Learning to Navigate for Secure UAV Communication

Xiangyu Zhang*†, Shu Xu*, Luxi Yang*†
* the National Mobile Communications Research Laboratory, School of Information Science and Engineering
Southeast University, Nanjing, China
†Purple Mountain Laboratories, Nanjing, China

***Abstract*—In this paper, we investigate the navigation for unmanned aerial vehicle (UAV)s in the secure communication system, where we design the UAV's navigation/trajectory to ensure the Quality of Service (QoS) with the Base Station (BS) in the existence of multiple unknown-location dynamical eavesdroppers and jammers. To this end, we formulate a UAV trajectory optimization problem to minimize its mission completion time with QoS and security constraints. The imperfect information, dynamic communication environment, and non-convexity make the problem intractable. For these reasons, we propose a novel solution approach, namely Model-Assisted Reinforcement Learning (MARL) algorithm, where the communication system model is embedded into Deep Reinforcement Learning (DRL) framework to ensure secure communication and shorten the learning process. Numerical results show that our proposed methodology can safeguard security and find the shortest way to finish the mission.**

I. Introduction

As a vital part of IoT, the unmanned aerial vehicle (UAV) creates a new paradigm shift that can provide more swift and flexible, and high-speed network service for 5G and Beyond-5G in the sky since the characteristic of fully controllable UAV mobility and high probability to have strong short-distance Line-of-Sight (LoS) communication links [1], [2]. However, UAV communication faces the danger of being eavesdropped on and interfered with because of the open nature of air-to-ground wireless channels [3]. The common way to ensure communication security in the wireless communication systems is by adopting cryptographic mechanisms and encoding mechanisms against unauthorized access and intentional interference. Nevertheless, the management of the cryptographic system and encoding system is complex. A novel method to achieve better performance is the exploration of physical-layer security, which leverages the intrinsic characteristics of wireless channels to establish a more robust channel with the BS and degrade the eavesdroppers and jammers' channels [4].

In this paper, we investigate the navigation problem of the UAV to elude the eavesdroppers while keeping communication with the ground Base Station(BS), which can be easily implemented in real scenarios. Some works have proved the feasibility and effectiveness of the enhancement of the physical-layer security in the UAV communication by optimizing the trajectory. The paper [5] ensures security via a new trajectory design by utilizing the block coordinate descent and successive convex optimization methods. Similarly, the paper [6] jointly optimizes the trajectory and transmit power of UAVs to maximize the average worst-case secrecy rate. Besides, utilizing the UAVs as jammers to interfere with the eavesdroppers can effectively protect the communication links from being eavesdropped by ground nodes. Some researches protect the UAV from being eavesdropped by using the friendly jammer UAV to interfere the eavesdropper [7]–[11]. However, the jammer UAV also causes interference to the valid users because of the strong LoS channel between the friendly jammers and valid users. Certainly, these works have concentrated on the security UAV communication problem, but it still has two major challenges. First, the eavesdroppers' information or the opposing jammers are assumed to be available previously, which is not realistic. Second, their results are based on the assumption that the eavesdroppers or opposing jammers are static or using fixed interference strategies, leading to failure when facing dynamic scenarios. To sum up, the requirement of perfect environment information makes the traditional method hard to solve the problem in reality.

The progress of deep reinforcement learning (DRL) algorithm provides a "end-2-end" method for planning the UAV's trajectory to enhance the performance of UAV communication networks [12], [13]. The work [14] proposes a multi-agent DRL algorithm in the UAV security communication scenario, where a cooperative jamming UAV helps the UAV transmitter defend against eavesdroppers. However, the security constraint is hard to ensure by DRL in the training processing, which is the main challenge in the application of DRL in UAV security. The main reason root in the essential of DRL algorithms is that the DRL algorithm plan the trajectory in numerous errors, but the error could not be allowed due to security constraints.

Solving the above challenges is indeed the motivation of this work. Specifically, we propose a new approach named model-assisted reinforcement learning algorithm that is based on the Deep Reinforcement Learning (DRL) [15] architecture and utilizes the communication systems model to assist the RL algorithm in planning the trajectory. Our main innovation contains two aspects. The first is that we formulate the trajectory planning problem of security UAV communication problem in dynamic environment as an Markov Decision Process (MDP) and solve it by the MARL algorithm, which overcomes the

This work was supported by the National Natural Science Foundation of China under Grants 61971128 and U1936201, and the National Key Research and Development Program of China under Grant 2020YFB1804901.

above two critical issues in conventional studies. The other is that we embed the communication system model into the RL architecture, namely, the Model Assist Learning (MAL), to enforce the solution satisfies the constraints. Meanwhile, we prove that the MAL also improves performance and shortens the learning processing significantly.

We formulate the UAV trajectory optimization problem to minimize mission complete time while ensuring the security and QoS constraint. And then, we proposed the Model-Assisted Reinforcement Learning (MARL) method to solve the problem. The MARL contains two parts, the RL part and the model-assisted learning (MAL). In the RL part, we formulate the problem as an MDP and employ a state-of-the-art algorithm, proximal policy optimization (PPO) [16], which is the data-driven method. Moreover, in the MAL part, we treat the problem as an optimization problem and utilize the communication system model to train the policy directly, which is the model-driven method. Compared to such existing path/trajectory designs, the assistance of the communication system model will keep the UAV in a safe state and reduce the useless exploration for shortening the learning process. Numerical results verify the performance of the proposed algorithm. It is observed that our proposed method can keep a higher security rate and less mission time than any other traditional algorithms.

The rest of the paper is organized as follows. We introduce the system model and problem formulation in Section II. Section III presents the proposed algorithms, including the PPO algorithm for UAV trajectory planning. Section IV presents the numerical results, and finally, we conclude the paper in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Description

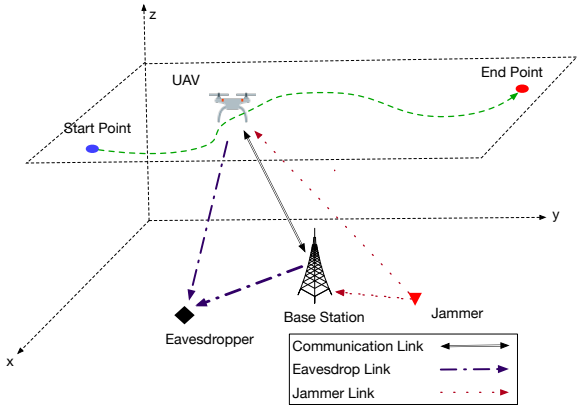


Fig. 1. An UAV MIMO communication system with multiple single-antenna jammers and eavesdroppers, multiple potential eavesdroppers.

As shown in Fig. 1, we consider a UAV-based MIMO communication system, which consists of a multi-antenna BS, multi-antenna UAV, K unknown-location eavesdroppers with a single antenna, and M jammers. In this case, our main objective is to design the trajectory of the UAV to fly to the aim point without being tapped by eavesdroppers, while maintaining connectivity with the BS. We assume that the UAV, the BS, and all the eavesdroppers share the same frequency band while communicating. Furthermore, the location of eavesdroppers can be estimated by the UAV via using accessory equipment, such as radar, imaging, or infrared sensing system.

B. Mobility Model

Let T denote the task's total completion time, where the UAV flies from the initial location $\mathbf{q}_I$ to the destination $\mathbf{q}_F$. Then, we discretize the time duration T with fixed time intervals δ_t into $N = \frac{T}{\delta_t}$ time slots. For ease of expression, we formulate the model in a three-dimensional (3D) Cartesian coordinate system. We assume that the BS, jammers, and eavesdroppers are located on the ground, the UAV flies at a constant altitude H_U. The horizontal trajectory of UAV at time n can be defined as $\mathbf{q}_U[n] = [x_U[n], y_U[n]]^T \in \mathbb{R}^{2\times 1}$. Then, we have $\mathbf{q}_U[0] = \mathbf{q}_I$ and $\mathbf{q}_U[N] = \mathbf{q}_F$ Thus, the velocity of the UAV can be expressed as

$$\mathbf{v}_U[n] = \mathbf{q}_U[n+1] - \mathbf{q}_U[n], n = 0, 1, ..., N-1. \quad (1)$$

$$\|\mathbf{v}_U[n]\|_2 \leq V_{\max}, n = 0, 1, ..., N-1. \quad (2)$$

where V_{max} denotes the maximum speed of the UAV.

We assume the eavesdroppers are movable ground equipment and randomly patrol with an unknown strategy to search for the UAV. The trajectory of eavesdropper k can be denoted as $\mathbf{q}_k^E[n] = \left[x_k^E[n], y_k^E[n]\right]^T \in \mathbb{R}^{2\times 1}$. Besides, we assume that the BS and jammers are fiexd during N time slots. Then, the location of aiming point, BS and jammer m can be denoted as $\mathbf{q}_A = [x_A, y_A, H_U]$ $\mathbf{q}_B = [x_B, y_B, H_B]$, and $\mathbf{q}_m^J = \left[x_m^J, y_m^J, H_m^J\right]$, respectively.

C. Channel Model

Consider the transmission scenario in an urban area, where a Probabilistic LoS Channel Model is applied to characterize the Air-to-Ground communication, i.e., the communication link between the UAV and the BS, well as the UAV and eavesdroppers [17]. The expected channel power gain from the BS to the UAV at time n is given by

$$g_{UB}[n] = \left[P_{LoS} + (1 - P_{LoS})\kappa\right]^{-1} \beta_0 d_{UB}^{-\alpha}[n], \quad (3)$$

where $d_{UB}[n] = \|\mathbf{q}_U[n] - \mathbf{q}_B\|_2$ denotes the distance between the UAV and the BS, the factor κ is the attenuation coefficient due to the NLoS link, the factor $\beta_0 = \left(\frac{\lambda}{4\pi}\right)^2$ is the channel power gain at the reference distance of 1m, λ is the carrier wavelength, and P_{LoS} denotes the probability of having LoS link, which is the function of the elevation angle $\boldsymbol{\theta}$ as

$$P_{LoS} = \frac{1}{1 + a\exp\left(-b\left(\sin^{-1}\left(\frac{H_U - H_B}{d_{UB}[n]}\right) - a\right)\right)}, \quad (4)$$

where a and b are modeling parameters. A similar expression can be written to calculate the expected channel power gain $g_k^{UE}[n]$ from the UAV to the BS.

Additionally, we only consider the large-scale fading channel model for the Ground-to-Ground communication model. Thus, the expected channel power gain for the BS to eavesdroppers can be expressed as

$$g_{BE}[n] = \beta_0 d_{BE}^{-\alpha}[n], \tag{5}$$

D. Antenna Model

In this paper, we assume that the UAV and the BS is equipped with a uniform linear array (ULA) with M_U and M_B antenna elements, respectively. The antenna array with the Angel of Departure (AoD) can be denoted as

$$\mathbf{a} = \left[1, e^{j\Psi}, e^{j2\Psi}, ..., e^{j(N-1)\Psi}\right] \tag{6}$$

where $\Psi = \frac{2\pi}{\lambda} d\cos\boldsymbol{\theta}$ denotes the wave-number variable, λ is the wavelength of the transmitted signals, and d is the spacing between the antenna elements. Thus, the array factor for such an array can be expressed as

$$\begin{aligned} AF &= 1 + e^{j\Psi} + e^{j2\Psi} + ... + e^{j(N-1)\Psi} \\ &= e^{j\frac{(N-1)}{2}\Psi}\left[\frac{\sin\left(\frac{N}{2}\Psi\right)}{\sin\left(\frac{1}{2}\Psi\right)}\right] \end{aligned} \tag{7}$$

The sensor fusion-based (such as GPS) beam tracking technology is applied for both the UAV and the BS, in order to assist the beam alignment [18]. Therefore, the expected SINR from the UAV to the BS can be expressed as

$$\gamma_{UB}[n] = \frac{AF_{UB}[n] P_U g_{UB}[n]}{I_J[n] + \sigma^2}, \tag{8}$$

where σ^2 denotes the power of the Additive White Gaussian Noise, and $I_J[n]$ is the interference produced by jammers. Similarly, the expected SINR from the UAV to eavesdroppers $\gamma_k^{UE}[n]$, from the BS to the UAV $\gamma_{BU}[n]$, from the BS to eavesdroppers $\gamma_k^{BE}[n]$ can be calculated.

E. Problem Formulation

Our task is to minimize the total time steps for the UAV flies from the given initial location to the final location, while avoiding being disconnected from the cellular network and being detected by eavesdroppers. Thus, the optimization problem can be formulated as

$$(P1): \min_{\{\mathbf{q}_U[n]\}} \{N\} \tag{9}$$

$$s.t.\ \mathbf{q}_U[0] = \mathbf{q}_I \tag{10}$$

$$\mathbf{q}_U[N] = \mathbf{q}_F \tag{11}$$

$$\gamma_{UB}[n] \geq \gamma_{th}, \forall n \tag{12}$$

$$\gamma_{BU}[n] \geq \gamma_{th}, \forall n \tag{13}$$

$$\gamma_k^{UE}[n] < \gamma_{th}, \forall n, k \tag{14}$$

$$\gamma_k^{BE}[n] < \gamma_{th}, \forall n, k \tag{15}$$

$$(1), (2)$$

We assume that the communication link is established when the minimum SINR threshold γ_{th} is satisfied. The optimization problem $P1$ is a highly non-convex problem due to the objective function related to the link's current transmission quality, and eavesdroppers are swapping around dynamically.

In the following, we propose an efficient approach based on the MARL algorithm and DRL framework to solve the UAV trajectory design problem.

III. PROPOSED ALGORITHMS

In this section, we first give a very brief background of the RL algorithm. We then formulate the UAV security communication problem as an MDP process for suiting the RL algorithm. Afterwards, we present our MARL method to solve the trajectory planning problem.

A. Basics of Reinforcement Learning

The RL algorithms primarily target solving the MDP problem, where an agent iteratively interacts with the environment. In each iterative step, the agent observes the environment and chooses an action according to the state to obtain the environment's reward. Mathematically, an MDP can be specified by 4-tuple $< \mathbf{S}, \mathbf{A}, \mathbf{P}, \mathbf{R} >$, where:

- $\mathbf{S}$ is the state set that $\forall s \in \mathbf{S}$, named state space;
- $\mathbf{A}$ is the action set that $\forall a \in \mathbf{A}$, named action space;
- $\mathbf{P}(s, a, s')$ is the state transition probability, that represent the probability that agent step into the state s' when the agent chooses the action a at the state s.
- $\mathbf{R}(s, a, s')$ represents the reward that the agent obtains from the environment when the immediate transform $s \rightarrow a \rightarrow s'$ occur.

The main objective of Reinforcement Learning algorithm is finding a policy $\pi(a|s)$, that gives the probability of taking action $a \in A$ when in state $s \in S$, to maximize the expectation of discounted future reward, which is :

$$\eta(\pi(a|s)) = \mathbb{E}_{s_0, a_0, ...}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right], \tag{16}$$

where γ is the discount factor.

However, in the virtual environment, the policy $\pi(a|s)$ always faces challenges to be formulated due to the environment's complexity. The DRL takes full advantage of the neural network to approximate the policy function $\pi(a|s)$ and trains $\pi(a|s)$ with the interactive experience. Thereby, the MDP problem can be solved by finding a suitable parameter vector $\boldsymbol{\theta}$ of the neural work, which is $\pi_{\boldsymbol{\theta}}(a|s)$.

B. UAV trajectory planning as a MDP

In our problem, we model the UAV as the agent in RL and the BS, jammers, and eavesdroppers as the environment. In each time step n, the UAV moves according to the environment. The problem objective of $P1$ can only be obtained when the mission ends, which is too sparse for RL to solve. Meanwhile, the expression of constraints (12)-(15) is hard to be integrated into the punishment function for MAL. As such, we change the objective into the maximum distance to the aim

point in each step. Meanwhile, by introducing the Lagrange multipliers, the problem $P1$ can be transformed into:

$$(P2): \max \sum_{i=1}^{n} J[i] \tag{17}$$

$$s.t.\ \mathbf{q}_U[0] = \mathbf{q}_I, \mathbf{q}_U[N] = \mathbf{q}_F \tag{18}$$

where the $J[i]$ is the punishment function in each step, which is

$$\begin{aligned} J[i] = &(|\mathbf{q_U}[i] - \mathbf{q}_A| - |\mathbf{q_U}[i-1] - \mathbf{q}_A|) \\ &- \lambda_1 \mathbb{F}_{relu}(\gamma_{th} - \gamma_{UB}[i]) - \lambda_2 \mathbb{F}_{relu}(\gamma_{th} - \gamma_{BU}[i]) \\ &- \sum_{k=1}^{K} \lambda_3 \mathbb{F}_{relu}(\gamma_k^{UE}[i] - \gamma_{th}) - \sum_{k=1}^{K} \lambda_4 \mathbb{F}_{relu}(\gamma_k^{BE}[i] - \gamma_{th}) \end{aligned} \tag{19}$$

The $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weighted coefficients, and the $\mathbb{F}_{relu}$ in the function above is defined as

$$\mathbb{F}_{relu} = \begin{cases} x & x > 0 \\ 0 & x < 0 \end{cases}.$$

As such, a typical MDP for the trajectory planning problem can be defined as follow:

- $\mathbf{S}$: each state $s \in S$ contains the location of UAV, the location of aim point, the reference signal received power from the BS to UAV, the reference signal received power from the UAV to BS, the received power of interference and noise, and the estimated received power of eavesdroppers.
- $\mathbf{A}$: The action corresponds to the speed $\mathbf{v}$;
- $\mathbf{P}$: The state transition probability for the location of UAV depends on the environment, such as wind, for the received power depending on the channel, which is shown in equation (1),(2),(8). Besides, the estimation error will influence the estimated received power of eavesdroppers.
- $\mathbf{R}$: The objective of equation (19).

C. Model assisted Reinforcement Learning for Security Trajectory Planning

As discussed before, our policy learning algorithm contains two mutually independent parts, the RL and the MAL. The two parts cooperatively train the policy represented by a fully-connected neural network. The RL targets solving the MDP problem that we formulated in Section III.C. However, the MAL part treats the problem $P2$ as an optimization problem that directly updates the policy. Both of them are offline methods, in which the agent updates the parameter of the neural network after collecting several interacting data.

In the RL part, we leverage the PPO algorithm. The PPO precisely updates the policy to the direction of better performance in each step, though constraining the Kullback-Leibler divergence between the policy $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}_{\text{old}}}$, where $\pi_{\boldsymbol{\theta}_{\text{old}}}$ is the vector of policy parameters before the update. Hence, the loss function for the update the policy is:

$$L(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_t|s_t)}\hat{A}_t - \beta \mathbb{K}\left[\pi_{\boldsymbol{\theta}_{\text{old}}}(\cdot|s_t), \pi_{\boldsymbol{\theta}}(\cdot|s_t)\right]\right], \tag{20}$$

where the $\hat{A}_t$ is an estimator of the advantage value at timestep t, which can be estimated by reward. $\mathbb{K}$ means the Kullback-Leibler divergence. The paper [16] details the whole processing.

Applying the RL in trajectory planning could face several challenges. The first thing is that communication security is hard to protect because the RL algorithm needs to learn the security knowledge from being eavesdropped with or interfered samples. The second thing is that the RL chooses the action for long-term consideration because the advantage value contains the afterward step reward estimation. The situation could happen, that the agent could ignore the current step security to keep the following steps within a safe place. Besides, the RL is time-consuming due to the low data efficiency and uncertain exploration.

Usually, the RL agent explores the environment by treating the environment as a block box. However, in our problem, the communication model has the corresponding mathematical model. Thus, we propose MAL utilizing the punishment $J[i]$ that is generated by the communication system model to constrain the policy. It is obvious that the punishment $J[i]$ is related to the state s_t and a_t, and the $a_t = \pi_{\boldsymbol{\theta}}(s_t)$. By utilizing this character, we employ the gradient ascent algorithm to update the parameter $\boldsymbol{\theta}$. Hence, the policy is enforced to satisfy the constraint of problem $P1$.

The MAL shares several features that can make up for the shortage of RL. The first thing is that the MAL mainly focuses on the action choice for the current step and can enforce the policy to choose the safe action at the current step. The second thing is that the MAL can update the model more flexibly with the loss function generated by the communication model, which is data-efficient.

$$L_{\text{ma}} = \sum_{j \in E} |\gamma_j[n] - \gamma_{th}| + |\gamma_{UAV}[n] - \gamma_{th}| + |\gamma_{BS}[n] - \gamma_{th}| \tag{21}$$

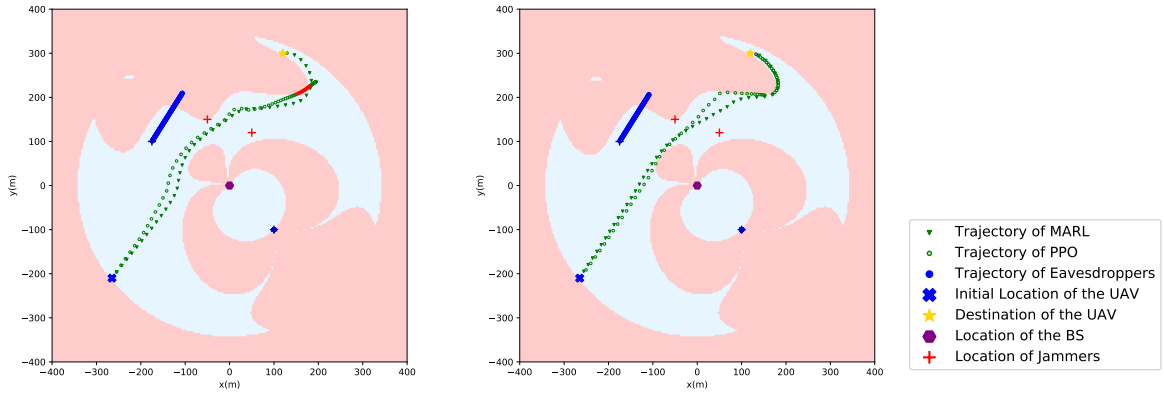
By a combination of RL and MAL, the learning process is shortened significantly. Hence, the parameter of a neural network can be updated by:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_r \nabla_{\boldsymbol{\theta}} L + \alpha_m \nabla_{\boldsymbol{\theta}} J. \tag{22}$$

To balance the update of RL and MAL, we set the same limitation for the gradient norm of parameter $\boldsymbol{\theta}$ to RL and MAL for each step.

D. Overall Algorithm

The proposed algorithm for security UAV navigation with MARL is summarized in Algorithm 1. The MAL is implemented in each step to ensure the security, while the RL updates the policy in each round.



(a) Training for 400 epochs (b) Training for 1000 epochs (c)

Fig. 2. The environment and the trajectories of UAV

Algorithm 1 MARL for security UAV navigation

Randomly initialize the policy $\pi(S|\boldsymbol{\theta}^{\pi})$. Set the parameter α_r, α_m, the gradient norm boundary β and the max allowed step T_{max};
for Epoch = 1 to T_e **do**:
  $t = 0$
  while the UAV not reach the aim point or $t < T$:
    Using the policy π interact with environment;
    while the equation (17) is not satisfied:
      Calculate with the loss function (19) and calculate the gradient $\frac{\nabla J}{\nabla \boldsymbol{\theta}}$, Clip the gradient norm to β and apply the gradient to $\boldsymbol{\theta}$;
    end while
  end while
  Calculate with the loss function (18) and calculate the gradient $\frac{\nabla L}{\nabla \boldsymbol{\theta}}$, Clip the gradient norm to β and pply the gradient to $\boldsymbol{\theta}$;
End For

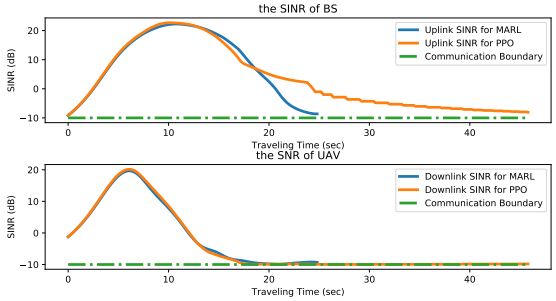
IV. NUMERICAL RESULTS

In this section, we provide numerical results to verify the performance of our proposed MARL algorithm for security UAV navigation. We compare our proposed method with the PPO algorithm. We trained the neural network 750 rounds for both algorithms, we assume we can train the UAV in a simulation environment. In our simulation, we assume that the maximum speed of the UAV is $v_{\max} = 30m/s$. We provide the predefined patrol trajectory for eavesdroppers for easy comparison between different algorithms. However, it should be pointed out that our algorithm can suit any trajectory of eavesdroppers. We set the time slot $\delta = 0.2s$. For the positional information parameters, the altitude of the UAV, the BS, eavesdroppers and jammers are $H_U = 100m$, $H_B = 30m$, $H_E = 20m$, $H_J = 40m$, respectively. The initial location and the destination of the UAV is set as $\mathbf{q_I} = [-265, -210]^T$, and $\mathbf{q_I} = [120, 300]^T$. For the channel parameters, the attenuation coefficient $\kappa = 0.05$, the AWGN power assumed to be $\sigma^2 = -130dBmW$, the SINR threshold for the UAV and the BS is $\gamma_{th1} = -10dB$, and that of eavesdroppers is $\gamma_{th2} = 5dB$. The modeling parameters for P_{LoS} is $a = 9.53$ and $b = 0.41$, respectively. The UAV and the BS is assumed to transmit data at their maximum transmission power, denoted as $P_U = 13dBmW$, $P_B = 13dBmW$, respectively. The UAV is equipped with a 4-antenna elements ULA, and the BS is equipped with an 8-antenna elements ULA. We assume the beams are always aligned.

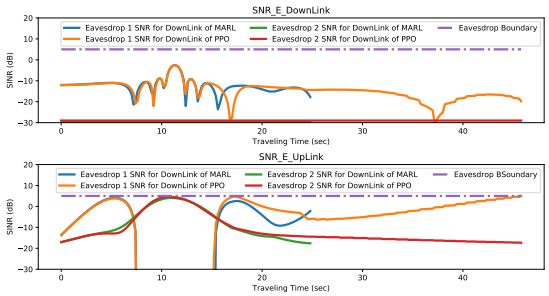
Fig. 2 shows the trajectories of UAVs and eavesdroppers. The UAV's mission is to fly from the blue cross to the golden star while communicating with the BS represented with a purple diamond. The red plus marker represents the jammers' location, and the blue point represents the trajectory of eavesdroppers' or eavesdroppers' location. The region in blue represents the area served safely by the BS, while the region in red corresponds the opposite meaning. Thus, the green line represents the area that can satisfy all constraints of problem $P1$. On the trajectory line, the blue markers correspond that the UAV is in the eavesdropping state, the red marker corresponds that the UAV is interfered with by jammers, and the green marker corresponds that the UAV is secured to communication. The trajectory line with triangle marker is the result of MARL, and the line with square marker is the result of PPO. It is easy to find that the MARL can directly keep safe, while the PPO needs the eavesdropped and interfered samples to learn to satisfy the constraint.

Fig.3 shows the SNR for both algorithms in 400 epochs to indicate the constrain of each step. The dash lines mean the communication boundary. Fig.3(a) shows the SINR for UAV and BS. It is clear that the PPO is out of boundary in the 150th step to the 200th step. Fig.3(b) shows the SNR for eavesdroppers. Both algorithms ensure the UAV has not been eavesdropped.

The specific information shows in Fig. 4. In Fig. 4, the left y-axis represents the average epoch rewards, and the right y-axis represents the average length to achieve the aim point. We



(a) The corresponding SNR for BS and UAV of trajectory



(b) The corresponding SNR for eavesdroppers of trajectory

Fig. 3. The SNR crave for trajectory

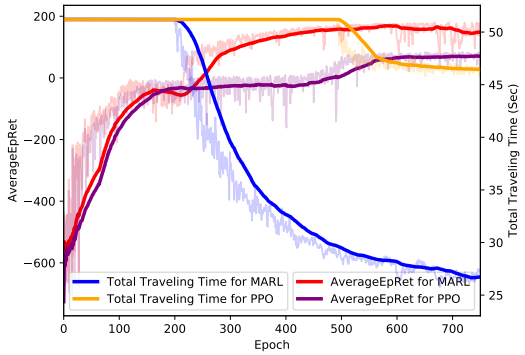


Fig. 4. The learning crave for both algorithms

can see the trajectory of MARL uses 26 seconds to achieve the aiming point while keeping the UAV always fling in the safe flying area, while the PPO achieves this aim for 47 seconds. It is easy to find that the MARL can learn the policy faster and achieve better performance from the red and purple line of Fig 4.

V. Conclusion

This paper studied the UAV navigation problem in a secrecy MIMO communication system by exploiting the UAV trajectory design. The UAV trajectory was designed to minimize the mission time subject to the security constraints as well as the QoS constraint of BS and UAV. We proposed an efficient learning algorithm to solve this design problem. Numerical results show that the new approach of adjusting the UAV trajectory can significantly improve the physical layer security performance of UAV communication systems. In the future work, we will investigate a more general scenario and improve the generalization performance of the algorithm.

References

[1] F. Qi, X. Zhu, G. Mang, M. Kadoch, and W. Li, "UAV Network and IoT in the Sky for Future Smart Cities," *IEEE Network*, vol. 33, no. 2, pp. 96–101, Mar. 2019.

[2] Y. Zeng, Q. Wu, and R. Zhang, "Accessing From The Sky: A Tutorial on UAV Communications for 5G and Beyond," *arXiv:1903.05289 [cs, eess, math]*, Mar. 2019.

[3] X. Sun, D. W. K. Ng, Z. Ding, Y. Xu, and Z. Zhong, "Physical Layer Security in UAV Systems: Challenges and Opportunities," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 40–47, Oct. 2019.

[4] B. Li, Z. Fei, Y. Zhang, and M. Guizani, "Secure UAV Communication Networks over 5G," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 114–120, Oct. 2019.

[5] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV Communications via Trajectory Optimization," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec. 2017, pp. 1–6.

[6] M. Cui, G. Zhang, Q. Wu, and D. W. K. Ng, "Robust Trajectory and Transmit Power Design for Secure UAV Communications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 9042–9046, Sep. 2018.

[7] Y. Zhou, P. L. Yeoh, H. Chen, Y. Li, R. Schober, L. Zhuo, and B. Vucetic, "Improving Physical Layer Security via a UAV Friendly Jammer for Unknown Eavesdropper Location," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11 280–11 284, Nov. 2018.

[8] Y. Cai, Z. Wei, R. Li, D. W. K. Ng, and J. Yuan, "Joint Trajectory and Resource Allocation Design for Energy-Efficient Secure UAV Communication Systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4536–4553, Jul. 2020.

[9] X. Zhou, Q. Wu, S. Yan, F. Shu, and J. Li, "Uav-enabled secure communications: Joint trajectory and transmit power optimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4069–4073, 2019.

[10] H. Lee, S. Eom, J. Park, and I. Lee, "UAV-Aided Secure Communications With Cooperative Jamming," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9385–9392, Oct. 2018.

[11] 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 Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7761–7775, 2019.

[12] 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, Sep. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8432464/

[13] U. Challita, W. Saad, and C. Bettstetter, "Cellular-Connected UAVs over 5G: Deep Reinforcement Learning for Interference Management," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.

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

[15] V. Mnih, K. Kavukcuoglu, and D. Silver, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv:1707.06347 [cs]*, Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[17] Y. Zeng, Q. Wu, and R. Zhang, "Accessing From the Sky: A Tutorial on UAV Communications for 5G and Beyond," *Proceedings of the IEEE*, vol. 107, no. 12, pp. 2327–2375, 2019.

[18] J. Zhao, F. Gao, L. Kuang, Q. Wu, and W. Jia, "Channel tracking with flight control system for uav mmwave mimo communications," *IEEE COMMUNICATIONS LETTERS*, pp. 1–1, 2018.