Trajectory Optimization and Power Control Of UAV-Assisted Mobile Edge Devices

Yunfeng Xia, Qingling Liu(✉), Shihao Wang, and Kuixian Li

Harbin Engimneering University, Harbin, China
liuqingling@hrbeu.edu.cn

Abstract. Mobile edge computing (MEC) is an important aspect of 5G networks, and the deployment of mobile edge devices on unmanned aerial vehicles (UAVs) is becoming an increasingly popular trend in the future. However, when the communication bandwidth resources are limited, the co-frequency link interference in the UAV-assisted MEC may lead to poor communication quality. Thus, this paper proposes a trajectory planning and power allocation scheme based on multi-agent deep reinforcement learning (DRL). To begin, we design a joint optimization goal, which aims to minimize link interference while simultaneously maximizing communication link throughput. Next, we propose an improved DRL algorithm for a multi-agent scenario that enables autonomous trajectory planning and power control of UAVs, ensuring the reasonable allocation of resources. Experimental results demonstrate that the proposed algorithm has better convergence and greater revenue compared to other benchmark algorithms. Overall, this paper presents a promising approach for improving the performance of UAV-assisted MEC through intelligent decision-making and resource allocation.

Keywords: Mobile edge server · Trajectory optimization · Power control · Deep reinforcement learning · Multi-agents

1 Introduction

5G mobile communication has expanded rapidly. Ultra-low latency (ULL), enhanced mobile broadband (eMBB), mass machine type communication (mMTC) and high reliability features have improved the standards of 5G and future 6G communications [1]. As a new distributed computing method, MEC can improve the computing and storage capacity of terminals, reduce communication delay and energy consumption, and play an increasingly important role

This work is supported by the National Natural Science Foundation of China (No: 62201172), the National Key Research and Development Program of China (2022YFE0136800). This work is also supported by Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin, China.

J. Wang (Ed.): MobiMedia 2023, LNICST 576, pp. 387–399, 2025.
https://doi.org/10.1007/978-3-031-60347-1_32

in 5G and future 6G networks. With its flexible mobility and LoS transmission, the UAV platform can break the limitations of traditional network infrastructure and be more effectively applied to multi-point and multi-hop remote transmission networks with poor communication quality, high transmission pressure or difficult to provide communication coverage [2].

So far, there have been many researches on providing communication services or edge computing services for ground users through UAVs.

There are many ways to study trajectory optimization. [3] extends DDPG to multi-agent scenarios, and proposes a multi-agent RL method for dynamic trajectory optimization and bandwidth allocation of UAVs. [4] proposes a Dinkelbach algorithm and successive convex approximation technology to maximize the energy efficiency of UAV. [5] improves service quality through semi-markov decision process from the perspective of user tasks. In [6], a knowledge tracking model and RL path recommendation algorithm are proposed to improve the efficiency of path recommendation. In [7], an improved genetic algorithm (GA) is proposed to implement the trajectory planning of UAV. Aiming at the possibility of eavesdropping attack when relay transmits data, a relay selection algorithm is proposed to plan the connection trajectory and improve the security connection performance in [8].

There are also many researches on power control of UAV. [9] establish a revenue model based on users, and consider the relevant economic cost of MEC services. In [10], an adaptive cooperative Q-learning power allocation algorithm is proposed to improve the adaptability of changeable scenes. In [11], the method of combining DRL with short and long term memory (LSTM) is adopted to improve the system throughput. [12] proposed a DRL channel allocation scheme for multi-UAV system, which used LSTM network to improve the learning ability of agents. [13] proposes a multi-agent layered network architecture, and designs a Multiple Agent Reinforcement Learning (MARL) algorithm to jointly dispatch sensing, computing and communication resources. [14] proposed a new framework based on stochastic game theory, and proposes a MARL algorithm to jointly optimize the power level and sub-channel selection strategy of UAV. While improving the system capacity and spectral efficiency of MEC and D2D communication, Gaussian distribution is adopted as a parameterization strategy to avoid interference to cellular users in [15].

However, in the case of limited bandwidth resources, multiple edge devices working at the same frequency will inevitably cause link interference to each other. But it is hoped the quality of service (QoS) of each link will be guaranteed. In view of this, we propose a method of UAV power control and trajectory optimization based on DRL. Key contributions to this paper include:

(1) Aiming at the problem of co-frequency link interference in MEC system, we propose a trajectory optimization and power control model for multiple UAVs. The goal is to increase the system channel capacity while reducing the interference between links.

(2) A joint optimization scheme for UAV trajectory and transmit power is proposed, with energy efficiency as the overall optimization objective. A multi-agent DRL method is proposed to maximize energy efficiency.

(3) We propose an enhanced multi-agent RL algorithm that builds on the QMIX algorithm. Our approach employs a mix network that takes the output of each critic network as input, producing a system evaluation that maximizes overall revenue.

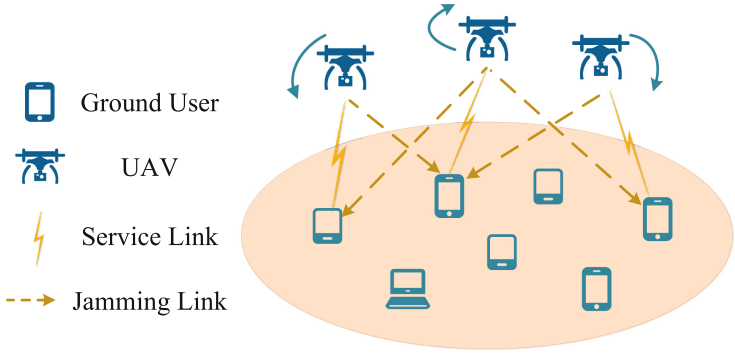

Fig. 1. System scenario

2 Proposed Method

2.1 System Model

As shown in Fig. 1, the system scenario includes several UAV-assisted mobile edge devices, as well as several ground terminal users. Each UAV can only communicate with one ground terminal user at a time, which is expressed as $L = \{1, 2, \cdots, l\}$. UAVs are represented by the set $U = \{1, 2, \cdots, u\}$, and ground users are represented by the set $M = \{1, 2, \cdots, m\}$. Assuming that the number of UAVs is less than that of ground users, there are at most U communication links at the same time. According to the position of ground users, each UAV dynamically adjusts its trajectory to improve the maximum information transmission rate. At the same time, UAVs also control their transmit power to reduce the interference to other links. Therefore, while adjusting the trajectory, the UAV also needs to adjust its transmit power.

We express the horizontal position of the ground user as $w_m = [x_m, y_m] \in \boldsymbol{R}^{2\times 1}, \forall m \in M$. The horizontal coordinate of UAV is expressed as $w_u = [x_u, y_u] \in \boldsymbol{R}^{2\times 1}, \forall u \in U$. Assumed the height H of UAV is fixed. And a period of time T is divided into N slots, expressed as $n = 1, 2, \cdots, N$ The change of flight trajectory is controlled by the speed of UAV, which is expressed as $v_n^u = [v_x^u, v_y^u] \in R^{2\times 1}, \forall u \in U$. The speed of each UAV is limited:

$$V_{\min} < \boldsymbol{v}_n^u < V_{\max} \tag{1}$$

Then the change of UAV space position can be expressed as:

$$w_{n+1}^u = w_n^u + v_n^u dt \tag{2}$$

where $dt = T/N$ is the time slot, which means n. For the sake of safety, there are also restrictions on the space position spacing between UAVs:

$$\| w_n^u - w_n^{u'} \|_2 \geq d_{min}, u \neq u' \tag{3}$$

where d_{min} is the minimum acceptable distance between UAVs.

In the information transmission between UAVs and ground users, we consider the path loss in the communication link. The path loss in free space is related to frequency and distance. We study the co-interference among UAVs, focusing on optimizing the power and flight trajectory variables while not considering frequency. Thus, we treat frequency as a non-variable parameter. From another perspective, we consider the impact of path loss due to frequency as a constant value, and is applicable to any frequency. Therefore, the path loss can be simplified as follows:

$$g_n^{u,m} = \lambda d_{u,m}^{-2} \tag{4}$$

where λ represents the path loss factor and $d_{u,m}^{-2} = \sqrt{H^2 + \| w_n^u - w_n^m \|_2^2}$ is the Three-dimensional Euclidean distance between UAV u and UE m .

Generally, when communication links are closer, the loss caused by the path becomes smaller. However, this proximity can lead to mutual interference between the links. At this moment, reducing the transmission power can reduce the corresponding interference, but it will also affect the channel capacity of the communication link and the quality of service. Therefore, the UAV will adjust its transmit power while planning its trajectory. The transmit power can be expressed as:

$$P_{\min} < \boldsymbol{p}_n^u < P_{\max} \tag{5}$$

Considering the mutual interference between links, the Signal to Interference plus Noise Ratio (SINR) of the communication link can be obtained by using (4) and (5):

$$\gamma_n^{u,m} = \frac{p_n^u g_n^{u,m}}{\sum\limits_{u' \neq u, u' \in U} p_n^{u'} g_n^{u',m} + \sigma^2} \tag{6}$$

where σ^2 is noise power, $p_n^{u'}$ is transmit power of other UAVs and $g_n^{u',m}$ is channel gain of UAVs of other links to the current link.

According to (6), the maximum information transmission rate of each link is:

$$R_n^{u,m} = B\log_2(1 + \gamma_n^{u,m}) \tag{7}$$

where B is the channel bandwidth, and $R_n^{u,m}$ represents the maximum information transmission rate between the UAV u and ground user m.

Then the energy efficiency EE_n^u can be defined as:

$$EE_n^u = \frac{R_n^{u,m}}{p_n^u} \tag{8}$$

2.2 Problem Formulation

Assumed that the positions of all users are randomly distributed in a square fixed area 2km×2km, and the initial position of the UAV is in the center of the area. According to the position of requesting users, each UAV jointly optimizes its flight trajectory and transmit power to maximize the system energy efficiency. The optimization problem is expressed as:

$$\begin{aligned} &\max EE = \frac{1}{N} \sum_{n \in N} \sum_{u \in U} EE_n^u \\ &\text{s.t.}(1), (3), (5) \end{aligned} \tag{9}$$

Equation (9) is a non-convex optimization problem. We jointly optimize the flight speed and transmit power of the UAV through DRL.

3 Madrl-Based Trajectory Planning and Power Control

In this section, we first introduce the framework of the multi-agent algorithm MADDPG based on DDPG. Then, an improved MADDPG algorithm is proposed. Finally, we propose the state, action and reward functions designed for the system model.

3.1 Algorithm Based On DDPG

The algorithm structure of DDPG is actually an actor-critic network structure, that is, its network is mainly divided into actor network and critic network, and each network is divided into current network and target network. The update strategy of critic network is to minimize the mean square error of current network and target network. The actor network is based on the critic network to maximize the output value of the actor network. In short, in the actor-critic network structure of the DDPG algorithm, the output value of the actor network is used as the action of the agent, and the decision-making process of the agent is guided by the evaluation of the critic network, which improves the accuracy and stability of the decision-making. At the same time, the output value of the actor network is also used to update the critic network, improving its evaluation ability of the environmental state and action, and achieving mutual promotion and optimization between the actor and critic networks.

MADDPG is a multi-agent version of DDPG algorithm, and its algorithm framework is consistent with DDPG. MADDPG also uses the actor-critic algorithm framework, but the critic network of MADDPG is different. Its input is

not only the observation space and action space of each agent itself, but also the observation space and action space of all agents. Through the sharing mechanism between agents, centralized training and distributed operation can be achieved, so that each agent is less affected by other agents changing the policy, so that the environment can be regarded as stable under the known information conditions.

3.2 MIX-MADDPG Algorithm

MADDPG has relatively poor scalability when dealing with large-scale Multi-agent system. As the number of intelligent agents increases, the interactions in algorithms become more complex, which may make the learning process more difficult and inefficient. In addition, sharing observation information may result in higher communication overhead, and in contrast, the QMIX algorithm can better handle this situation. Because the QMIX algorithm utilizes a centralized value function for decision-making, the communication cost between intelligent agents is relatively low.

Therefore, we introduce the MIX network in the QMIX algorithm on the basis of MADDPG. After the DDPG algorithm calculates the Q value, the Q values of all agents are used as inputs to a hybrid network, and a total evaluation value Q is output. The Q value is used for forward feedback propagation of the network.

MIX-MADDPG is an improved multi-agent RL algorithm based on MADDPG. The algorithm adopts bottom-to-top reward allocation strategy to maximize the reward return. Specifically, we added a mix network, which takes the critic network output of each agent as input, and outputs an overall evaluation value through the linear layer network of mix. Through the mix network, each agent's evaluation value is linearly changed. Mix networks assign different network weights to each agent's evaluation values, which prevents the agent from getting good rewards if it does not perform well. It is worth noting that each agent's critic network does not share an observation space. Each agent outputs an evaluation value based on its own observations and actions, and evaluates it through a mix network. The pseudo-code is shown in Algorithm 1.

In Algorithm 1, each type of network has a current network and a target network. However, the proposed algorithm requires an update to the mix network after updating the critic network. To achieve this, we combine the update of the mix network and critic network into the same optimizer, as both aim to bring the current Q value closer to the target Q. The loss function for the hybrid network is expressed as:

$$L(\theta) = \sum_{j=1}^{D} \left[\left(y_i^{\text{tot}} - Q_{tot}(s, a; \theta) \right)^2 \right] \tag{10}$$

where D represents the number of samples, and $Q_{tot}(s, a; \theta)$ is the output of mix network. y_i^{tot} is the target value of mix network, and the variance of Q_{tot} and y_i^{tot} continuously decreases during training.

Algorithm 1: MIX-MADDPG for N agents

while *episode* $\leq$ *max_episodes* **do**
  Initial the environment *Env*.
  Create the agents U.
  while *step* $\leq$ *max_steps* **do**
    Each agent u takes action $\boldsymbol{a}$ from observation $\boldsymbol{o}$;
    Interact with env, get next observation $\boldsymbol{o'}$ and reward $\boldsymbol{r}$;
    Store experience sample $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s'})$ in replay buffer $\mathcal{D}$;
    Update $\boldsymbol{s} \leftarrow \boldsymbol{s'}$;
    for *agent* $i = 1, 2, \cdots, N$ **do**
      Randomly sample $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s'})$ from $\mathcal{D}$;
      Calculate $y_{tot} = \sum_i \left(r_i + \gamma Q_i' \left(\boldsymbol{s'}, \boldsymbol{a'}; \omega'\right)\right)$
      Update estimate mix network:
      $\mathcal{L}(\theta) = \sum_j \left(y_{tot}^j - Q_{tot}^{\mu}\left(\mathbf{s}^j, \boldsymbol{a}_i^j\right)\right)^2$
      Update estimate actor network:

$$\nabla_{\theta_i} J \approx \frac{1}{S} \sum_j \nabla_{\theta_i} \boldsymbol{\mu}_i \left(o_i^j\right) \nabla_{a_i} Q_i^{\boldsymbol{\mu}}$$

$$\left(\mathbf{s}^j, \boldsymbol{a}_i^j\right) \big|_{a_i = \boldsymbol{\mu}_i\left(o_i^j\right)}$$

    end
    Update target network:
    $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$
  end
end

The update of actor network is consistent with DDPG, which aims to maximize the evaluation value. However, the difference is to maximize the q value of the mix network.

3.3 Algorithm Elements

1) Observed State

The state space is the system state that each agent can observe in time slot n, that is, the observation space of each agent.

- **UAV trajectory parameters:** UAV trajectory parameters include UAV flight speed $v_n^u = [v_x^u, v_y^u] \in \boldsymbol{R}^{2\times 1}, \forall u \in U$ and horizontal coordinates $w_n^u = [x_n^u, y_n^u] \in \boldsymbol{R}^{2\times 1}, \forall u \in U$.
- **Distance parameters:** The distance parameter includes communication link distance, which is expressed as $d_n^{u,m} = \parallel w_n^u - w_n^m \parallel_2, \forall u \in U, \forall m \in M$. And the distance between UAVs, which is expressed as $d_n^{u,u'} = w_n^u - w_n^{u'}, u \neq u' \in U$.
- **Link parameters:** The link parameters include the current transmit power p_n^u of the UAV and SINR $\gamma_n^{u,m}$ of the current communication link.

2) Action

- **Horizontal speed:** The horizontal speed is expressed as $v_n^u = [v_x^u, v_y^u] \in R^{2\times 1}, \forall u \in U$.
- **Transmit power:** The power p_n^u is designed as a continuous action space, which is not higher than $P_{\max}$.

3) Reward

Reward is the feedback obtained by the agent applying the action to the environment, and is an important bridge connecting the algorithm and the scene. The design of reward value includes communication link energy efficiency, transmit power penalty, boundary penalty and collision penalty, which is represented as follows:

$$r = r_{ee} + r_{power} + r_{bound} + r_{coll} \tag{11}$$

- **Communication link energy efficiency:** The energy efficiency is designed as a positive reward, which can be calculated by SINR and transmit power::

$$r_{ee} = \frac{10\log_2(R_n^{u,m} + 10^{-9})}{p_n^u / P_{\max} + 1} \tag{12}$$

- **Transmit power penalty:** The transmit power penalty means when the power exceeds the specified range, a negative reward will be obtained. It is represented as:

$$r_{power} = \begin{cases} -20, & p_n^u < P_{\min} \text{ or } p_n^u > P_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

- **Boundary penalty:** Boundary penalty means that when the UAV reaches outside the fixed area, it will get a negative reward, which is defined as:

$$r_{bound} = \begin{cases} -20, & \text{out of bounds} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

- **Collision penalty:** Collision penalty means that when the distance between two UAVs is less than a threshold, they will be punished. It is expressed as:

$$r_{coll} = \begin{cases} -50, & d_{u,u'} < d \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

4 Simulations

In this section, we provide experimental simulation to verify the feasibility of our proposed scheme and the effectiveness of the algorithm.

We refer to the experimental parameters in [3]. Table 1 provides the parameter settings of our algorithm in simulation. We choose "Leaky Relu" and "Tanh" functions as our network activation layer.

Table 1. Parameters settings

Parameter	**Value**
Maximum episodes	50000
Maximum running steps per episode	500
Target network weight update coefficient τ	0.01
Reward discount factor γ	0.8
Batch size	256
Memory size	1e6
Learning rate of policy network optimizer θ_a	1e-4
Learning rate of value network optimizer θ_c	1e-3
Learning rate of value network optimizer θ_m	1e-3
Size of Hidden Layer	(256, 128)
Size of Mix Layer	(128, 128)

For the experimental scenario, we design a square area of 2km×2km. The number of ground users in the field is $M = 10$ randomly distributed, and the number of UAVs is $U = 3$. The time slot $dt = 1s$, and the UAV can only provide edge services to one terminal user at the same time. The requesting users in each episode randomly select three and establish communication links with three UAVs. This ensures the randomness of the link and satisfies the user service requests at any location. Assume that the altitude of all UAVs is fixed at $H = 200$ m, the maximum speed of UAV is 10 m/s, and the maximum transmit power is 200 mW. In addition, the minimum value of speed constraint and power constraint is 0. The collision penalty distance between UAVs is 25 m.

The path loss factor in the communication link is set to -40 dB, the PSD of white noise is set to -167 dB, and the bandwidth of each link is 1 MHz.

Note that the Fig. 2 shows the actor network loss curve of the three algorithms. Because the experimental set is too large, we collect the loss value every 100 episodes. Fig. 3 and Fig. 4 are the same.

Compared with DDPG, multi-agent RL algorithm has faster convergence speed and lower convergence loss. The reason is DDPG will change the state of the environment when making action decisions. The actions of other subjects are only aimed at the environment before the change, not necessarily the best strategy after the change. Therefore, the actions of each agent will affect the change of the environment, thus affecting the agent's decision. Each agent can share the observation space and reward value with other agents, so as to avoid mutual interference of action strategies between agents.

As can be seen from Fig. 2, the three algorithms have fully converged in about 10000 episodes. The loss value of MADDPG algorithm begins to decline at the beginning of training and converges around 5000 sets. Moreover, their loss value fluctuates slightly, and their convergence speed is much faster than that of MIX-MADDPG.

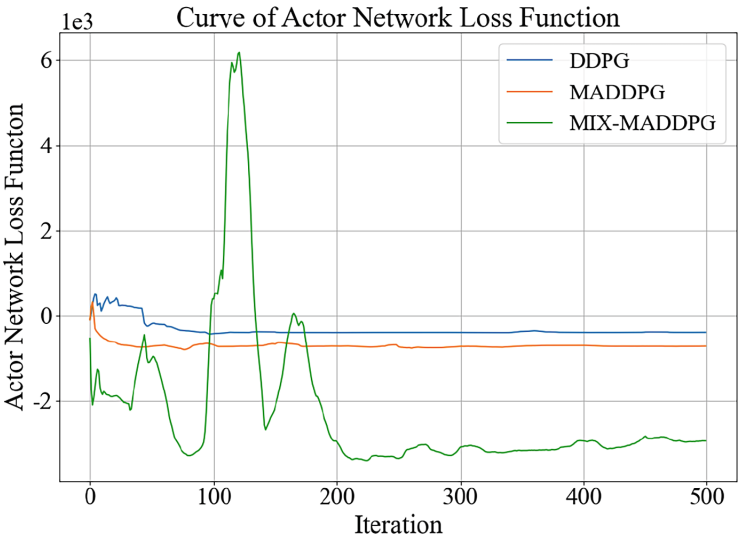

Fig. 2. Convergence curve of actor network.

The loss value of MIX-MADDPG algorithm before convergence fluctuates greatly. It may be due to the large difference between the critic output value and the mix network output value in the early training stage, leading to over-exploration in the early stage. Additionally, when updating the critic network, the large discrepancy between the target Q value (which includes the episode reward value) and the current Q value results in a large mean square error (MSE), causing some disparity between the target action and the current action. However, compared with the other two algorithms, MIX-MADDPG has better convergence. Its actor convergence value is far lower than the other two.

Figure 3 shows reward convergence curve of the three algorithms. As shown in the figure, the three algorithms converged before 10000 episodes. Compared with the other two algorithms, DDPG has slower convergence speed and lower reward value. In contrast, the MADDPG algorithm converges faster and attains higher and most stable reward value, exhibiting the fastest convergence rate among the three algorithms. The sharing of observation space among agents can enhance the benefits of their individual decision-making processes, enabling the algorithm to quickly find the optimal strategy and ultimately converge.

On the other hand, our improved algorithm exhibited significant reward value fluctuations in the early stages, even experiencing a sharp decline at around 8000 episodes. This may be attributed to over-exploration, which makes it challenging for the algorithm to find the optimal direction in a brief period. Nevertheless, once the algorithm converges, it outperforms the other two algorithms in terms of reward value. Despite not sharing observations, it corrects environmental instability caused by individual agents through the Mix network. As a result, it can obtain a higher reward value compared to MADDPG.

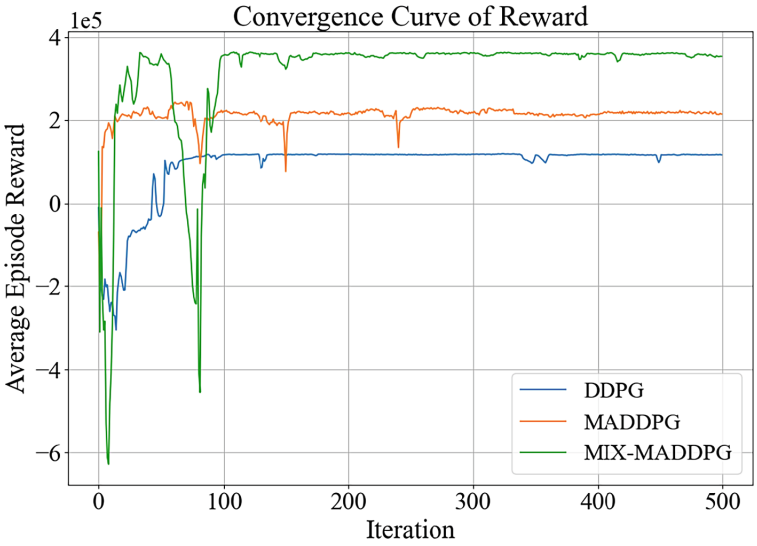

Fig. 3. Convergence curve of reward.

In Fig. 4, the throughput change curve of three agents is depicted. Prior to algorithm convergence, the throughput of the three agents are relatively small, possibly due to the absence of a clear trajectory direction during the initial exploration phase of each UAV. As the position of each UAV changes, the power will also change. In addition, each UAV will also consider the interference to other links. Thus, in the early stage of exploration, it is easy to fall into two extremes, that is, the large power of UAV causes the long link distance, or the small power makes UAV close to the user.

Subsequently, after the algorithm converged, one of the three UAVs exhibits a larger throughput, while the other two have smaller throughputs. The UAV has been able to find the optimal direction of the trajectory, so the UAV is more focused on power control. To avoid link interference, when the power output of one UAV is high, the other two decrease their output, resulting in changes to their throughput values.

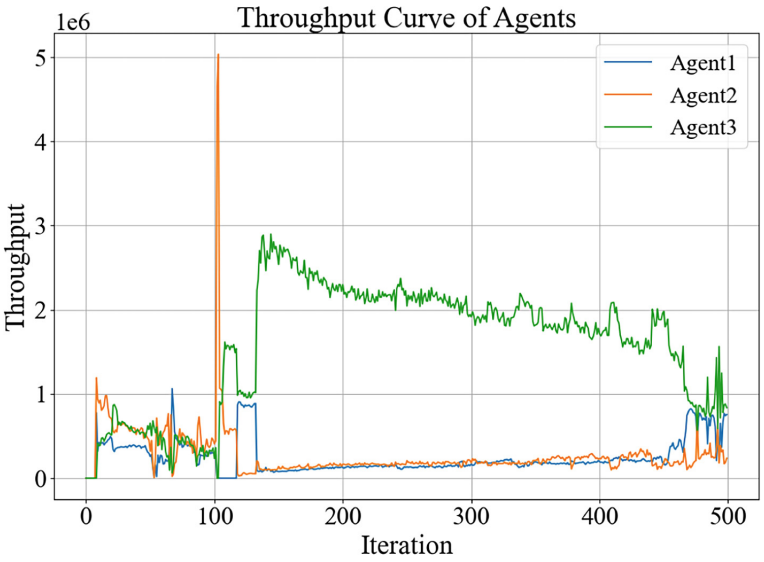

Fig. 4. Throughput curve of agents.

5 Conclusion

This paper studies the trajectory planning and power control of multi-UAV system, aiming at improving the energy efficiency. Specifically, we seek to jointly optimize the trajectory and transmit power of UAVs to improve the maximum throughput of the system and reduce mutual interference of co-frequency links. We propose an improved MADDPG-based multi-agent DRL algorithm. We adopt centralized training and decentralized execution strategies to promote better learning strategies for multiple agents. The results show that compared with DDPG and MADDPG, our algorithm has lower loss function value, which means it has better convergence. The reward value of the algorithm is higher than that of other algorithms, indicating that it has better benefits. Of course, our algorithm did not perform well in the early stage, which is also the point we need to improve. Future work will also focus on improving the scenario by considering other aspects of energy consumption.

References

1. Wu, Q., Zeng, Y., Zhang, R.: Joint trajectory and communication design for multi-UAV enabled wireless networks. IEEE Trans. Wirel. Commun. **17**(3), 2109–2121 (2018)
2. Kawamoto, Y., Nishiyama, H., Kato, N., et al.: A traffic distribution technique to minimize packet delivery delay in multilayered satellite networks. IEEE Trans. Veh. Technol. **62**(7), 3315–3324 (2013)

3. Wang, W., Lin Y.: Trajectory design and bandwidth assignment for UAVs-enabled communication network with multi - agent deep reinforcement learning. IEEE 94th Veh. Technol. Conf. (VTC2021-Fall), Norman, OK, USA, pp. 1–6 (2021)
4. Li, M., Cheng, N., Gao, J., Wang, Y., Zhao, L., Shen, X.: Energy-efficient UAV-assisted mobile edge computing: resource allocation and trajectory optimization. IEEE Trans. Veh. Technol. **69**(3), 3424–3438 (2020)
5. Li, J., Liu, Q., Wu, P., Shu, F., Jin, S.: Task offloading for UAV-based mobile edge computing via deep reinforcement learning. IEEE Int. Conf. Commun. China (ICCC), Beijing, China, pp. 798-802 (2018)
6. Cai, D., Zhang, Y., Dai, B.: Learning path recommendation based on knowledge tracing model and reinforcement learning. IEEE 5th international conference on computer and communications (ICCC), pp. 1881–1885. IEEE (2018)
7. Asim, M., Mashwani, W.K., Shah, H., et al.: An evolutionary trajectory planning algorithm for multi-UAV-assisted MEC system. Soft Comput. —bf 26(16), 7479–7492 (2022)
8. Si, G., Dou, Z., Lin, Y., Qi, L., Wang, M.: Relay selection and secure connectivity analysis in energy harvesting multi-hop D2D networks. IEEE Commun. Lett. **26**(6), 1245–1248 (2022)
9. Xue, J., Wu, Q., Zhang, H.: Cost optimization of UAV-MEC network calculation offloading: a multi-agent reinforcement learning method. Ad Hoc Netw. **136**, 102981 (2022)
10. Dou, Z., Si, G., Lin, Y., et al.: A power allocation algorithm based on cooperative Q-learning for multi-agent D2D communication networks. Phys. Commun. **47**, 101370 (2021)
11. Lin, Y., Wang, M., Zhou, X., Ding, G., Mao, S.: Dynamic spectrum interaction of UAV flight formation communication with priority: a deep reinforcement learning approach. IEEE Trans. Cognit. Commun. Netw. **6**(3), 892–903 (2020)
12. Zhou, X., Lin, Y., Tu, Y., et al.: Dynamic channel allocation for multi-UAVs: a deep reinforcement learning approach. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2019)
13. Li, T., Leng, S., Wang, Z., et al.: Intelligent resource allocation schemes for UAV-swarm-based cooperative sensing. IEEE Int. Things J. **9**(21), 21570–21582 (2022)
14. Cui, J., Liu, Y., Nallanathan, A.: Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. IEEE Transactions on Wireless Communications **19**(2), 729–743 (2020)
15. Wang, D., Qin, H., Song, B., et al.: Resource allocation in information-centric wireless networking with D2D-enabled MEC: a deep reinforcement learning approach. IEEE Access **7**, 114935–114944 (2019)