[t]$ | Jamming indication of the $m^{\text{th}}$ UAV to $e^{\text{th}}$ EU at the $t^{\text{th}}$ time slot | $p_n[t]$ | Transmission power of the $n^{\text{th}}$ GU at the $t^{\text{th}}$ time slot |
| $\mathbf{u}_m[t] \in \mathbb{Z}^{N\times 1}$ | Service matrix of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot | $\mathbf{p}_n \in \mathbb{R}^T$ | Transmission power vector of the $n^{\text{th}}$ GU throughout the duration $T$ |
| $\mathbf{v}_m[t] \in \mathbb{Z}^{E\times 1}$ | Jamming matrix of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot | $\mathbf{p}^{\mathrm{J}}_m \in \mathbb{R}^T$ | Jamming power vector of the $m^{\text{th}}$ UAV throughout the duration $T$ |
| $\mathbf{U} \in \mathbb{Z}^{M\times T\times N}$ | Service matrix of all UAVs | $\mathbf{P} \in \mathbb{R}^{T\times(N+M)}$ | Power matrix of all UAVs and GUs |
| $\mathbf{V} \in \mathbb{Z}^{M\times T\times E}$ | Jamming matrix of all UAVs | $g_{n,m}$ | Channel gain between the $m^{\text{th}}$ |
| $\boldsymbol{q}_m[t]$ | Position of the $m^{\text{th}}$ UAV at the $t^{\text{th}}$ time slot | $g_{m,e_n}$ | Channel gain between the $m^{\text{th}}$ UAV and the $e_n^{\text{th}}$ EU. |
| $\boldsymbol{q}_n$ | Position of the $n^{\text{th}}$ GU | $g_{n,e_n}$ | Channel gain between the $n^{\text{th}}$ GU and the $e_n^{\text{th}}$ EU |
| $\boldsymbol{q}_e$ | Position of the $e^{\text{th}}$ EU | $g_{m,m'}$ | Channel gain between the $m^{\text{th}}$ UAV and the $m'^{\text{th}}$ UAV. |
| $\boldsymbol{q}^{\mathrm{I}}_m$, $\boldsymbol{q}^{\mathrm{F}}_m$ | Initial and final location of the $m^{\text{th}}$ UAV, respectively | $R_{n,m}$ | Transmission rate of the $n^{\text{th}}$ GU to the $m^{\text{th}}$ UAV |
| $\mathbf{q}_m \in \mathbb{R}^{2\times T}$ | Way points vector of the $m^{\text{th}}$ UAV throughout the duration $T$ | $R_{n,e_n}$ | Wiretapping rate of the $e_n^{\text{th}}$ EU to the $n^{\text{th}}$ GU |
| $\mathbf{q}_{\mathrm{C},m}$ | Way points set of the $m^{\text{th}}$ UAV acting as collector | $D_n$ | Task or Data size of the $n^{\text{th}}$ GU to be offloaded |
| $\mathbf{q}_{\mathrm{J},m}$ | Way points set of the $m^{\text{th}}$ UAV acting as jammer | $B_\omega$ | Bandwidth of each channel |
| $T_0$ | Total flight period | $T$ | Number of time slots |

where $\delta_L^2$ denotes the additive white Gaussian noise (AWGN) power, and $g_{m',m}[t] = \frac{\beta_0}{\|\boldsymbol{q}_m[t]-\boldsymbol{q}_{m'}[t]\|^2}$ is the air-to-air interference channel gain from UAV $m'$ to UAV $m$ ($m' \neq m$). Thus the summation $\sum_{m'=1,m'\neq m}^{M} v_{m',e}[t]p^{\mathrm{J}}_{m'}[t]g_{m',m}[t]$ in denominator is the interference of all jammer UAVs to the $m^{\text{th}}$ UAV which currently acts as a collector.

Jammer UAVs also generate interference to EUs, so the wiretapping rate of the $e^{\text{th}}$ EU to the $n^{\text{th}}$ GU can be formulated as:

$$R_{n,e}[t] = \log_2\left(1 + \frac{p_n[t]g_{n,e}}{\sum_{m=1,m'\neq m}^{M} v_{m',e}[t]p^{\mathrm{J}}_{m'}[t]g_{m',e}[t] + \delta_E^2}\right) . \tag{2}$$

**Lemma 1.** *Assume that EUs will non-selectively wiretap the channel with the largest received power (RP). Then the worst case for a legitimate GU is to be wiretapped by the most nearby eavesdropper, i.e.,* $e_n = argmax\ g_{n,e}$ ($e \in \mathcal{E}$, $\forall n \in \mathcal{N}$).

*Proof.* If EU $e'$ rather than $e_n$ detects a larger received power at the channel occupied by the $n^{\text{th}}$ GU, and the $m^{\text{th}}$ jammer UAV is sending AN to interfere with the channel, then according to the 'largest received power' wiretapping strategy, there is:

$$\| p^{\mathrm{J}}_m g_{m,e'} + p_n g_{n,e'} + {\delta_E}^2 \|^2 \geq \| p^{\mathrm{J}}_m g_{m,e_n} + p_n g_{n,e_n} + {\delta_E}^2 \|^2 . \tag{3}$$

Since $e_n$ is supposed to be the most nearby eavesdropper for the $n^{\text{th}}$ GU, i.e., $g_{n,e_n} > g_{n,e'}$, there must be $g_{m,e'} > g_{m,e_n}$ to hold the inequality (3) true. Because of $g_{n,e_n} > g_{n,e'}$, $g_{m,e'} > g_{m,e_n}$ and the monotonic increasing of $\log_2(\cdot)$ function, there is:

$$\log_2(1 + \frac{P_n g_{n,e'}}{P^{\mathrm{J}}_m g_{m,e'}}) < \log_2(1 + \frac{P_n g_{n,e_n}}{P^{\mathrm{J}}_m g_{m,e_n}}) , \tag{4}$$

which finally leads to $R_{n,e'} < R_{n,e_n}$, i.e., the theoretical wiretapping rate of EU $e'$ is still less than that of $e_n$. In other words, EU $e'$ tends to wiretap the channel with the largest received power, even though the channel is not occupied by the closest GU currently. If doing so, its theoretical wiretapping rate will be smaller than that of the most nearby eavesdropper $e_n$. It implies that no matter how eavesdroppers change their wiretapped channels, it is still the most effective choice for jammer UAVs to suppress the eavesdropper closest to one specific GU. □

### E. Role Switching of UAVs

Based on **Lemma 1**, the $n^{\text{th}}$ GU can be automatically paired with the most nearby eavesdropper $e_n$. When the number of GUs is more than EUs, the absent EUs can be easily treated as infinite far away. Since there is the possibility for the $n^{\text{th}}$ GU to be wiretapped during the transmission, the $m^{\text{th}}$ UAV should be safeguarded by another jammer UAV $m'$ as long as it works as a collector, i.e., $u_{m,n} = 1$. So there is the constraint:

$$\sum_{m=1}^{M} u_{m,n}[t] - \sum_{m'=1}^{M} v_{m',e_n}[t] = 0, \forall n, t\ , \tag{C13}$$

which implies that the number of UAVs who act as collectors and jammers should be the same at a time slot. It is worth noting that the coupling between 'collector-jammer' UAVs undergoes dynamic changes across different time slots, and there is no requirement for the number of UAVs to be even. Specifically, once a pair of 'collector-jammer' UAVs completes their service for the current GU, they will be decoupled and have the possibility to form a new pair with other UAVs, being reassigned new roles to serve another GU.

Because of the half duplex operation, each UAV can only play one role at the same time, i.e., either an collector or a jammer by dynamic role switching.

$$0 \leq \sum_{n=1}^{N} u_{m,n}[t] + \sum_{e_n=1}^{E} v_{m,e_n}[t] \leq 1, \forall m, t\ . \tag{C14}$$

The secure transmission can be effectively improved by the cooperation of multiple UAVs involved in the system. In specific, when collector UAVs move close to GUs for a better offloading rate by trajectory optimization, other jammer UAVs should generate AN to against possible EUs nearby at the same time.

### F. Problem Formulation

The worst-case achievable secrecy transmission rate of the $n^{\text{th}}$ GU can be expressed by:

$$R_n^{\text{sec}}[t] = \left[0, \sum\nolimits_{m=1}^{M} u_{m,n}[t]R_{n,m}[t] - R_{n,e_n}[t]\right]^+, \forall n, t, \tag{5}$$

where $[\cdot]^+ \overset{\Delta}{=} max\{\cdot\}$.

Since UAVs are battery or fuel driven, it is essential to make sure that the data offloading of GUs should be accomplished within the limited time slots $T$:

$$\sum_{t=1}^{T}\sum_{m=1}^{M} B_{\omega} u_{m,n}[t] R_{n,m}[t] \geq D_n, \forall n \in \mathcal{N} \ , \tag{C15}$$

where $D_n$ is the byte size of the offloading tasks of the $n^{\text{th}}$ GU and $B_{\omega}$ is the bandwidth for legitimate transmission.

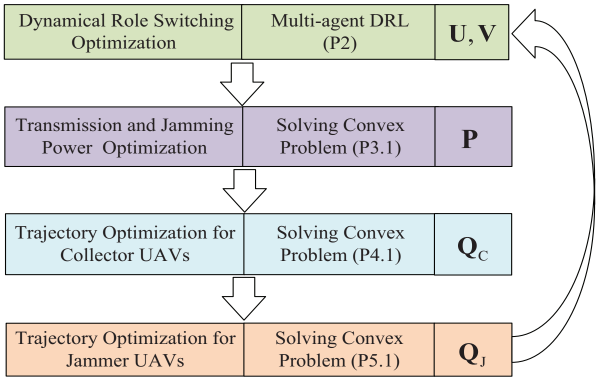
Let $\mathbf{P} = \left[\mathbf{p}_1^{\mathrm{T}}, \cdots, \mathbf{p}_N^{\mathrm{T}}, \mathbf{p}_1^{\mathrm{J}\,\mathrm{T}}, \cdots, \mathbf{p}_M^{\mathrm{J}\,\mathrm{T}}\right]$ be the power vector concatenation of GUs and jammer UAVs, $\mathbf{Q} = \left[\mathbf{q}_1^{\mathrm{T}}, \cdots, \mathbf{q}_M^{\mathrm{T}}\right]$ be the trajectory matrix of UAVs, and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M] \in \mathbb{Z}^{M\times T\times N}$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M] \in \mathbb{Z}^{M\times T\times E}$ be the role switching matrix of UAVs. The problem can be formulated to maximize ASR of all GUs by jointly optimizing all the variables above:

$$\begin{aligned} \mathrm{P1}: \max_{\mathbf{Q},\mathbf{U},\mathbf{V},\mathbf{P}} \ \min \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} R_{\text{sec}}^{n}[t] \ , \\ \text{s.t.} \quad \text{(C1)-(C15)}, \end{aligned} \tag{6}$$

where (C1)-(C4), (C13)-(C14) are the requirements for the uniqueness of UAVs' role assignment, (C5)-(C8) are the trajectory Kinetic limitations of UAVs' movement, (C9-(C12) are the average and maximum transmitting power limitation for GUs and UAVs, and (C15) means that GUs should finish data offloading within limited time slots $T$.

Notice that the objective function (6) is non-smooth due to the operator $[\cdot]^{+}$ and the binary matrix $\mathbf{U}$ and $\mathbf{V}$. Besides, (C8) and (C15) are non-convex. Therefore, P1 is a non-convex MINLP problem which is difficult to be optimally solved in general.

## III. DRL-SCA for multi-UAV Assisted Secure Communication



Dynamical Role Switching Optimization
Multi-agent DRL (P2)
U, V
Transmission and Jamming Power Optimization
Solving Convex Problem (P3.1)
P
Trajectory Optimization for Collector UAVs
Solving Convex Problem (P4.1)
Q_C
Trajectory Optimization for Jammer UAVs
Solving Convex Problem (P5.1)
Q_J

Fig. 2: DRL-SCA optimization framework.

As shown in Fig. 2, P1 can be divided into four sub-problems by block coordinate descent (BCD) approach for iterative optimization. In specific, the role switching of UAVs which is indicated by binary variables is learned by multi-agent DRL denoted as P2. The continuous variables, i.e., the transmitting power of GUs and jammer UAVs, are modeled as P3, and the flight trajectory of UAVs acting as collectors and jammers is formulated as P4 and P5, respectively. Such DRL-SCA algorithm for multi-UAV assisted secure communication with RSS is detailed in the following.

For notation brevity, the time slot indication $t$ is omitted in the following unless otherwise stated.

### A. Dynamic Role Switching Optimization by MADRL

Since the role assignment of UAVs is independent with other continuous variables in P1, it can be solved by multi-agent DRL (MADRL) in advance with the given trajectory of UAVs and given transmitting power of GUs and jammer UAVs. The sub-problem can be formulated as:

$$\begin{aligned} \mathrm{P2}: \max_{\mathbf{U},\mathbf{V}} \ \min \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} R_{n}^{\text{sec}} \ , \\ \text{s.t.} \quad \text{(C1)-(C4), (C13)-(C14)} \ . \end{aligned} \tag{7}$$

Shown as Fig. 3, each UAV acts as an individual agent and adopts double deep $Q$-network (DDQN) to learn interacting with the environment. In specific, UAVs can be trained to map a specific *State* consisting of the power vector $\mathbf{P}$ and the trajectory vector $\mathbf{Q}$ to a proper *Action* denoted by $\mathbf{U}$ and $\mathbf{V}$ during the flight to make a better *Reward*. It is worthy noting that the matrixes $\mathbf{U}$ and $\mathbf{V}$ not only indicate the role of UAVs, but also imply which GUs or EUs should be assigned to a specific UAV. It is a basic condition for the optimization of following sub-problems.

- *State* $\mathcal{S}_m^k = \{\mathbf{Q}^k, \mathbf{P}^k\}$ is the global observation of the $m^{\text{th}}$ UAV to the overall system. All agents share the same state.
- *Action* $\mathcal{A}_m^k = \{\mathbf{u}_m^k, \mathbf{v}_m^k\}$ is the role and service assignment for each agent.
- *Reward* $r_m^k$ is defined on account of ASR and returned from the environment when the agent takes action $\mathcal{A}_m^k$.

$$r_m^k = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{n}^{N} u_{m,n}[t]R_{n,m}[t] - \sum_{e=1}^{E} v_{m,e}[t]R_{n,e_n}[t]\right) \ , \tag{8}$$

where the superscript $k$ is used to denote the interaction step number.

On the right side of (8), the first term is the cumulative transmission capacity of the $m^{\text{th}}$ UAV acting as a collector, while the second term represents the blocking effectiveness of the $m^{\text{th}}$ UAV which is taken as a negative value of the wiretapping rate. The summation reward $\sum_{m=1}^{M} r_m^k = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} R_{\text{sec}}^{n}[t]$ is identical to the object of P1.

DDQN, inherited from DQN, adopts double networks, i.e., an on-line network $Q(\mathcal{S},\mathcal{A}|\theta)$ and a target network $\hat{Q}(\mathcal{S},\mathcal{A}|\hat{\theta})$, to decouple the action generation and Q-value evaluation, which improves the learning stability and overcomes the over-optimistic in large-scale problems [25], where $\theta$ and $\hat{\theta}$ denote the corresponding network parameters, respectively.

Another essential technique of DDQN is equipped with the replay buffer (RB) to store the (state, action,reward, next state) transition $\{\mathcal{S},\mathcal{A},r,\mathcal{S}'\}$, which will be fetched out randomly for the further calculation of loss function. Together with mini-batch, RB can effectively avoid the highly correlation of successive updating.

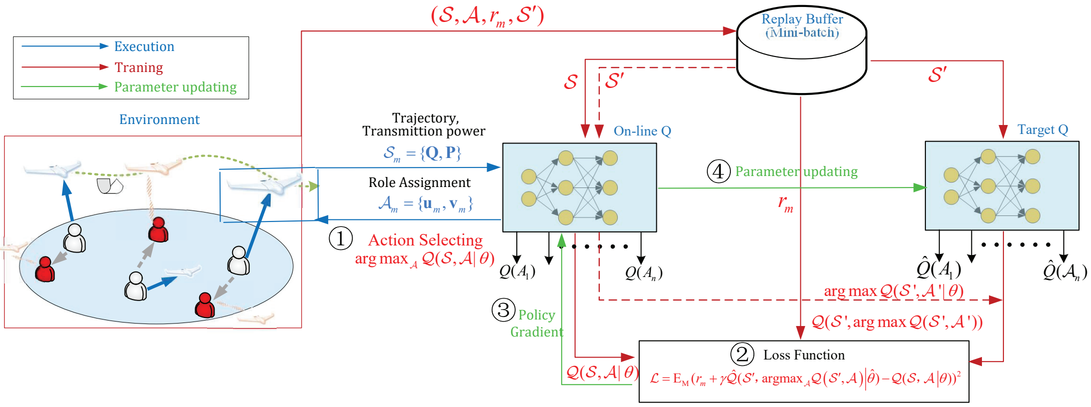

Execution
Traning
Parameter updating
Environment
$(\mathcal{S},\mathcal{A},r_m,\mathcal{S}')$
Replay Buffer (Mini-batch)
$\mathcal{S}$
$\mathcal{S}'$
$\mathcal{S}'$
Trajectory, Transmittion power
$\mathcal{S}_m=\{\mathbf{Q},\mathbf{P}\}$
Role Assignment
$\mathcal{A}_m=\{\mathbf{u}_m,\mathbf{v}_m\}$
① Action Selecting
$\arg\max_{\mathcal{A}} Q(\mathcal{S},\mathcal{A}|\theta)$
On-line Q
④ Parameter updating
$r_m$
Target Q
$Q(\mathcal{A}_1)$
$Q(\mathcal{A}_n)$
$\hat{Q}(\mathcal{A}_1)$
$\hat{Q}(\mathcal{A}_n)$
$\arg\max Q(\mathcal{S}',\mathcal{A}'|\theta)$
$Q(\mathcal{S}',\arg\max Q(\mathcal{S}',\mathcal{A}'))$
③ Policy Gradient
$Q(\mathcal{S},\mathcal{A}|\theta)$
② Loss Function
$\mathcal{L}=\mathrm{E}_{\mathrm{M}}(r_m+\gamma\hat{Q}(\mathcal{S}',\mathrm{argmax}_{\mathcal{A}}Q(\mathcal{S}',\mathcal{A})|\hat{\theta})-Q(\mathcal{S},\mathcal{A}|\theta))^2$

Fig. 3: DDQN for role switching of UAVs.

❶ On-line network selects the next action according to the Q-value of each state-action pair with greedy policy, i.e., $\arg\max_{\mathcal{A}} Q(\mathcal{S},\mathcal{A}|\theta)$.

❷ Target network aims to evaluate such action by calculating the temporal difference (TD) error defined as:

$$\mathcal{L}=\mathbb{E}_{\mathbb{M}}\left(r_m+\underbrace{\gamma\hat{Q}(\mathcal{S}',\arg\max_{\mathcal{A}} Q(\mathcal{S}',\mathcal{A})|\hat{\theta})}_{①}-\underbrace{Q(\mathcal{S},\mathcal{A}|\theta)}_{②}\right)^2, \tag{9}$$

which is the mean square error (MSE) of $\mathbb{M}$ number of $(\mathcal{S},\mathcal{A},r_m,\mathcal{S}')$ tetrads randomly selected as a mini-batch. Term ① in (9) is the expected Q-value of the next state-action pair with a discount factor $\gamma$, which is approximated by the target network based on its current policy $\hat{\theta}$. The next action is selected by the on-line network, i.e., $\arg\max_{\mathcal{A}} Q$. While term ② is the predicted Q-value of the current state-action pair by the neural network with parameter $\theta$, and $\mathbb{E}_{\mathbb{M}}$ is the mathematical expectation over the mini-batch. Therefore, the action selection and evaluation can be independent to avoid the over-estimation of Q-value [25].

❸ The on-line network parameter $\theta$ can be iteratively updated by the gradient decent of TD-error $\nabla_\theta\mathcal{L}$, i.e., $\theta_m^{k+1}=\theta_m^k+\alpha_\theta\nabla_\theta\mathcal{L}$, where $\alpha_\theta$ is the step size for parameter updating, and $\nabla_\theta\mathcal{L}$ is the gradient of $\mathcal{L}$ with regard to $\theta$, which can be easily obtained by back propagation.

❹ With regard to the parameter $\hat{\theta}$ of the target network, it will be copied from the on-line network every $\tau$ times iterations, i.e., $\hat{\theta}_m^k=\theta_m^{k-\tau}$. As a result, the role switching and service assignment of UAVs can be obtained once the on-line network has been well trained. The detailed DDQN learning is in Algorithm 1.

### B. Transmission and Jamming Power Optimization

Once the role switching for UAVs has been assigned, the transmission power for GUs and jammer UAVs can be next optimized by convex theory. The sub-problem can be formulated as:

$$\mathrm{P3}:\max_{\mathbf{P}}\ \min\ \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}R_n^{\mathrm{sec}}, \tag{10}$$
$$\text{s.t. (C9)-(C12), (C15) .}$$

**Algorithm 1:** Double DQN for Role Switching

```
Input: Tasks vector $\mathcal{D}=\{D_1,D_2,\cdots,D_N\}$, transmission
       power vector $\mathbf{P}$ and trajectory vector $\mathbf{Q}$ ;
Output: On-line network $Q(\cdot|\theta_m)$;
Obtain the initial observation state $\mathcal{S}$;
if training = true then
    Randomly initialize the on-line $Q_m$ network parameter
      $\theta_m$;
    Initialize the target $\hat{Q}_m$ network parameter $\hat{\theta}_m \leftarrow \theta_m$;
    Empty replay buffer RB;
    Initialize Gaussian noise $\mathbb{N}$;
    for agent $m=1$ to $M$ do
        for step $k=1$ to max-episode-length do
            Select action by
              $\mathcal{A}_m^k=\arg\max_{\mathcal{A}} Q(\mathcal{S}^k,\mathcal{A}_m|\theta_m^k)+\mathbb{N}$;
            Interact with the environment, and obtain the
              reward $r_m^k$ and next state $\mathcal{S}^{k+1}$ to store into
              RB as $(\mathcal{S},\mathcal{A},r_m,\mathcal{S}')$ ;
            Sample a random mini-batch of $\mathbb{M}$ from RB;
            Calculate Q-value $Q(\mathcal{S}^k,\mathcal{A}^k|\theta_m^k)$ according to
              the sampled $(\mathcal{S}^k,\mathcal{A}^k)$ pairs in the mini-batch,
              and obtain the second term of (9);
            Select the action $\arg\max_{\mathcal{A}}$ according to the
              sampled $\mathcal{S}'$ in the mini-batch, and find the
              corresponding target Q-value in $\hat{Q}$, i.e., obtain
              the first term of (9);
            Calculating the TD error for tetrad
              $(\mathcal{S},\mathcal{A},r_m,\mathcal{S}')$ in mini-batch according to (9);
            Update the online network parameters $\theta$ for
              minimizing TD error by gradient decent ;
            Update the target network parameters $\hat{\theta}$ every $\tau$
              iteration interval;
            $k \leftarrow k+1$;
else
    Output action by $\mathcal{A}_m=\arg\max_{\mathcal{A}} Q(\mathcal{S},\mathcal{A}_m|\theta_m)$
```

For clarity, two auxiliary variables $X_n$ and $Y_n$ are introduced to facilitate the derivation. Then $R_n^{\mathrm{sec}}$ in (5) can be rewritten as:

$$R_n^{\mathrm{sec}}=\sum_{m=1}^{M}u_{m,n}\left(\log_2(X_n+\delta_L^2+p_n g_{n,m})-\log_2(X_n+\delta_L^2)\right. \tag{11}$$
$$\left.-\log_2(Y_n+\delta_E^2+p_n g_{n,e_n})+\log_2(Y_n+\delta_E^2)\right),$$

where $X_n=\sum_{m'=1,m'\neq m}^{M}v_{m',e_n}p_{m'}^{\mathrm{J}}g_{m',m}$ and $Y_n=\sum_{m'=1,m'\neq m}^{M}v_{m',e_n}p_{m'}^{\mathrm{J}}g_{m',e_n}$.

Both $X_n$ and $Y_n$ are linear functions of $p_{m'}^{\mathrm{J}}$ and (C9)-(C12) with respect to $p_n$ and $p^{\mathrm{J}}_{m'}$ are convex. However, (C15) is not convex and the objective function (10) is non-concave with respect to $p_n$ and $p^{\mathrm{J}}_{m'}$. Successive convex approximation is used to tackle such issue.

❶ The first-order Taylor expansion is used to approximate the second and the third terms by the global upper-bounded inequality [3]:

$$\begin{aligned}&\sum_{m=1}^{M} u_{m,n}\log_2(X_n+\delta_L^2)+\log_2(Y_n+\delta_E^2+p_n g_{n,e_n})\\&\le\sum_{m=1}^{M} u_{m,n}\log_2(\hat{X}_n+\delta_L^2)+\log_2(\hat{Y}_n+\hat{p}_n g_{n,e_n}+\delta_E^2)\\&+\sum_{m=1}^{M} u_{m,n}\frac{X_n-\hat{X}_n}{\ln 2(\hat{X}_n+\delta_L^2)}+\frac{(Y_n-\hat{Y}_n)+g_{n,e_n}(p_n-\hat{p}_n)}{\ln 2(\hat{Y}_n+\hat{p}_n g_{n,e_n}+\delta_E^2)}\\&\triangleq f_1(X_n,Y_n)\ ,\end{aligned}\tag{12}$$

where $\hat{X}_n=\sum_{m'=1,m'\neq m}^{M} v_{m',e_n}\hat{p}^{\mathrm{J}}_{m'}g_{m',m}$ and $\hat{Y}_n=\sum_{m'=1,m'\neq m}^{M} v_{m',e_n}\hat{p}^{\mathrm{J}}_{m'}g_{m',e_n}$.

Since $\hat{X}_n,\hat{Y}_n$ could be the value of $X_n$, $Y_n$ at any feasible point of $\hat{p}^{\mathrm{J}}_{m'}$, they can be treated as constants. Thus $f_1$ is the linear function of $p_n$, $X_n$ and $Y_n$.

❷ According to (12), define $\hat{R}^{\mathrm{sec}}_n$ as the global lower-bound of $R^{\mathrm{sec}}_n$, which is a concave function to $p_n$ and $p^{\mathrm{J}}_{m'}$.

$$\begin{aligned}R^{\mathrm{sec}}_n\ge&\sum_{m=1}^{M} u_{m,n}\Big(\log_2(X_n+\delta_L^2+p_n g_{n,m})\Big)\\&+\log_2(Y_n+\delta_E^2)-f_1(X_n,Y_n)\triangleq\hat{R}^{\mathrm{sec}}_n.\end{aligned}\tag{13}$$

❸ Similarly, there is the global inequality for $R_{n,m}$ :

$$\begin{aligned}R_{n,m}&=\log_2(X_n+\delta_L^2+p_n g_{n,m})-\log_2(X_n+\delta_L^2)\\&\ge\log_2(X_n+\delta_L^2+p_n g_{n,m})\\&-\left(\log_2(\hat{X}_n+\delta_L^2)+\frac{X_n-\hat{X}_n}{\ln 2(\hat{X}_n+\delta_L^2)}\right)\\&\triangleq R^{\mathrm{lb}}_{n,m}(X_n,\hat{X}_n)\ ,\end{aligned}\tag{14}$$

where $R^{\mathrm{lb}}_{n,m}(X_n,\hat{X}_n)$ is the lower-bound of $R_{n,m}$, and is concave and linear to $p_n$ and $p^{\mathrm{J}}_{m'}$.

❹ In summary, P3 can be approximated by maximizing of the global lower-bound of $R^{\mathrm{sec}}_n$:

$$\begin{aligned}\mathrm{P3.1}:\max_{\mathbf{P}}\ \ &\min\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\hat{R}^{\mathrm{sec}}_n,&(15)\\\text{s.t.}\ \ &(\text{C9})\text{-}(\text{C12})\ ,\\&\sum_{t=1}^{T}\sum_{m=1}^{M}B_{\omega}u_{m,n}R^{\mathrm{lb}}_{n,m}\ge D_n,\forall n\in\mathcal{N}\ .&(\text{C15.a})\end{aligned}$$

The object is to maximize the concave function (15) with respect to $\mathbf{P}$ (equivalent to minimize a convex function), and (C15.a) is a convex set [4]. Therefore, P3.1 is transferred to a convex problem which can be effectively solved by the existing convex toolbox such as CVX.

### C. Trajectory Optimization For Collector UAVs

Based on the role switching assignment and the transmitting power control solved by DRL and SCA respectively, the trajectory of UAVs can be further planned. It should be noticed that the purpose of trajectory optimization for collector UAVs and jammer UAVs are different. In specific, collector UAVs tend to move close to GUs for a better offloading rate. But when UAVs act as jammers, they fly close to EUs to deteriorate the wiretapping. Therefore, UAVs' trajectory are discriminately optimized in Sec.III-C and Sec.III-D, respectively.

The trajectory of the $m^{\text{th}}$ UAV denoted by $\mathbf{q}_m$ can be split into two subsets of vectors, i.e., $\mathbf{q}_{\mathrm{C},m}$ and $\mathbf{q}_{\mathrm{J},m}$, which consist of the way points of the $m^{\text{th}}$ UAV acting as a collector and a jammer, respectively. For all UAVs, there are $\mathbf{Q}=[\mathbf{q}_1^{\mathrm{T}},\cdots,\mathbf{q}_{\mathrm{M}}^{\mathrm{T}}]$, $\mathbf{Q}_{\mathrm{C}}=[\mathbf{q}_{\mathrm{C},1}^{\mathrm{T}},\cdots,\mathbf{q}_{\mathrm{C},M}^{\mathrm{T}}]$ and $\mathbf{Q}_{\mathrm{J}}=[\mathbf{q}_{\mathrm{J},1}^{\mathrm{T}},\cdots,\mathbf{q}_{\mathrm{J},M}^{\mathrm{T}}]$. Due to constraint (C14), a way point in the trajectory of the $m^{\text{th}}$ UAV can not belong to both $\mathbf{q}_{\mathrm{C},m}$ and $\mathbf{q}_{\mathrm{J},m}$ at the same time.

$$\begin{cases}\text{if}\sum_{n=1}^{N}u_{m,n}=1,\boldsymbol{q}_m\in\mathbf{q}_{\mathrm{C},m}\ ,\\\text{if}\sum_{n=1}^{N}v_{m,n}=1,\boldsymbol{q}_m\in\mathbf{q}_{\mathrm{J},m}\ .\end{cases}\forall m\in\mathcal{M}\tag{16}$$

If the trajectory of jammer UAVs $\mathbf{Q}_{\mathrm{J}}$ is fixed, the trajectory optimization for collector UAVs can be written as:

$$\begin{aligned}\mathrm{P4}:\max_{\boldsymbol{q}_m\in\mathbf{Q}_{\mathrm{C}}}\ \ &\min\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}R^{\mathrm{sec}}_n,\\\text{s.t.}\ \ &(\text{C5})\text{-}(\text{C8}),\ (\text{C15}).\end{aligned}\tag{17}$$

where (C8) and (C15) are non-convex, and the object function (17) is still non-concave with regard to $\boldsymbol{q}_m$. Such optimization can be solved by convex approximation similar with P3.

❶ Constraint (C8) is non-convex[5]. Recalling that any convex function is global lower-bounded by its first-order Taylor expansion at any given feasible point $\hat{\boldsymbol{q}}_m$. So there is:

$$\begin{aligned}\|\boldsymbol{q}_m-\boldsymbol{q}_{m'}\|^2\ge&\|\hat{\boldsymbol{q}}_m-\boldsymbol{q}_{m'}\|^2+2(\hat{\boldsymbol{q}}_m-\boldsymbol{q}_{m'})^{\mathrm{T}}(\boldsymbol{q}_m-\hat{\boldsymbol{q}}_m)\\&\triangleq f_2^{\mathrm{lb}}(\boldsymbol{q}_m,\hat{\boldsymbol{q}}_m,\boldsymbol{q}_{m'})\ ,\end{aligned}\tag{18}$$

where (C8) is strengthened by its global lower-bound $f_2^{\mathrm{lb}}$ which is an affine function of $\boldsymbol{q}_m$.

❷ The wiretapping rate $R_{n,e_n}$ of the $e_n^{\text{th}}$ EU in (2) is only related to the trajectory of the $m'^{\text{th}}$ jammer UAV via the path-loss $g_{m',e_n}$ in the denominator. So it can be seemed as a constant during the trajectory optimization for collector UAVs.

For briefly, the auxiliary variables $H_n$ and $I$ are introduced. Consequently, $R_{n,m}$ and $R^{\mathrm{sec}}_n$ can be rewritten as:

$$R_{n,m}(\boldsymbol{q}_m)=\log_2(1+\frac{1}{H_n I})\ ,\tag{19}$$

$$R^{\mathrm{sec}}_n(\boldsymbol{q}_m)=\left(\sum_{m=1}^{M}u_{m,n}\log_2(1+\frac{1}{H_n I})-R_{n,e_n}\right)\ ,\tag{20}$$

[3] $-\log_2(\cdot)$ is convex, and any convex function is global lower-bounded by its first-order Taylor expansion.

[4] The superlevel sets of a concave function are convex [26].

[5] Even though $\|\boldsymbol{q}_m-\boldsymbol{q}_{m'}\|^2$ is convex with respect to $\boldsymbol{q}_m$, the resulting set is not a convex set since the superlevel sets of a convex quadratic function is not convex in general [14], [26].

where $H_n(\boldsymbol{q}_m) = \frac{\|\boldsymbol{q}_n-\boldsymbol{q}_m\|^2}{p_n\beta_0}$ and $I(\boldsymbol{q}_m) = \sum_{m'=1,m'\neq m}^{M} v_{m',e_n} \frac{p^{\mathrm{J}}_{m'}\beta_0}{\|\boldsymbol{q}_{m'}-\boldsymbol{q}_m\|^2} + \delta_{\mathrm{L}}^2$.

Since (20) is convex to $H_n$ and $I$, the global lower-bound of the first term can be obtained by its first-order Taylor expansion at any given feasible point $\hat{H}_n$ and $\hat{I}$.

$$\log_2(1+\frac{1}{H_nI}) \geq f_3^{\mathrm{lb}}(H_n, I)\ , \tag{21}$$

where $f_3^{\mathrm{lb}}(H_n, I) \triangleq \log_2(1+\frac{1}{\hat{H}_n\hat{I}})-\left(\frac{H_n-\hat{H}_n}{\ln 2(\hat{H}_n+\hat{H}_n^2\hat{I})}+\frac{I-\hat{I}}{\ln 2(\hat{I}+\hat{I}^2\hat{H}_n)}\right)$.

Notice that $H_n$ and $I$ are convex to $\boldsymbol{q}_m$ because their secondary derivation is always positive. Thus that $f_3$ is a concave function to $\boldsymbol{q}_m$ now.

❸ With regard to (C15), similarly, there is $R_{n,m} = \log_2(1+\frac{1}{H_nI}) \geq f_3^{\mathrm{lb}}(H_n, I)$.

❹ In summary, with any give feasible point $\hat{\boldsymbol{q}}_m$ as well as the lower-bound calculated by (18) and (21), P4 can be approximately transferred to the following convex problem:

$$\begin{aligned} \mathrm{P4.1}: \max_{\boldsymbol{q}_m\in\mathbf{Q}_{\mathrm{C}}} \quad & \min \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} \tilde{R}_n^{\mathrm{sec}}\ , && (22)\\ \text{s.t.} \quad & \text{(C5)-(C7)}\ , \\ & f_2^{\mathrm{lb}}(\boldsymbol{q}_m, \hat{\boldsymbol{q}}_m, \boldsymbol{q}_{m'}) \geq (d^{\min})^2, && \text{(C8.a)}\\ & \sum_{t=1}^{T}\sum_{m=1}^{M} B_\omega u_{m,n} f_3^{\mathrm{lb}}(H_n, I) \geq D_n, \forall n \in \mathcal{N}\ . \\ & && \text{(C15.b)} \end{aligned}$$

where $R_n^{\mathrm{sec}} \geq \left(\sum\limits_{m=1}^{M} u_{m,n} f_3^{\mathrm{lb}}(H_n, I) - R_{n,e_n}\right) \triangleq \tilde{R}_n^{\mathrm{sec}}$.

Therefore, the maximization of $R_n^{\mathrm{sec}}$ can be approximated by maximizing its lower-bound $\tilde{R}_n^{\mathrm{sec}}$, which is a concave function of $\boldsymbol{q}_m \in \mathbf{Q}_{\mathrm{C}}$. Furthermore, (C8.a) is a linear set and (C15.b) is a convex set. Thus P4.1 turns to a convex problem which can be effectively solved by the existing convex toolbox such as CVX.

### *D. Trajectory Optimization for Jammer UAVs*

The trajectory optimization of jammer UAVs can be rewritten as follows with all the other variables fixed:

$$\begin{aligned} \mathrm{P5}: \max_{\boldsymbol{q}_m\in\mathbf{Q}_{\mathrm{J}}} \quad & \min \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} R_n^{\mathrm{sec}}\ , && (23)\\ \text{s.t.} \quad & \text{(C5)-(C8), (C15),} \end{aligned}$$

where (23) is non-concave with regard to $\boldsymbol{q}_{m'}$ and (C8), (C15) are non-convex.

To facilitate the derivation, four auxiliary variables are introduced as follows: $a_n = \frac{p_n g_{n,m}}{\delta_{\mathrm{L}}^2}$, $b_n = \frac{p_n g_{n,e_n}}{\delta_{\mathrm{E}}^2}$, $A_m = \frac{\sum_{m'=1,m'\neq m}^{M} v_{m',e_n} p_{m'}\beta_0}{\|\boldsymbol{q}_{m'}-\boldsymbol{q}_m\|^2\delta_{\mathrm{L}}^2} + 1$ and $B_m = \frac{\sum_{m'=1,m'\neq m}^{M} v_{m',e_n} p_{m'}\beta_0}{(\|\boldsymbol{q}_{m'}-\boldsymbol{q}_{e_n}\|^2)\delta_{\mathrm{E}}^2} + 1$. Notice that $a_n$ is only related to the trajectory of collector UAVs which has been planned by sub-problem P4 above and $b_n$ is not related to any UAV's trajectory, so they can be treated as constants.

On this basis, $R_n^{\mathrm{sec}}$ can be rewritten as:

$$R_n^{\mathrm{sec}} = \left(\sum\nolimits_{m=1}^{M} u_{m,n}\log_2(1+\frac{a_n}{A_m})\right) - \log_2(1+\frac{b_n}{B_m})\ . \tag{24}$$

❶ Since $a_n$ and $b_n$ are non-negative parameters, both $\log_2(1+\frac{a_n}{A_m})$ and $\log_2(1+\frac{b_n}{B_m})$ are convex functions with respect to $A_m$ and $B_m$, respectively. Therefore, the global inequality must hold:

$$\begin{aligned} \log_2(1+\frac{a_n}{A_m}) &\geq f_4^{\mathrm{lb}}(A_m, \hat{A}_m) \\ &\triangleq \log_2(1+\frac{a_n}{\hat{A}_m}) - \frac{a_n(A_m-\hat{A}_m)}{\ln 2(\hat{A}_m^2+a_n\hat{A}_m)}\ , \end{aligned} \tag{25}$$

where $\hat{A}_m$ is any given feasible point of $A_m$. Thus there is:

$$R_n^{\mathrm{sec}} \geq \left(\sum\nolimits_{m=1}^{M} u_{m,n} f_4^{\mathrm{lb}}(A_m, \hat{A}_m)\right) - \log_2(1+\frac{b_n}{B_m}) \triangleq \breve{R}_n^{\mathrm{sec}}\ . \tag{26}$$

Note that $A_m$ is convex to $\boldsymbol{q}_{m'}$ due to the secondary derivation to $\boldsymbol{q}_{m'}$ is always positive. While the second term $-\frac{a_n}{\ln 2(\hat{A}_m^2+a_n\hat{A}_m)}$ of (25) is always negative. Therefore, $f_4^{\mathrm{lb}}(A_m, \hat{A}_m)$ is a concave function to $\boldsymbol{q}_{m'}$.

❷ Similarly, $B_m$ is convex to $\boldsymbol{q}_{m'}$ due to the secondary derivation of $B_m$ is always positive, and $\log_2(1+\frac{b_n}{B_m})$ is convex and monotonic decreasing to $B_m$. Thus that the second term of (26), i.e., $-\log_2(1+\frac{b_n}{B_m})$, is a concave function to $\boldsymbol{q}_{m'}$.

Therefore, the global lower-bound $\breve{R}_{\mathrm{sec}}^n$ in (26) is a concave function to $\boldsymbol{q}_{m'}$ now.

❸ With regards to (C15), similarly, there is $R_{n,m} = \log_2(1+\frac{a_n}{A_M}) \geq f_4^{\mathrm{lb}}(A_m, \hat{A}_m)$.

❹ In summary, with any give feasible point $\hat{\boldsymbol{q}}_{m'}$ as well as the lower-bound expressed by (18) and (26), P5 can be approximately transferred to a convex problem:

$$\begin{aligned} \mathrm{P5.1}: \max_{\boldsymbol{q}_m\in\mathbf{Q}_{\mathrm{J}}} \quad & \min \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} \breve{R}_n^{\mathrm{sec}}\ , && (27)\\ \text{s.t.} \quad & \text{(C5)-(C7)}\ , && (28)\\ & f_2^{\mathrm{lb}}(\boldsymbol{q}_m, \hat{\boldsymbol{q}}_m, \boldsymbol{q}_{m'}) \geq (d^{\min})^2\ , && \text{(C8.a)}\\ & \sum_{t=1}^{T}\sum_{m=1}^{M} B_\omega u_{m,n} f_4^{\mathrm{lb}}(A_m, \hat{A}_m) \geq D_n, \forall n \in \mathcal{N}\ . \\ & && \text{(C15.c)} \end{aligned}$$

Therefore, the maximization of $R_n^{\mathrm{sec}}$ can be approximated by maximizing its lower-bound $\breve{R}_n^{\mathrm{sec}}$ which is a concave function to $\boldsymbol{q}_{m'} \in \mathbf{Q}_{\mathrm{J}}$. Furthermore, (C8.a) is a linear set and (C15.c) is a convex set. Thus P5.1 turns to be a convex problem which can also be solved by CVX.

Details of the overall optimization algorithm are summarized in Algorithm 2.

## IV. Convergence and Complexity Analysis

### *A. Convergence Analysis*

**Lemma 2.** *Algorithm 2 can be converged to a local suboptimal solution at least in finite iterations.*

*Proof.* The original problem P1 is divided into four subproblems and iteratively solved by applying BCD. In specific,

**Algorithm 2:** DRL-SCA for Multi-UAV Assisted Secure Communication with RSS

```
1  Initialization: the maximum tolerance $\epsilon$, random transmit power $\mathbf{P}^0$, random trajectory for each UAV $\mathbf{Q}^0$, the iteration index $l = 0$.
2  repeat
3      Solve Problem P2 by MADRL with any given $\{\mathbf{U}^l, \mathbf{V}^l, \mathbf{P}^l, \mathbf{Q}^l\}$, and output the feasible solution of role arrangement $\mathbf{U}^{l+1}$ and $\mathbf{V}^{l+1}$ according to Algorithm 1 ;
4      Split the trajectory $\mathbf{Q}^l$ into $\mathbf{Q_C}^l$ and $\mathbf{Q_J}^l$ according to (16) based on $\mathbf{U}^{l+1}$ and $\mathbf{V}^{l+1}$;
5      Solving Problem P3.1 by CVX for any given $\{\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \mathbf{P}^l, \mathbf{Q_C}^l, \mathbf{Q_J}^l\}$, and denote the feasible solution of the transmit power of GUs and jammer UAVs as $\mathbf{P}^{l+1}$;
6      Solve Problem P4.1 by CVX with any given $\{\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \mathbf{P}^{l+1}, \mathbf{Q_C}^l, \mathbf{Q_J}^l\}$, and denote the feasible solution of the trajectory of collector UAVs as $\mathbf{Q_C}^{l+1}$;
7      Solve Problem P5.1 by CVX with any given $\{\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \mathbf{P}^{l+1}, \mathbf{Q_C}^{l+1}, \mathbf{Q_J}^l\}$, and denote the feasible solution of the trajectory of jammer UAVs as $\mathbf{Q_J}^{l+1}$;
8      $l \leftarrow l + 1$;
9  until The fractional increasement of the objective value is small enough;
10 Output: $\mathbf{P}, \mathbf{Q}, \mathbf{U}, \mathbf{V}$
```

problems P2, P3.1, P4.1 and P5.1 are alternatively optimized to obtain the suboptimal solution with the initial feasible points. The obtained solution in each iteration is used as the input feasible points for the next iteration.

Let $\eta(\mathbf{U}^l, \mathbf{V}^l, \tilde{\mathbf{P}}^l, \tilde{\mathbf{Q}}^l)$ be the solution of the original objective function 6 at the $l^{\text{th}}$ iteration.

In Step 3, DDQN will output a better solution $\mathbf{U}^{l+1}$ and $\mathbf{V}^{l+1}$ satisfying:

$$\eta(\mathbf{U}^l, \mathbf{V}^l, \tilde{\mathbf{P}}^l, \tilde{\mathbf{Q}}^l) \leq \eta(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^l, \tilde{\mathbf{Q}}^l) \ . \tag{29}$$

Then, in Step 5, the suboptimal solution for transmission power of both GUs and jammer UAVs $\tilde{\mathbf{P}}^{l+1}$ can be obtained by solving P3.1:

$$\begin{aligned} &\eta(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^l, \tilde{\mathbf{Q}}^l) \overset{\text{(a)}}{=} \eta^{\text{lb}}(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^l, \tilde{\mathbf{Q}}^l) \\ &\overset{\text{(b)}}{\leq} \eta^{\text{lb}}(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^{l+1}, \tilde{\mathbf{Q}}^l) \overset{\text{(c)}}{\leq} \eta(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^{l+1}, \tilde{\mathbf{Q}}^l) \end{aligned} \tag{30}$$

where $(a)$ holds the fact that the first-order Taylor expansions are tight at the optimal; $(b)$ follows that P3.1 can be solved optimally due to its convexity; $(c)$ holds because the optimization objective in (13) is the lower bound of the original objective function. The inequality of (30) induces that problems P2 and P3.1 with regarding to ASR are always non-decreasing after each iteration.

The proof of Step 6 and Step 7 is similar to that of (30), and the result follows:

$$\eta(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^{l+1}, \tilde{\mathbf{Q}}^l) \leq \eta(\mathbf{U}^{l+1}, \mathbf{V}^{l+1}, \tilde{\mathbf{P}}^{l+1}, \tilde{\mathbf{Q}}^{l+1}) \ . \tag{31}$$

Notice that the objective function is non-decreasing after each iteration. Owing to the limitation of constraints, the maximum sum ASR is upper bounded by a finite value. Therefore, Algorithm 2 is guaranteed to converge to at least a local suboptimal solution. □

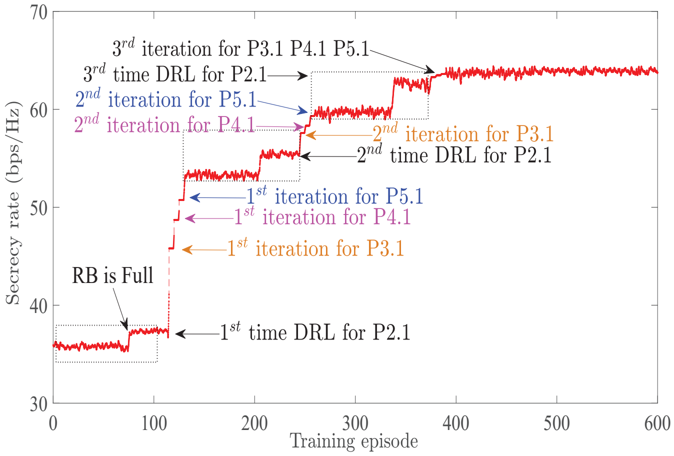


Fig. 4: Algorithm convergence by DRL-SCA.

### B. Complexity of DRL for Role Switching

The computation complexity of DDQN is in terms of the floating point of operations(FLOPs) [27]. As aforementioned, each agent consists of two isomorphic neural networks, i.e., the target network and the on-line network constructed by basic fully-connected multi-layer perceptron (MLP). The neural networks configuration is detailed in TABLE III.

Let $e_i$ be the number of neurons in the $i^{\text{th}}$ layer of the on-line network, where $i \in \{0, \cdots, I\}$ and $I$ is the number of layers. For a fully-connected layer of MLP with $e_i$ neurons as the input and $e_{i+1}$ neurons as the output, the dot product of FLOPs computation from the $i^{\text{th}}$ to the $(i+1)^{\text{th}}$ layer is $(2e_i - 1) \times e_{i+1}$, i.e., multiply $e_i$ times and add $(e_{i+1} - 1)$ times for each neuron.

Let $\kappa$ be the corresponding parameter determined by the type of activation function. For example, the Sigmoid function has $\kappa_{\text{Sigmoid}} = 4$ FLOPs because the function $\delta(z) = 1/1 + e^{-z}$ has four mathematical operations, i.e., division, summation, exponentiation and subtraction, and each of them needs one FLOPs. Similarly, $\kappa_{\text{Relu}} = 1$ FLOPs. Therefore, the computation complexity of DDQN is:

$$\begin{aligned} &2\sum\nolimits_{i=0}^{I-1} ((2e_i - 1)e_{i+1} + \kappa_i e_i) = O(\sum\nolimits_{i=0}^{I-1} e_i e_{i+1}) \\ &= O\left(2e_1 MT + (e_1 + e_2)NT + e_2 ET\right). \end{aligned} \tag{32}$$

### C. Complexity of SCA for Trajectory and Power Optimization

The following problems P3.1, P4.1 and P5.1 solved by SCA are convex optimization [26]. For example, P3.1 involves logarithmic operation which has the complexity of $O((NT + N + T)^{3.5} \log \frac{1}{\epsilon})$ and can be solved in polynomial time, where $(NT+N+T)$ denotes the total number of variables and $\epsilon$ is the solving accuracy.

Similarly, the complexity of P4.1 and P5.1 are both $O((NT + N + MT + T)^{3.5} \log \frac{1}{\epsilon})$. Therefore, the overall computation complexity of SCA for the trajectory and power optimization is:

$$O\left(\left(2(NT + N + MT + T)^{3.5} + (NT + N + T)^{3.5}\right) \log \frac{1}{\epsilon}\right) \ . \tag{33}$$

TABLE III: DDQN neural networks configuration

| Name | Neurons Num. and Active Fun. | Type | Notes |
|---|---|---|---|
| Input layer | $e_0 = (2M+N)T$ for $\mathcal{S}_m^{k\star}$, ReLU | On-line | $\star$ Input is two-tuples $\mathcal{S}_m^k = \{\mathbf{Q}^k, \mathbf{P}^k\}$. Thus that the input dimension is $e_0 = (2M+N)T$.<br>$\dagger$ Output is $\mathcal{A}_m^k = \{\mathbf{u}_m^k, \mathbf{v}_m^k\}$, thus $e_3 = (E+N)T$. |
| Hidden Layer | 2 layers with $e_{a,1} = e_{a,2} = 256$, ReLU6 | | |
| Output layer | $e_{a,3} = (E+N)T$ for $\mathcal{A}_m^{k\dagger}$, Sigmoid | | |

TABLE IV: Simulation parameters

| Parameter | Notation | Simulation value | Parameter | Notation | Simulation value |
|---|---|---|---|---|---|
| Max speed of UAVs | $v^{\max}$ | 50m/s | Average transmit power of UAVs | $P_{\mathrm{UAV}}^{\mathrm{ave}}$ | 25dBm |
| Safe distance among UAVs | $d^{\min}$ | 5m | Peak transmit power of UAVs | $P_{\mathrm{UAV}}^{\max}$ | 30dBm |
| Number of time slots | $T$ | 50 | Noise power | $\delta_L^2$ | -110dBm |
| Time slot duration | $\triangle t$ | 1s | Noise power | $\delta_E^2$ | -110dBm |
| Altitude of UAVs | $h$ | 150m | Data size of GUs | $D_n$ | 1000Mbits |
| Channel power gain | $\beta_0$ | -60dB | Bandwidth | $B_\omega$ | 1MHz |
| Terrestrial pass-loss exponent | $\alpha$ | 3 | Size of replay buffer | RB-size | 80 |
| Average transmit power of GUs | $P_{\mathrm{GU}}^{\mathrm{ave}}$ | 15dBm | Size of mini-batch | $\mathbb{M}$ | 32 |
| Peak transmit power of GUs | $P_{\mathrm{GU}}^{\max}$ | 20dBm | Discount factor | $\gamma$ | 0.95 |

## V. Simulation Results

Supposing that there are $N = 5$ GUs and $E = 5$ EUs deployed in the range of $[1600m \times 1600m]$. While $M = 4$ UAVs cooperatively collect offloading data for GUs and generate AN to EUs during the flight with a fixed initial and final location. As shown in Fig. 5, UAV 1 and UAV 2 start and end at the same position, i.e., $[0, 800]$ and $[0, -800]$ respectively. UAV 3 flies from $[-800, 800]$ to $[-800, -800]$, while UAV 4 starts from $[800, -800]$ and ends at $[800, 800]$. All UAVs fly at the fixed altitude of 150m. The detailed parameters are provided in TABLE IV.

### A. Convergence and Effectiveness

Fig. 4 shows that the value of ASR improves with the iteration and finally converges at a steady value by the proposed DRL-SCA optimization.

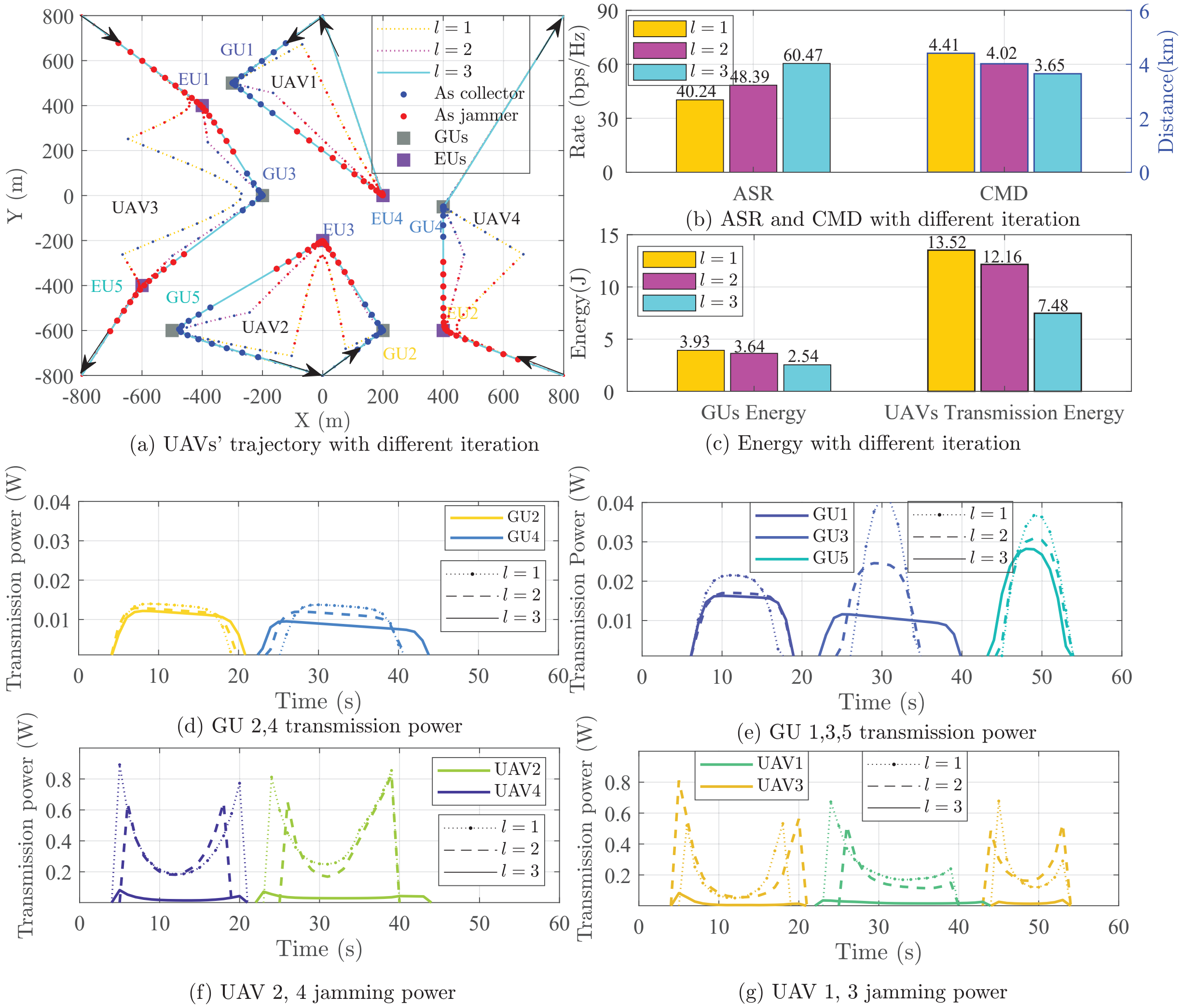
- Once RB is full (Algorithm 1 Line 11), each agent begins to be trained for a better reward with a random selected mini-batch from RB. Along with the training, only the best action can be retained in RB which brings the leap of ASR. Although DDQN only needs to be trained once to handle the following changes of environment, the stage (red line) is specifically retained in each iteration for clarity [6].
- Even though, compared with SCA approach for P3.1, P4.1 and P5.1, the training for role switching is still time-consuming which is an inevitable drawback of reinforcement learning based approaches. It implies that DRL transfers the optimization difficulty to the training of networks by its powerful non-linear capability, but sacrifices the real-time requirements. However, the introduction of convex theory which can be proved to converge with finite iterations, helps to speed up the training.

Fig. 5 details how the system performance is gradually improved along with the number of iterations. For clarity, the trajectory of UAVs by the 1st and 2nd iteration is drawn by yellow and purple dashed lines with small dots respectively, while the trajectory of UAVs by the 3rd iteration is drawn by green solid line with bigger dots.

- Fig. 5 (a) not only depicts the trajectory of UAVs by each iteration, but also distinguishes the role switching during the flight by way points with different colors (blue for collectors and red for jammers). It is worth noticing that when one GU is offloading tasks to a specific collector UAV, the nearest EU will be suppressed by another UAV acting as a jammer. By the cooperation of multiple UAVs, ASR of GUs can be effectively guaranteed.
- Fig. 5 (d)-(g) reveal that the transmission power of GUs and jammer UAVs are also descended step by step. Since the offloading time for multiple GUs would be overlapped sometimes, four sub-figures are used to jointly detail the transmission power of each GU and UAV. For example, when GU 2 is offloading data to UAV 2 during 4s to 20s (shown in (a) and (d)) , UAV 4 is correspondingly acting as the jammer to suppress the wiretapping of EU 2 at the same time (shown in (a) and (f)).
- The statistics results are shown in (b) and (c). Along with the iterations, both collector and jammer UAVs are allowed to fly closer to their objects with less cumulative moving distance (CMD). As a result, not only the energy consumption of GUs and UAVs is reduced, but also ASR is greatly enhanced. Actually, no matter the fixed-wing [28] or rotary-wing [29] UAVs, they have to consume a large proportion of energy for propulsion, which increases approximately linear with the moving distance [30]. So the propulsion energy cost is sensitive to the little variation of CMD [7].

[6] Actually, once the state $\mathcal{S}$ has accomplished the ergodic trajectory during the training, the network of DDQN is able to output a proper action without retraining in the following iterations.

[7] Existing literatures have put forward that the trajectory optimization with the consideration of propulsion energy of UAVs, can be generally approximated as a convex problem solved by SCA with regard to the moving velocity and further transferred to the moving distance by integral transformation [4]–[6], [8], [10].



(a) UAVs' trajectory with different iteration

(b) ASR and CMD with different iteration

(c) Energy with different iteration

(d) GU 2,4 transmission power

(e) GU 1,3,5 transmission power

(f) UAV 2, 4 jamming power

(g) UAV 1, 3 jamming power

Fig. 5: Detailed optimization process along with iterations.

### B. Impact of Flight Duration

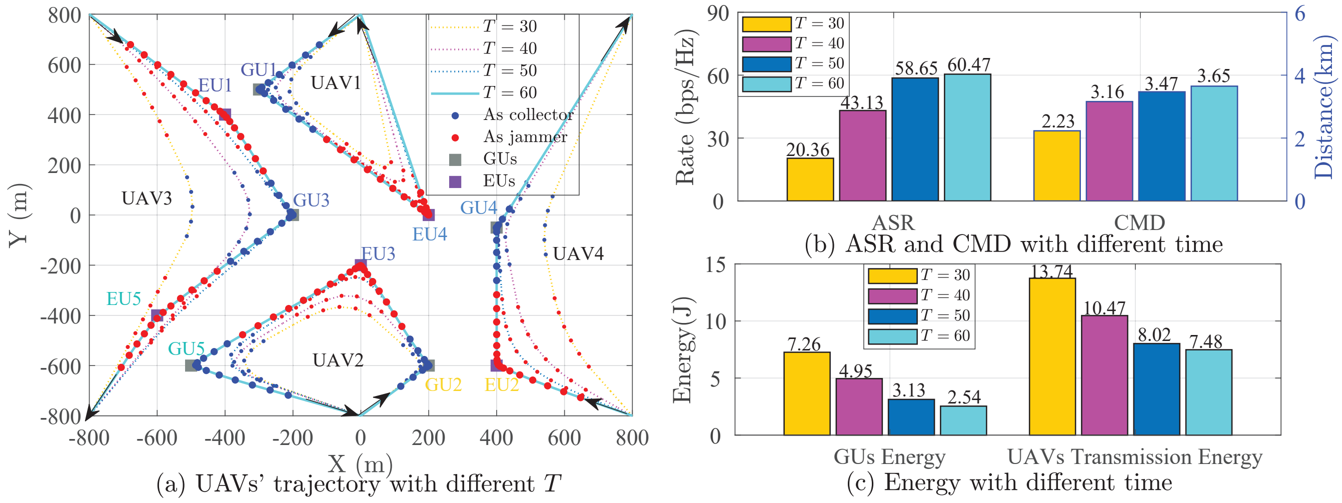
Further simulations shown as Fig. 6 discuss the system performance with different flight duration time $T = 30$, 40, 50 and 60, respectively. Similar with the experiments above, the trajectory of UAVs with $T = 30$, 40 and 50 is marked by dashed lines of different color with small dots, and the trajectory of UAVs with $T = 60$ is drawn by solid line with bigger dots.

- When the flight duration $T$ changes from 30 to 50, there are more time for UAVs to move closer to the target users, no matter GUs for data collecting or EUs for wiretapping suppressing. Although the moving distance of UAVs is slightly increased, the sacrifice is meaningful in exchange for a great improvement of ASR as well as transmission energy consumption of GUs and UAVs.
- However, the system performance gap between $T = 50$ and $T = 60$ is insignificant. It can be inferred that when the flight duration is large enough for obtaining the optimum solution, further increment in $T$ can not bring benefits any more because UAVs have to hover more time idly which will cause extra but unnecessary propulsion energy. So the number of time slots $T$ is set as 50 in the following simulations unless otherwise stated.

### C. Impact of Transmission Parameters

Transmission parameters also play an important role on the average secure rate. Since the average transmission power is linear with the maximum value, the results shown in Fig. 7 only exhibit the system performance along with the maximum transmission power of GUs and UAVs.

- It is easy to interpret that when GUs are allowed to enhance the threshold of maximum transmission power, the legitimate channel gain i.e., $g_{n,m}$ in (1) will be augmented which leads to better ASR.
- At the first glance, eavesdroppers can be more effectively suppressed when jammer UAVs generate AN with larger transmission power according to (2). However, more interference will be induced correspondingly which finally aggravates ASR instead. So parameters related to the transmission power are set as $P_{\text{GU}}^{\text{ave}} = 15\text{dBm}$, $P_{\text{GU}}^{\max} = 20\text{dBm}$, $P_{\text{UAV}}^{\text{ave}} = 25\text{dBm}$ and $P_{\text{UAV}}^{\max} = 30\text{dBm}$ unless otherwise stated.



(a) UAVs' trajectory with different $T$

(b) ASR and CMD with different time

(c) Energy with different time

Fig. 6: Impact of flight duration $T$.

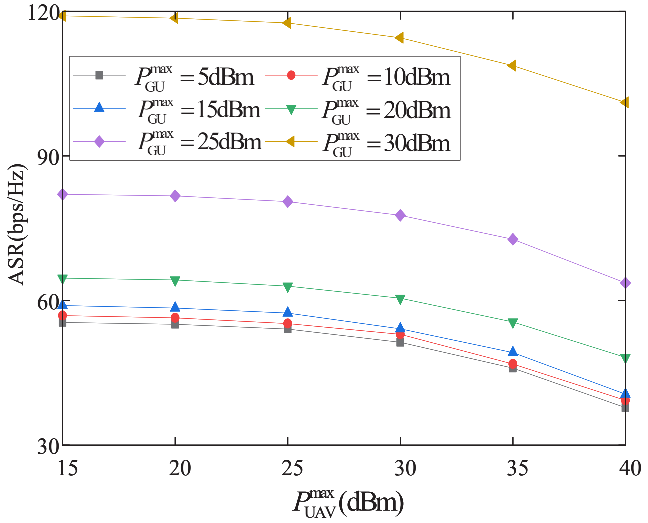


Fig. 7: Impact of the maximum transmission power.

### D. Algorithms Comparison

To demonstrate the superiority of the proposed DRL-SCA RSS algorithm, two more algorithms are considered for comparison.

- **DRL-SCA RSS**, the algorithm proposed in the paper.
- **DRL-SCA RFS**, where once the role of one specific UAV is assigned by DRL, it has to be stayed unchanged during the whole flight.
- **Binary-Relax RSS**, where the binary variables **U** and **V** are relaxed into the continuous range and then the whole optimization problem can be solved by SCA iteration [5], [6]. However, such approach may lead the solution dropping into local minimum.

A simple scenario with 2 UAVs and 4 GU-EU pairs is under the consideration of Fig. 8.

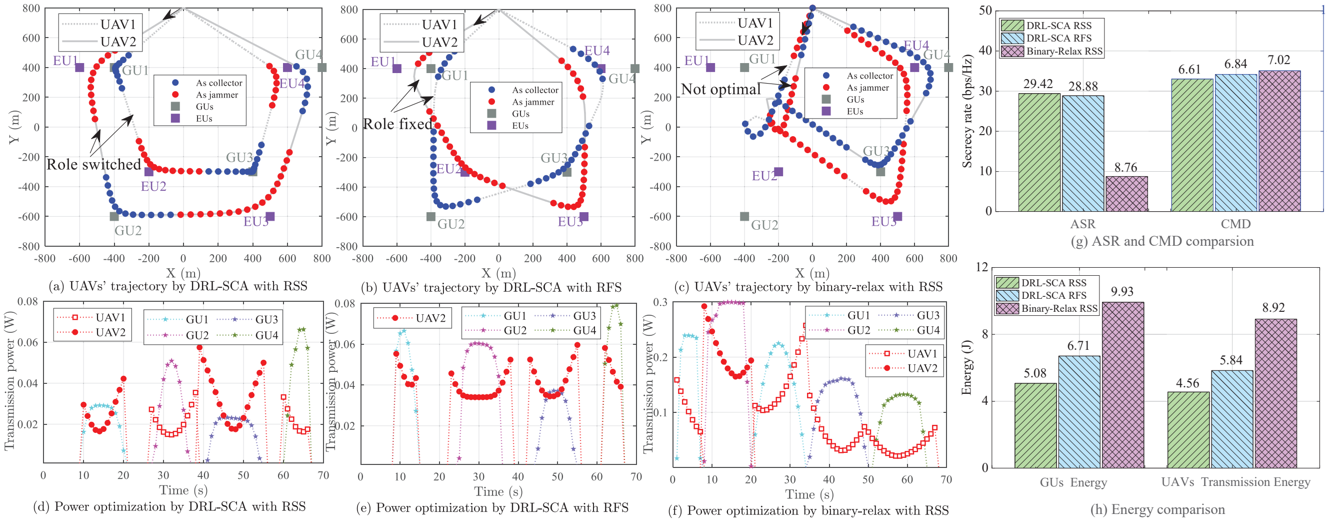
- Since each UAV is assigned with a fixed role by DRL-SCA RFS, it is impossible to ensure that one specific collector UAV is always closer to GUs than EUs in pairs (there exists the similar circumstance for jammer UAVs). For example, when UAV 1 finishes serving GU 1 as a collector in (b), it is actually closer to EU 2 than GU 2. However, it still has to fly more distance to GU 2 as a collector than switch the role to a jammer to suppress EU 2 instead. As a result, DRL-SCA RFS suffers worse ASR of GUs and longer CMD of UAVs with more energy consumption.
- Although role switching is still available for UAVs by Binary-Relax RSS in (c), they can not fly close enough to targets. Sometimes, the trajectory even shifts to other adjacent GUs/EUs. As a result, even though there is only a slight increment of CMD of UAVs, GUs and UAVs have to consume much more energy compared with the former two approaches, and ASR deteriorates irretrievably to a much smaller value.
- A more complicated scenario involving 4 UAVs and 5 GU-EU pairs is considered in Fig. 9, where the superiority of DRL-SCA RSS is further expanded compared with the other two approaches. However, it is worth to notice that the performance gap between DRL-SCA RFS and Binary-Relax RSS is reduced instead, which implies that RSS will be more effective along with the increasing complexity of system.

## VI. Discussion and Future Work

Since the paper focuses on the joint optimization of trajectory and power control with dynamic role switching scheme, we would like to restrict the current contributions to a more concise case, where the communication channels are dominated by LoS links without obstacles, and each entity is equipped with a single antenna and deployed in 2D plane.

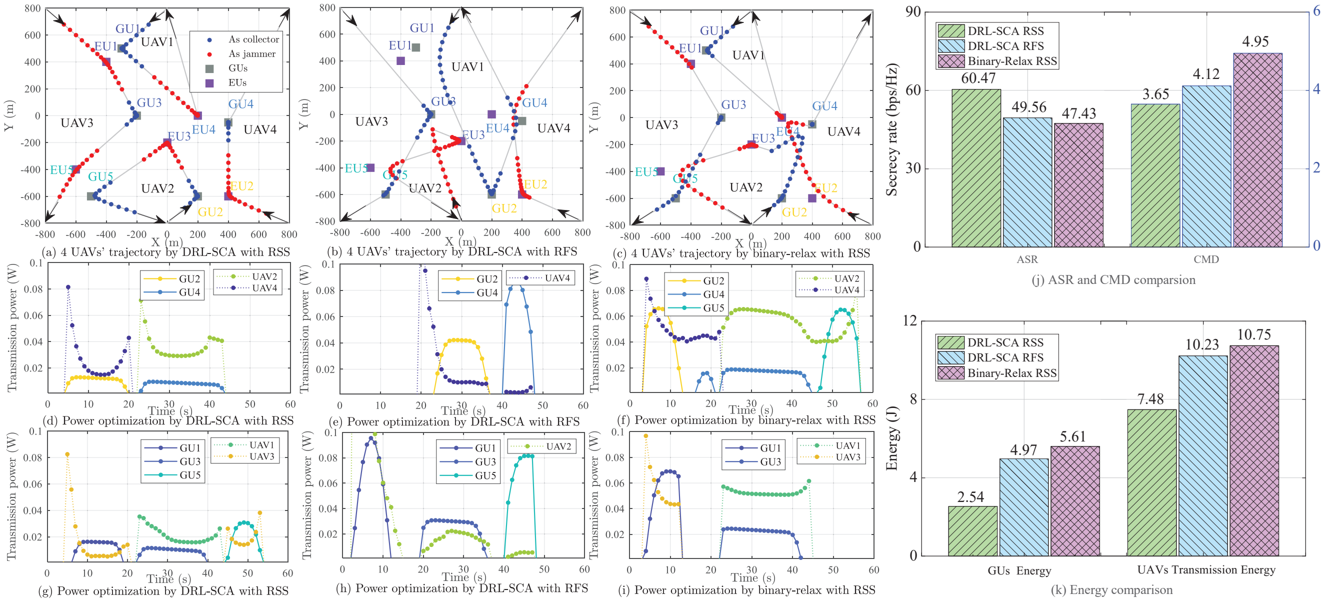
### A. Deployment

It should be noticed that the proposed DRL-SCA algorithm is deployed in centralize. All legitimate GUs and UAVs are connected to a central controller (such as a base station or a leader UAV) through an interference-free command channel.



(a) UAVs' trajectory by DRL-SCA with RSS (b) UAVs' trajectory by DRL-SCA with RFS (c) UAVs' trajectory by binary-relax with RSS (g) ASR and CMD comparsion (d) Power optimization by DRL-SCA with RSS (e) Power optimization by DRL-SCA with RFS (f) Power optimization by binary-relax with RSS (h) Energy comparison

Fig. 8: Algorithms comparison in a simple scenario (2 UAVs for 4 GU-EU pairs) .



(a) 4 UAVs' trajectory by DRL-SCA with RSS (b) 4 UAVs' trajectory by DRL-SCA with RFS (c) 4 UAVs' trajectory by binary-relax with RSS (j) ASR and CMD comparsion (d) Power optimization by DRL-SCA with RSS (e) Power optimization by DRL-SCA with RFS (f) Power optimization by binary-relax with RSS (g) Power optimization by DRL-SCA with RSS (h) Power optimization by DRL-SCA with RFS (i) Power optimization by binary-relax with RSS (k) Energy comparison

Fig. 9: Algorithms comparison in a more complicated scenario (4 UAVs for 5 GU-EU pairs).

Therefore, the position and task size of GUs, as well as the initial and final position of UAVs can be easily obtained.

Besides, the position of EUs can be aware by detecting the leaked oscillator power [31] or other alternative techniques such as cameras [32], synthetic aperture radar [33] and RF sensing [34] equipped on-board, and then reported to the central controller. On the basis, the central controller can construct a virtual environment to play the training by itself, and the reward related to the secure communication in the worst case defined in (5) can be figured out.

Once well trained, DDQN can output a proper action (i.e., the role assignment $\mathbf{U}, \mathbf{V}$ for all UAVs) with the given global observation $\{\mathbf{Q}, \mathbf{P}\}$. In other words, DDQN only needs to be trained once, and then can handle the following changes of environment (including the trajectory of UAVs and transmission power of GUs) according to the iteration of Step 5, 6, 7 in Algorithm 2.

### B. *Future Work*

Quite a number of literatures have assumed the free space path-loss model to simplify the mathematical analysis of UAV-assisted communication networks. However, there still exists the possibility of NLoS links caused by the terrain and aircraft structure reflection. Such mobility related uncertainty may lead to the performance deterioration, and block the straightforward usage of the proposed DRL-SCA algorithm. It is still an open issue to consider the composite channel model and perform optimization using only the large-scale CSI for UAV communications [35].

In general, it will be more effective for UAVs to work in full-duplex mode, i.e., the UAV acting as a full-time jammer can serve as a collector synchronously. The scheme will be possible if UAVs are equipped with multi-antenna to compensate the self-interference by active beamforming. However, it is unpractical for small size UAVs to accommodate antenna array with large inter-element distance. Smaller carrier wavelength such as mmWave can make it feasible but at the expense of higher path loss and power consumption [36]. Even though, some studies from the aspect of beam pattern optimization have exploited MIMO in the airborne environment [16], [17], [35], [37], which will be an interesting topic and one of our further work.

Actually, no matter DDQN proposed in the manuscript, or other descendant DRL approaches such as PPO [38], DDPG [39], MADDPG [40] and TD3 [41] are capable of solving such complicated optimized problems. However, there is an inevitable drawback of such DRL based approaches that it is time-consuming to train the neural networks. Another common concern for machine learning is the lack of interpretation for the results [42], which prevents them to be completely trusted. The combination of DRL with other theories, such as convex [43], matching [44] and game theory [20] to pursue the near closed-form solution with low-complexity, may be a feasible way to balance the environment uncertainty and the time-consuming training. Moreover, inspired by the multi-actor-attention-critic (MAAC) scheme [45], self-attention mechanism can be integrated into the neural networks to interpret the agents' interaction in complex environments.

## VII. Conclusion

A novel role switching scheme is proposed in the paper for secure transmission in UAV-assisted communication networks by the cooperation of multiple UAVs, where UAVs can dynamically switch their roles either as aerial collectors to collect offloading tasks of GUs or as friendly jammers to suppress the potential wiretapping of malicious EUs. The maximization of achievable secrecy rate of GUs can be formulated as a non-convex MINLP problem by jointly optimizing the trajectory and power control of GUs and UAVs, which is hard to be solved in general.

A DRL combined SCA algorithm is further designed to tackle such non-trivial issue. In specific, each UAV works as an individual agent to learn the role switching during the flight. While the trajectory and power control can be sequentially solved by SCA. The results demonstrate that the introduction of convex theory can converge with finite iterations and help to speed up the training. Due to the capacity of exploring better trajectory and avoiding dropping into local minimum, the proposed DRL-SCA RSS is superior to the other two approaches, i.e., DRL-SCA RFS and Binary-Relax RSS in achieving better ASR with less energy consumption of GUs and UAVs, especially when there are more UAVs and GUs involving in the system.

## References

[1] X. Sun, D. W. K. Ng, Z. Ding, Y. Xu, and Z. Zhong, "Physical layer security in uav systems: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 40–47, Oct. 2019.

[2] Q. Wu, W. Mei, and R. Zhang, "Safeguarding wireless network with uavs: A physical layer security perspective," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 12–18, Oct. 2019.

[3] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing uav communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb. 2019.

[4] M. Cui, G. Zhang, Q. Wu, and D. W. K. Ng, "Robust trajectory and transmit power design for secure uav communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 9042–9046, Sep. 2018.

[5] R. Zhang, X. Pang, W. Lu, N. Zhao, Y. Chen, and D. Niyato, "Dual-uav enabled secure data collection with propulsion limitation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7445–7459, Nov. 2021.

[6] R. Zhang, X. Pang, W. Lu, N. Zhao, M. Liu, Y. Chen, and D. Niyato, "Cooperative uav-assisted secure uplink communications with propulsion power limitation," in *IEEE Int. Conf. Commun.(ICC), Montreal, Canada*, Jun. 2021.

[7] Y. Cai, F. Cui, Q. Shi, M. Zhao, and G. Y. Li, "Dual-uav-enabled secure communications: Joint trajectory design and user scheduling," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1972–1985, Sep. 2018.

[8] C. Zhong, J. Yao, and J. Xu, "Secure uav communication with cooperative jamming and trajectory control," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 286–289, Feb. 2019.

[9] A. Li, Q. Wu, and R. Zhang, "Uav-enabled cooperative jamming for improving secrecy of ground wiretap channel," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 181–184, Feb. 2019.

[10] M. Li, X. Tao, N. Li, H. Wu, and J. Xu, "Secrecy energy efficiency maximization in uav-enabled wireless sensor networks without eavesdropper's csi," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3346–3358, Jul. 2022.

[11] K. Xu, M.-M. Zhao, Y. Cai, and L. Hanzo, "Low-complexity joint power allocation and trajectory design for uav-enabled secure communications with power splitting," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1896–1911, Mar. 2021.

[12] Y. Chen and Z. Zhang, "Uav-aided secure transmission in misome wiretap channels with imperfect csi," *IEEE Access*, vol. 7, pp. 98 107–98 121, Jul. 2019.

[13] M. Hua, Y. Wang, Q. Wu, H. Dai, Y. Huang, and L. Yang, "Energy-efficient cooperative secure transmission in multi-uav-enabled wireless networks," *IEEE Trans. Veh. Technol*, vol. 68, no. 8, pp. 7761–7775, Aug. 2019.

[14] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-uav enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[15] S. Sun, G. Zhang, H. Mei, K. Wang, and K. Yang, "Optimizing multi-uav deployment in 3-d space to minimize task completion time in uav-enabled mobile edge computing systems," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 579–583, Feb. 2021.

[16] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "Uav-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 599–11 611, Oct. 2020.

[17] Y. Zhang, Z. Zhuang, F. Gao, J. Wang, and Z. Han, "Multi-agent deep reinforcement learning for secure uav communications," in *IEEE Wireless Commun. Networking Conf.(WCNC), South Korea,*, May 2020.

[18] C. Wen, Y. Fang, and L. Qiu, "Securing uav communication based on multi-agent deep reinforcement learning in the presence of smart uav eavesdropper," in *IEEE Wireless Commun. Networking Conf.(WCNC), Austin, USA*, Apr. 2022.

[19] H. Kang, X. Chang, J. Mišić, V. B. Mišić, J. Fan, and J. Bai, "Improving dual-uav aided ground-uav bi-directional communication security: Joint uav trajectory and transmit power optimization," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10 570–10 583, Oct. 2022.

[20] A. Gao, Q. Wang, W. Liang, and Z. Ding, "Game combined multi-agent reinforcement learning approach for uav assisted offloading," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12 888–12 901, Dec. 2021.

[21] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. Elkashlan, B. Vucetic, and Y. Li, "Secure communications for uav-enabled mobile edge computing systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 376–388, Jan. 2020.

[22] 3GPP, Technical Report, TR 36.777, v.15.0.0, "Technical specification group radio access network: Study on enhanced lte support for aerial vehicles," 2017.

[23] Qualcomm Technologies, Inc., "Lte unmanned aircraft systems," San Diego, CA, USA, Trial report v.1.0.1, 2017.

[24] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P. E. Wang, "The sky is not the limit: Lte for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 204–210, Apr. 2018.

[25] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *AAAI Conf. Artif. Intell.(AAAI-16), Phoenix, Arizona*, Feb. 2016.

[26] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[27] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (ddpg)-based energy harvesting wireless communications," *IEEE Internet of Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.

[28] Y. Zeng and R. Zhang, "Energy-efficient uav communication with trajectory optimization," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[29] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[30] 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 J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[31] A. Mukherjee and A. L. Swindlehurst, "Detecting passive eavesdroppers in the mimo wiretap channel," in *IEEE Int. Conf. Acoust Speech Signal Process (ICASSP)*, Kyoto, Japan, Mar. 2012.

[32] S. Minaeian, J. Liu, and Y.-J. Son, "Vision-based target detection and localization via a team of cooperative uav and ugvs," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 46, no. 7, pp. 1005–1016, Nov. 2016.

[33] I. Guvenc, F. Koohifar, S. Singh, M. L. Sichitiu, and D. Matolak, "Detection, tracking, and interdiction for amateur drones," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 75–81, Apr. 2018.

[34] M. Ezuma, F. Erden, C. Kumar Anjinappa, O. Ozdemir, and I. Guvenc, "Detection and classification of uavs using rf fingerprints in the presence of wi-fi and bluetooth interference," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 60–76, Nov. 2020.

[35] W. Feng, J. Wang, Y. Chen, X. Wang, N. Ge, and J. Lu, "Uav-aided mimo communications for 5g internet of things," *IEEE Internet of Things J.*, vol. 6, no. 2, pp. 1731–1740, Oct. 2019.

[36] A. A. Khuwaja, Y. Chen, N. Zhao, M.-S. Alouini, and P. Dobbins, "A survey of channel modeling for uav communications," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 4, pp. 2804–2821, Jul. 2018.

[37] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted uav communication: Joint trajectory design and passive beamforming," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, May 2020.

[38] S. John, W. Filip, D. Prafulla, R. Alec, and K. Oleg, "Proximal policy optimization algorithms," in *Proc. Mach. Learn. Res. (PMLR),Scotland, U.K.,*, Jun. 2017.

[39] P. L. Timothy, J. H. Jonathan, P. Alexander, H. Nicolas, E. Tom, T. Yuval, S. David, and W. Daan, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR), Puerto Rio,*, May 2016.

[40] L. Ryan, W. Yi, T. Aviv, H. Jean, A. Pieter, and M. Igor, "Multiagent actor-critic for mixed cooperative-competitive environments," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA,*, Dec. 2017.

[41] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Mach. Learn. Res. (PMLR), Stockholm, Sweden,*, Jun. 2018.

[42] C. Molnar, *Interpretable Machine Learning*, Aug. 2021, https://christophm.github.io/interpretable-ml-book/.

[43] A. Gao, Z. Shao, Y. Hu, and W. Liang, "Joint trajectory and energy efficiency optimization for multi-uav assisted offloading," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 508–511.

[44] K. Chen, A. Gao, W. Duan, and W. Liang, "Matching combined multi-agent reinforcement learning for uav secure data dissemination," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 3367–3370.

[45] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn (ICML)*, Jun. Long Beach, 2019.

**Ang Gao** received his Ph.D. degree in control theory and control engineering from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2011. He currently serves as an Associate Professor at the School of Electronics and Information, Northwestern Polytechnical University. His research interests include QoS control, resource management and deep reinforcement learning in wireless communication networks.

**Qinyu Wang** is currently a master student under the supervision of Prof. A. Gao with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. Her research interests contain convex optimization and deep reinforcement learning in uav-assisted wireless communication networks for resource allocation.

**Yansu Hu** received her Ph.D. degree in control theory and control engineering from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2012. She currently serves as an Associate Professor at the School of Electronics and Control, Chang'an University. Her research interests include networked control and resource allocation in cloud computing.

**Wei Liang** received her Ph.D. degree in wireless communication at University of Southampton, Southampton, U.K, in 2015. She was a Postdoctoral Research Fellow in Lancaster University from 2015 to 2018. She currently serves as an Associate Professor at the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include adaptive coded modulation, network coding, matching theory, game theory, non-orthogonal multiple access and machine learning.

**Jiangkang Zhang** (IEEE Senior Member) is a Senior Lecturer at Bournemouth University. Prior to joining in Bournemouth University, he was a senior research fellow at University of Southampton, UK. Dr Zhang was a lecturer from 2012 to 2013 and then an associate professor from 2013 to 2014 at Zhengzhou University. His research interests are in the areas of aeronautical communications and networks, evolutionary algorithms, machine learning algorithms and edge computing. He serves as an Associate Editor for IEEE ACCESS.