Author(s): Chang, Zheng; Deng, Hengwei; You, Li; Min, Geyong; Garg, Sahil; Kaddoum, Georges

Title: Trajectory Design and Resource Allocation for Multi-UAV Networks : Deep Reinforcement Learning Approaches

Year: 2022

Version: Published version

Please cite the original version:

Chang, Z., Deng, H., You, L., Min, G., Garg, S., & Kaddoum, G. (2022). Trajectory Design and Resource Allocation for Multi-UAV Networks : Deep Reinforcement Learning Approaches. IEEE Transactions on Network Science and Engineering, 10(5), 2940-2951. https://doi.org/10.1109/tnse.2022.3171600

Trajectory Design and Resource Allocation for Multi-UAV Networks: Deep Reinforcement Learning Approaches

Zheng Chang, *Senior Member, IEEE,* Hengwei Deng, Li You, *Member, IEEE,*
Geyong Min, *Senior Member, IEEE,* Sahil Garg, *Senior Member, IEEE,* and Georges
Kaddoum *Senior Member, IEEE.*

◆

Abstract—The future mobile communication system is expected to provide ubiquitous connectivity and unprecedented services over billions of devices. The unmanned aerial vehicle (UAV), which is prominent in its flexibility and low cost, emerges as a significant network entity to realize such ambitious targets. In this work, novel machine learning-based trajectory design and resource allocation schemes are presented for a multi-UAV communications system. In the considered system, the UAVs act as aerial Base Stations (BSs) and provide ubiquitous coverage. In particular, with the objective to maximize the system utility over all served users, a joint user association, power allocation and trajectory design problem is presented. To solve the problem caused by high dimensionality in state space, we first propose a machine learning-based strategic resource allocation algorithm which comprises of reinforcement learning and deep learning to design the optimal policy of all the UAVs. Then, we also present a multi-agent deep reinforcement learning scheme for distributed implementation without knowing a priori knowledge of the dynamic nature of networks. Extensive simulation studies are conducted and illustrated to evaluate the advantages of the proposed scheme.

Index Terms—Trajectory design, Resource allocation, Multi-agent reinforcement learning, Deep learning, UAV, Drone.

1 INTRODUCTION

1.1 Background and Motivation

The increasing demand for high quality wireless services urges the future wireless communication system to provide ubiquitous connectivity and coverage over all kind

of mobile devices. The diversity of network applications also poses strict requirements on network capacity, service latency and energy consumption for trillions of mobile devices. To realize the vision of essentially unlimited access to wireless data anywhere and anytime for anything, the recent emerging unmanned aerial vehicle (UAV)-based flying platforms are able to break the limitations of traditional network infrastructure [1], which urges to rethink the development of next generation communication systems. The UAV, also known as drone, has attracted many attentions due to its prominent in flexibility, easy and low cost deployment [2]. Because of its high flying attitude, the UAV-based platform can establish the effective Line-of-Sight (LoS) links with the ground-users (GUs), thus to reduce the energy consumption for reliable connectivity [3]. Therefore, an UAVs-based flying mobile communication system provides a cost- and energy-efficient solution with limited territorial cellular infrastructure for the GUs.

Developing an UAV-enabled wireless communications system has received attracted a large amount of research interests. To date, majority of the works have dedicated on the single UAV two-dimension (2-D) or 3-dimension (3-D) deployment/placement optimization problems, with the assumption that UAV can serve as aerial quasi-static base stations (BS) or relay. Although adding a single UAV into the cellular network has shown its potential on performance enhancement, it has limited communications, caching and computing capability in general, which is not preferred for mission-critical services and a large number of GUs. Correspondingly, deployment of a swarm of UAVs is motivated. In the multi-UAV communication system, multiple UAVs may cooperatively serve the GUs in a large area. Moreover, different GUs could be served simultaneously with lower latency and higher throughput, which could address some throughput- and latency-related problems brought by a single-UAV system.

On the other hand, current works on the multi-UAV network usually focus on the proposals of trajectory design and resource allocations in a static manner considering the UAVs can act as BSs. In order to provide long-term effective connectivity and reliable coverage, UAV-based network

with high mobility needs to be carefully designed and different UAVs should autonomously work as a team and their interactions should be explored. Therefore, establishing an efficient, smart and autonomous multi-UAV network emerges as a research topic with profound importance while is still under-investigated. Addressing such a topic is typically challenging. First, due to its high cost and limited communication capability, the mobility/route of different UAVs should be designed and coordinated with high accuracy to cover a large area over a long run. Moreover, fairness is also critical for the UAV network as the UAVs should move around to ensure the communication coverage. In addition, the energy consumption issues should be seriously considered as the UAVs are typically with limited energy supply and should be recharged from time to time. Last but not the least, the UAVs are usually deployed to where the network access is limited to execute mission-critical services. Certain degree of autonomy or self-organizing is highly preferred.

To address the aforementioned problems, and develop a smart and autonomous multi-UAV communication systems, we propose to leverage deep reinforcement learning (DRL) framework, which recently demonstrates a potential on improving the performance of wireless network. Due to the fact that RL can enable UAVs to choose their policies for optimizing the objectives without a priori knowledge of the environment, it is suitable to address the trajectory control and resource allocation in the multi-UAV wireless networks. Specifically, we consider that all the UAVs share the same spectrum to serve the GUs. By focusing on the downlink of the network, i.e. transmissions from the UAVs to GUs, the objective of this work is to maximize the system utility among all the GUs by jointly optimizing the power allocation, user association, and UAV trajectory in a given finite period. Addressing the formulated joint optimization is challenging, because the transmit power allocation, user association, and UAV trajectory design optimizations are actually coupled. Correspondingly, for the formulated problem, the DRL is able to provide a promising solution because it can solve the problem of high dimensionality in state-action space and also handle the time-varying environment [4]. The DRL uses Deep Neural Networks (DNNs) to the decision making process, which can offer significant performance improvement to many learning problems with limited or even zero knowledge. Moreover, developing decentralized approaches is becoming more needed than ever due to the complexity of the multi-UAV wireless networks. Though it can be very challenging to design them, decentralized approaches scale well, as they typically incur little to no communication and computational overhead while still performing relatively well. Thus, we also consider the decentralized feature of multi-UAV system, and propose to utilize the multi-agent DRL to design a distributed algorithm [5], which enables the way towards an autonomous UAV communications system.

1.2 Related works

The research on the UAV-based wireless communication systems have mainly concentrated on the UAV placement and resource optimization [3]- [19], with the assumptions that UAV can serve as aerial BSs or aerial relay to support GUs. For the trajectory design, the altitude of the UAV can be optimized with or without the horizontal location based on different considerations and QoS requirements. In [3], the authors aim to maximize the communication coverage by optimizing the altitude of the a single UAV wireless network. The authors of [6] utilize stochastic geometry-based approach to analyze two-tier wireless network consisting of BSs and aerial BS. General probabilistic LoS and NLoS propagation models are assumed and coverage probability and spectral efficiency are derived with the consideration of the height of the aerial BS. In [7], the authors jointly optimize the altitude of UAVs, the duration of transmission phases and the antenna configuration to maximize the coverage, under the assumptions of UAV and ground BS with distributed access points and multiple antennas.

In contrast, there are several papers working on the two-dimensional (2-D) trajectory design (e.g. the horizontal positions) of the UAV by fixing its altitude. To address the problem of control over a group of UAVs in a long term, the authors of [8] utilize the deep reinforcement learning to minimize the energy consumption of the overall network while maintaining the reliable connectivity. In [10], the authors consider the UAV flies to a given location for certain mission and it needs reliable communication with BSs at each time slot. The aim is to minimize the completion time of the UAV by 2-D trajectory optimization, subject to the connectivity constraint of BS-UAV link. The authors of [11] investigate the cooperation of a group of UAVs, and propose mode selection between UAV-to-infrastructure and UAV-to-UAV modes for data delivery. Then the resource allocation and speed optimization are propose to maximize the uplink data rate. In [12], the authors investigate the UAV-based secure communication. A two-UAV system is considered where one is for data transmission and the other one is to jam the eavesdroppers on the ground. The minimum worst-case secrecy data rate of the GUs is optimized by designing UAVs' trajectories and user scheduling.

As for the 3-D trajectory design, in [13], both periodic and temporal operation modes are considered for the UAV system. In each case, the aim is to minimize the duration of UAV flight or mission completion time. In [14], the authors propose to maximize the minimum throughput of all the GUs in order to achieve fair performance. The route design, power allocation and user scheduling schemes are presented. The authors of [15] consider UAV provides services for a group of GUs in a dynamic channel scenario, and propose a transmit power allocation and 3-D trajectory design optimization scheme to maximize the minimum throughput of the group in a given time duration. In [16], a drone-based small cell placement problem is explored to maximize the overall system utility. In [17] and [18], by considering joint optimization of the mobility and location of the UAVs, transmit power allocation and user association schemes are presented to improve reliability of the uplink. The authors of [19] investigate the trajectory design and resource allocation problem for maximizing the throughput of a solar powered UAV system over a given time period.

In general, the (deep) multi-agent reinforcement learning has been explored to address control-related problems [20]- [23]. There are increasing efforts to investigate the

potential of multi-agent reinforcement learning (MARL) on the resource allocations in the wireless communication system. The authors of [24] utilize the MARL to address the power allocation problem in D2D communications, while the MARL-based approach is applied to address computation offloading and interference coordination in [27]. The authors explore the MARL on improving the secure performance of wireless network in [28]. In addition, the spectrum access problems in different types of wireless network are addressed via MARL in [29] and [30]. Recently, MARL-based schemes have been gradually applied to the UAV networks [31] [32]. The authors of [31] has utilized the MARL to present distributed trajectory design of multi-UAV network. In [32], MARL-based scheme is also applied for trajectory design when considering a UAV-assisted edge computing system.

As one can observe, there is a lack of works utilizing learning-based schemes on the proposal of joint optimization of trajectory design, power allocation and user association, to effectively and efficiently operate multi-UAV network. Moreover, there is spare study towards an autonomous multi-UAV communication system, which is of profoundly importance towards fully utilizing UAVs in the development of wireless communication system.

1.3 Contribution

In this work, our main target is to utilize collaborative machine learning, i.e., DRL-based scheme and multi-agent DRL-based scheme to tackle the problem of power allocation, user association and trajectory design for multi-UAV communications system. Bearing in mind the above mentioned works, main contributions of this paper are summarized in the following.

- A multi-UAV communication system is considered to serve multiple GUs. A central base controller is assumed to carry out the learning process. With the objective to maximize the system utility, the problem of trajectory design, user association and power allocation is investigated. To address the problems related to the high dimensionality in state space, we first propose a machine learning-based strategic resource allocation algorithm which comprises of reinforcement learning and deep learning to explore the optimal policy of all the UAVs. The proposed centralized DRL process can be carried out at the central base and the UAVs are controlled via the signaling exchange with the base.
- Because the UAV-based network is expected to solve mission-critical problems in reality, an autonomous communication system is preferred. Thus, we further consider a complex scenario and propose to decentralize the considered multi-UAV system. In this setting, no UAVs can observe the underlying Markov state. Instead, each UAV only obtains a private observation correlated with that state. The UAVs are able to utilize dedicated limited-bandwidth channel to communicate with each other, and are fully cooperative and share the goal of maximizing the system utility. However, due to the partial observability and limitation of communication channel, the UAVs have
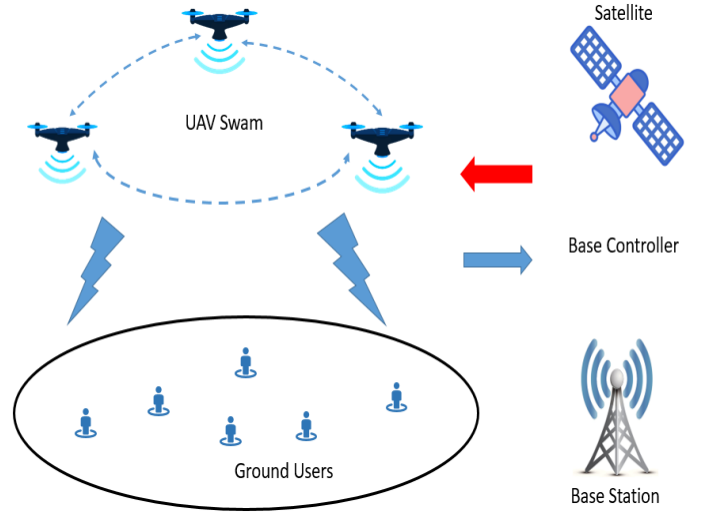

Fig. 1. UAV System Model

to find a communication protocol which is able to coordinate their behavior and policy.
- Consequently, we propose to utilize the centralized learning and decentralized execution. A deep multi-agent reinforcement learning is proposed where the UAVs are considered as the agents. In the proposed scheme, learning is performed via the centralized algorithm, while during execution, the UAVs can communicate through the dedicated limited-bandwidth channel and learn the communication protocol.

1.4 Organization

The reminder of this paper is organized as follows. In Section II, the system model is depicted. Section III present the problem formulation and we propose the DRL-based resource allocation and trajectory design algorithms in Section IV. In Section V, we conduct the performance evaluation through simulation study. Section VI concludes this work.

2 SYSTEM MODEL AND ASSUMPTION

2.1 System Model

The system model is depicted in Fig. 1. There are $M > 1$ UAVs sharing the same frequency spectrum and serving a group of $U > 1$ GUs. The UAV swam and GU set are denoted as $\mathcal{M}$ and $\mathcal{U}$, respectively. Apparently, we have $|\mathcal{M}| = M$ and $|\mathcal{U}| = U$. All the UAVs provide services to the users in consecutive time slots. We denote the time slot as t, and $t \in \{1, 2, ..., T\}$. The overall period is denoted as $\mathcal{T}$. In this work, we consider a 3-D Cartesian coordinate system where the fixed location of each GU u denoted by horizontal and vertical coordinates, e.g., $\phi_u = [x_u, y_u]^T \in \mathbb{R}^{2 \times 1}, u \in \mathcal{U}$. All UAVs are assumed to fly at a fixed altitude $d_h = H$ above ground and the coordinate of UAV m at time t is denoted by $\psi_m(t) = [x_m(t), y_m(t)]^T \in \mathbb{R}^{2 \times 1}$. We consider there is a base controller carrying out the proposed learning process, which can be satellite or BS. In addition, the UAVs are able to communicate within the swam.

We consider all the UAVs will fly back to the base so the trajectories should satisfy the following constraint

$$\psi_m(1) = \psi_m(T). \tag{1}$$

In addition, the trajectories of the UAVs are also subjected to certain constraints of speed and distance, which are

$$\|\psi_m(t+1) - \psi_m(t)\| \le V_{max}, \tag{2}$$

$$\|\psi_m(t) - \psi_j(t)\| \ge S_{min}, \tag{3}$$

where V_{max} is the maximum speed of the UAV and S_{min} is the minimum inter-UAV distance to avoid certain interference or collision. Accordingly, the distance between UAV m and user u in time slot t is given as

$$d_{m,u}(t) = \sqrt{H^2 + \|\psi_m(t) - \psi_u\|^2}. \tag{4}$$

2.2 Path Loss Model

As a flexible flying platform, the UAV is able to establish a LoS link with the GUs. However, due to the fact that the changes of practical environment (rural, suburban, urban etc) are usually unpredictable, the randomness associated with the LoS and Non-LoS (NLoS) in a certain time should be taken into consideration when designing the UAV system. Accordingly, it is practical to consider the GU connects with the UAV via a LoS link with certain probability which we refer as LoS probability. The LoS probability will depend on the environment, the position of the UAV and GU. One commonly used expression can be given as

$$\rho_{m,u}^{los}(t) = \frac{1}{1 + \xi_1 \exp[-\xi_2(\theta_{m,u}(t) - \xi_1)]}, \tag{5}$$

where ξ_1 and ξ_2 are constant, the value of which the value depend on the carrier frequency and environment. $\theta_{m,u}(t)$ is the elevation angle, and we have

$$\theta_{m,u}(t) = \frac{180}{\pi \sin(H/d_{m,u}(t))}. \tag{6}$$

The LoS and NLoS path loss models between the UAV m and user u is given as

$$\hat{L}_{m,u}(t) = \begin{cases} \eta_1 (\frac{4\pi f_c d_{m,u}(t)}{c})^\alpha, & \text{LoS link,} \\ \eta_2 (\frac{4\pi f_c d_{m,u}(t)}{c})^\alpha, & \text{NLoS link,} \end{cases} \tag{7}$$

where η_1 and η_2 are the excessive coefficients in LoS and NLoS links, respectively. f_c is the carrier frequency, α is the path loss exponent, and c is the speed of light. Given the locations of the UAVs and GUs, it is difficult to determine whether a LoS or NLoS path loss model should be used in the considered UAV system. Thus, we consider an average over both the LoS and NLoS links, i.e.,

$$\begin{aligned} L_{m,u}(t) = {} & \rho_{m,u}^{los}(t)\eta_1 \Big(\frac{4\pi f_c d_{m,u}(t)}{c}\Big)^\alpha \\ & + (1 - \rho_{m,u}^{los}(t))\eta_2 \Big(\frac{4\pi f_c d_{m,u}(t)}{c}\Big)^\alpha. \end{aligned} \tag{8}$$

2.3 Transmission Model

To express the user association between UAVs and GUs, a binary variable $\beta_{m,u}(t)$ is defined as the user association indicator, which is

$$\beta_{m,u}(t) = \begin{cases} 1, & \text{if GU } u \text{ associates with UAV } m, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In this work, we assume that one GU can only receive from one UAV in a given time slot, i.e. $\sum_{m=1}^{M} \beta_{m,u}(t) \le 1$. In addition, The transmit power of the UAV m for u is denoted as $p_{m,u}(t)$ and the channel gain between UAV m and user u is denoted as $h_{m,u}(t)$. Then, the data rate of GU u is expressed as

$$R_u(t) = \sum_{m=1}^{M} \beta_{m,u}(t) \log_2 (1 + \gamma_{m,u}(t)), \tag{10}$$

In (10), due to the fact that multiple UAVs can cause interference to GU u, $\gamma_{m,u}(t)$ modelled as Signal to Interference and Noise Ratio (SINR) of the link between m and u, which is

$$\gamma_{m,u}(t) = \frac{p_{m,u}(t)h_{m,u}(t)L_{m,u}^{-1}(t)}{\sum_{j=1,j\neq m}^{M} p_{j,u}(t)h_{j,u}(t)L_{j,u}^{-1}(t) + \sigma^2}, \tag{11}$$

where σ^2 is the noise variance. Note that essentially the trajectory of the UAVs, transmit power and channel state are continuous. Then after partitioning and and quantizing their values into different levels within their ranges, in each time slot t, the values of these variables can be understood as discrete counterparts.

3 PROBLEM FORMULATION

3.1 Utility Function

As there are multiple UAVs sharing the same frequency resources, the transmit power of one UAV may bring additional interference to the users served by other UAVs, which is shown in (11). Moreover, the association factor and trajectory also affects the data rate of the users, which can be observed by (4) and (10). Therefore, in the following, we consider to jointly optimize these three parameters. We have $\mathbf{B} = \{\beta_{m,u}(t), \forall m, u, t\}$ which is the association policy between UAVs and GUs, $\mathbf{\Psi} = \{\psi_m(t), \forall m, t\}$ which is trajectory of the UAVs and essentially determines the path loss, and $\mathbf{P} = \{p_{m,u}(t), \forall m, u, t\}$ which is the transmit power allocation. Based on the analysis, we can define the utility function $\Upsilon_{sys}(\mathbf{B}, \mathbf{\Psi}, \mathbf{B})$ of the overall multi-UAV system as follow.

$$\Upsilon_{sys}(\mathbf{P}, \mathbf{\Psi}, \mathbf{B}) = \log \left(\sum_t^T \sum_u^U R_u(t) \right). \tag{12}$$

3.2 Problem formation

In order to maximize system utility, in this work, we jointly optimize transmit power allocation $\boldsymbol{P}$, trajectory design $\boldsymbol{\Psi}$, and user association $\boldsymbol{B}$. With the above analysis, the formulated problem $\mathbf{P1}$ can be expressed as follows,

$$\mathbf{P1}: \max_{\mathbf{P}, \boldsymbol{\Psi}, \mathbf{B}} \Upsilon_{sys}(\mathbf{P}, \boldsymbol{\Psi}, \mathbf{B}), \tag{13}$$

$$\begin{aligned} s.t. \quad &\mathbf{C1}: \quad \beta_{m,u}(t) \in \{0,1\}, \\ &\mathbf{C2}: \quad \sum_{m=1}^{M} \beta_{m,u}(t) \leq 1 \\ &\mathbf{C3}: \quad 0 \leq p_{m,u}(t) \leq p_m^{max}, \\ &\mathbf{C4}: \quad \psi_m(1) = \psi_m(T), \\ &\mathbf{C5}: \quad R_u(t) \geq R_u^{min}. \end{aligned} \tag{14}$$

C1 and **C2** are the user association constraints, which ensure that one GU can only be served by one UAV in a time slot. The maximum transmit power constraint is given in **C3**, which means that the transmit power of the UAV should be smaller than its maximum power. **C4**- **C5** is to ensure the minimum data requirement of each GU.

P1 is a non-convex combinatorial integer programming problem and it is NP-hard. In general, a brute-force-like scheme can be employed to find the optimal solution with high computational cost, which however, is infeasible for a large scale system. In addition, the optimization problem needs to obtain the complete information of the future in order to achieve the optimal solution for the next time slot, which means absence of prior information may degrade its achievable performance. Therefore, we intend to utilize the RL-based algorithm to achieve near-optimal solution without aforementioned prior knowledge.

4 CENTRALIZED DEEP REINFORCEMENT LEARNING-BASED SOLUTION

In this section, we will utilize the DRL-based algorithm to address the formulated problem. We first introduce the basics of DRL, including the defined specific state, action and reward. Then, the single agent DRL is utilized where the base controller acts as the agent and control the behaviors of the UAVs, and we refer the scheme as centralized DRL (CDRL).

4.1 RL Framework Formulation

The RL problem comprises of a single or multiple agents and an environment. The agent can take actions based on a chosen policy to interact with the environment. Briefly, there are three elements in the RL framework: action $\mathbf{a}$, state $\mathbf{s}$ and reward $\mathbf{r}$. In our considered system, the agent can be the UAV central base controller or UAV itself, and the environment consists of all the GUs. The agent chooses an action a_t from the action space at time slot t, which decides trajectory and resource allocation. After applying an action, the agent receives a reward or punishment from the environment. This scheme aims at maximizing the cumulative received rewards within interactions.

4.2 State, Action and Reward

We define state space, action space and reward of the DRL-based framework at time slot t of the considered system as follows. For the considered DRL framework, the decision will be carried out at the central base controller.

4.2.1 State

As for the centralized scheme, the central base should know all information about UAVs, e.g., association state, transmit power and trajectory state. The we define the state at the time slot t consists of the data rate R_t and battery level E_t. The battery level can decide the transmit power. The R_t comprises of both the channel state (essentially the location of UAV) and the UAV-GU association. Then the state at the time slot t is

$$s_t = [E_t^1, R_t^1, ..., E_t^M, R_t^M]. \tag{15}$$

4.2.2 Action

In the considered system, the action consists of multiple parts, i.e., the user association strategy $\mathbf{B}$, power allocation factors $\mathbf{P}$ and the trajectory design $\boldsymbol{\Psi}$. The action space $\mathcal{A}$ should be the combination of all the possible values of these factors.

4.2.3 Reward

After executing the chosen action, the agent will obtain a reward in certain state in each time slot. As shown in (17), to enforce the agent to take proper action, the definition of the reward is compulsory. In general, the reward should be related to the objective function. According to the formulated problem in **P1**, the objective is to maximize the overall system utility while the QoS of each GU should be satisfied. In order to transform the objective function to a reward, we consider the following points.

- The main objective of **P1** is to maximize the overall system utility. As the target of the RL is to obtain reward maximization, the defined reward needs to be positively related to the objective function.
- To meet the QoS requirements of GUs, the loss of the throughput of the GUs from their required QoS should decrease the reward.

Accordingly, the immediate reward is defined as

$$r(s_t, a_t) = \varphi_a \Upsilon_{sys} + \varphi_b \left(R_u\left(t\right) - R_u^{min} \right), \tag{16}$$

where φ_a and φ_b are the weights of two parts.

4.3 Q-Learning Method

Q-learning is one of the classical RL schemes that records the Q-value. In the considered system, the base controller first watches the state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ at each time slot t according to a stochastic policy π. Then the base controller transmits control signals to the UAVs, obtains the reward $r(s_t, a_t)$, and transitions to the next state s_{t+1}. Q-learning advocates a value function $Q(s_t, a_t)$ that is the expected cumulative future discounted reward at state s_t and chooses action a_t. Then, each pair of state-action has a value $Q(s_t, a_t)$ for time slot t. For each time slot, the base controller calculates $Q(s_t, a_t)$, the value of which is considered as a long-term reward and stores it in a Q-table. $Q(s_t, a_t)$ is expressed as:

$$Q(s_t, a_t) = \mathbb{E}\left[\hat{r}_t | s_t, a_t\right], \tag{17}$$

Algorithm 1 Q-learning method

1: Initialize $Q(s, a)$

2: **for** each episode **do**
3:    Initialize s of each UAV randomly.
4:    **for** each time **do**
5:       Choose an action a_t from all actions of state s_t
6:       Execute chosen a_t, observe reward and next state s_{t+1}
7:       $Q(s_t, a_t) \leftarrow \kappa r(s_t, a_t) + \kappa \xi \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) + (1 - \kappa)Q(s_t, a_t)$
8:       Let $s_t \leftarrow s_{t+1}$.
9:    **end for**
10: **end for**

Fig. 2. Flow of the proposed DQN-based scheme

where $\hat{r}_t = \sum_t^T \xi r(s_t, a_t)$. We define ϖ as the discount parameter and $0 \leq \varpi \leq 1$. Note that if ϖ tends to 0, the base controller mainly takes the immediate reward into consideration and if ϖ tends to 1, the future is the focus of the base controller. In each step, the value of $Q(s_t, a_t)$ is iterated. When the optimal policy $\pi(s_t) = \max_{a_t} Q(s_t, a_t)$ that maps the state and action is satisfied, the optimal action-value function $Q^*(s_t, a_t)$ is achieved. It obeys the Bellman optimality equation:

$$Q^*(s_t, a_t) = r(s_t, a_t) + \xi \max_{a_{t+1}} Q^*((s_{t+1}, a_{t+1})|s_t, a_t), \quad (18)$$

where a_{t+1} and s_{t+1} are the action and state of the next time slot, respectively. With a defined learning rate κ, the procedure of Q-learning scheme in shown in Algorithm 1, where ξ is the time-varying learning rate.

4.4 Proposed Centralized DRL Solution

Although simply advocating Q-learning may obtain the solution of **P1**, it is not ideal. When using Q-learning, we need to obtain and store its corresponding Q-value in a Q-table for each state-action group as presented. However, in the considered UAV system, due to its mobility feature, there is a very high possibility that a thousands of states exist. Then, if all the values are stored, the matrix of Q-table would be very large. Then it can be difficult to get enough samples to traverse each state, which results in the failures of the algorithm. Therefore, instead of calculating Q-value for each pair, a DNN is used here to estimate $Q(\mathbf{s}, \mathbf{a})$ which is the main idea of Deep Q-Network (DQN).

DQN uses a neural network (NN) $Q(\mathbf{s}, \mathbf{a}; \theta)$ to represent Q-function, where θ is the weights of the NN. By updating θ at each iteration, the Q-network is trained to approximate the real Q-values. When it is applied to Q-learning, NN improve the performance of flexibility at the cost of stability [4]. In this context, DNN is proven to be with a more robust learning scheme and it has three major improvements comparing with the Q-learning.

The first one is that DNN is with multiple layer. The hierarchical layers of convolution filters in the DNN can be used to exploit the local spatial correlations. By such, the high-level features of input data are extracted. The second one is that experience replay can store its experience tuple $e(t) = (s_t, a_t, r_t, s_{t+1})$ at time slot t into a replay memory $\mathcal{O}$. The relay can randomly sample batches $\hat{\mathcal{O}}$ from the memory
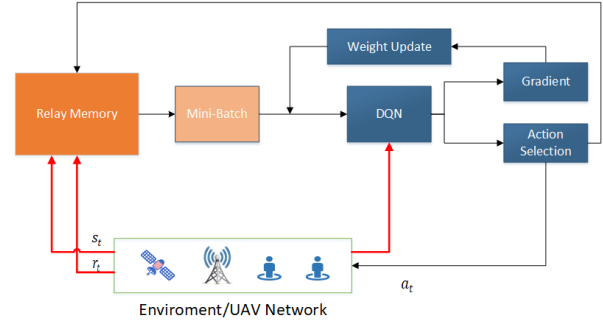
to train the DNN. Such a process enables DQN to learn from different past experience rather than from the current one. In addition, while using one network for estimating the Q-values, the target Q-values that compute the loss of each action in the training process can be generated by a second network. Such a procedure is able to make the DQN stable.

As presented, DQN uses NN with parameter θ to represent $Q(s_t, a_t)$ in each iteration. θ and policy π are updated according to the mini-batch of $\hat{\mathcal{O}}$ which is taken from experience memory $\mathcal{O}$ to train the DQN in a online manner. DQNs are optimized by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}[y_t - Q(s_t, a_t; \theta)^2] \quad (19)$$

where y_t is the target Q-value, and it is given as

$$y_t = r(s_t, a_t) + \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}; \theta^-). \quad (20)$$

In (20), θ^- is a target network parameter that is frozen for some iterations when the online network $-Q(\mathbf{s}, \mathbf{a}; \theta)$ is updated by gradient descent. Specially, the base controller chooses a_t at time slot t according to ((18)), obtains reward r_t and goes to the next state s_{t+1}. Accordingly, the base controller has a experience replay memory $\mathcal{O}$ to store the vector (s_t, a_t, r_t, s_{t+1}). We can utilize the greedy policy in order to balance the exploration and exploitation. That is, we aim to balance the reward maximization based on the known information with choosing new actions to get unknown information. Algorithm 2 presents the process and the flow is shown in Fig. 2.

5 MULTI-AGENT DEEP REINFORCEMENT LEARNING-BASED SOLUTION

The proposed CDRL-based scheme assumes that the UAV base actually performs the learning process and coordinate the actions of the entire UAV swam. However, on the way towards a smart UAV system, it is expected that the UAVs can be autonomous at a certain level. Thus, in the following, we focus on a setting with centralized learning but distributed execution towards establishing an autonomous UAV wireless communication system. Before we introduce the proposed scheme, some preliminaries are presented.

Algorithm 2 DQN-based online method

1: Initialize replay memory $\mathcal{O}$
2: Initialize parameter of the DNN θ with random weights

3: **for** each episode **do**
4:     Initialize the considered wireless UAV network
5:     Receive the initial observation on the state s_1.
6:     **for** each time slot t **do**
7:         Randomly select an action a_t with probability ϵ, otherwise, select $a_t = \arg\max_a Q(x, a; \theta)$.
8:         Execute chosen a_t, observe reward and s_{t+1}
9:         Store $(s_t, a_t; r_t, s_{t+1})$ in replay memory $\mathcal{O}$
10:        Sample a random batch of Z vectors $(s_i, a_i; r_i, s_{i+1})$ from $\mathcal{O}$
11:        Obtain the target Q-value y_i from the target DQN, as follows,

$$y_i = r_i + \xi \max_{a_{l+1}} Q(s_{i+1}, \arg\max_{a'} Q(s_{i+1}, a', \theta), \theta^-)$$

12:        Update the main DQN by minimizing the loss function $\mathcal{L}(\theta)$,

$$\mathcal{L}(\theta) = \frac{1}{Z} \sum_i (y_i - Q(s_i, a_i, \theta))^2).$$

13:        Perform a gradient descent step on $\mathcal{L}(\theta)$ with respect to θ.
14:     **end for**
15: **end for**
16: Output: the optimal resource allocation policy, i.e., the user association strategy $\mathbf{B}$, trajectory design $\mathbf{\Psi}$, and power allocation $\mathbf{P}$

5.1 Preliminary

5.1.1 Independent DQN

The single agent DQN can be extended to multi-agent cooperative settings. In this setting, the global state s_t can be observed by the agents. Then, the each agent chooses an individual action a_t^m and obtains a group reward r_t which is shared among all the agents. A platform combining independent Q-learning with DQN has been proposed . In this framework, each agent m learns its own Q-function $Q^m(s, a^m; \theta_i^m)$ independently and simultaneously. In [33], it is shown that there may be some convergence problems in independent Q-learning (since individual learning may result in non-stationary environment for the others). Nevertheless, it has been successfully applied to practical problems [33].

5.1.2 Deep Recurrent Q-Networks (DRQN)

For both DQN and independent DQN, it is assumed full observability, i.e., the global state s_t is the input. However, in practice, the dynamic environments are usually partially observable, i.e., the global state s_t cannot be observed. Instead, each of the agents can only obtain an observation o_t which is correlated with global state. In [34], the DRQN is proposed to address single-agent and partially observable case. In this work, instead of obtaining $Q(s, a)$ with a feed-forward network, $Q(o, a)$ is approximated with a recurrent

NN that maintains an internal state and aggregates all the personal observations over some time slots. This is done by adding a hidden state h_{t-1} as the input, and it results in $Q(o_t, h_{t-1}, a_t; \theta)$.

5.2 Assumption

In this case, we turn to investigate the formulated problem with different UAVs as multiple agents and partial observability is considered. The objective of maximizing the same discounted group rewards $r(t)$ are shared among all the UAVs. Although the global state s_t is not observable to the UAVs, each UAV m has its own observation o_t^m. In each time slot, each UAV selects an action $a^m \in \mathcal{A}$ that has impact on the environment and a communication action $\varsigma^m \in \Omega$ that is observed by other UAVs but does not directly affect the environment/reward. Such settings are of interests because usually in the multi-UAV system, partial observability is a practical case. We concentrate on the case with centralized learning and decentralized execution. This is to say, communications between UAVs and base controller is not limited during centralized learning while during execution the UAVs can communicate only via a dedicated signaling channel with limited-bandwidth. Then, during decentralized execution, each UAV uses its own copy of the learned network, evolving its own hidden state, selecting its own actions, and communicating with others only through the communication channel.

Towards an self-organized and autonomous system in a dynamic environment, the UAV must develop and agree on a communication protocol as the environment can change fast and the configured communication protocol may not work effectively .

Intuitively, the space dimension of communication protocols is extremely high, since they are the mappings from the histories of observation-action to sequences of communication signals over number of UAVs. Therefore, discovering an effective protocol is challenging. In addition, due to the UAVs' requirement of coordinating the transmission and decoding of communication messages, exploring within this space becomes more difficult. For example, if a UAV transmits something useful to another UAV, it can obtain a positive reward only when receiving UAV successfully decodes and takes action accordingly. If the receiving UAV cannot decode the message correctly, the sending UAV will be hindered from transmitting again. Therefore, positive rewards can be achieved *iff* transmitting and decoding are successful, which is difficult to be achieved via a random search.

5.3 Proposed decentralized solution

In this following, we propose the reinforced inter-UAV learning which combines independent Q-learning with DRQN to select environment and communications actions. Each UAV's Q-network is denoted as $Q^m(o_t^m, \varsigma_{t-1}^m, h_t^m, a^m)$, which conditions on that UAV's individual hidden state and observation. To avoid $|\Omega||\mathcal{A}|$ outputs, we divide the Q-network into Q_a^m for the environment action and Q_ς^m for the communication action, respectively. By utilizing ϵ-greedy policy, the action selector separately picks $a^m(t)$ and $\varsigma^m(t)$ from Q_a and Q_ς, respectively. Correspondingly, only
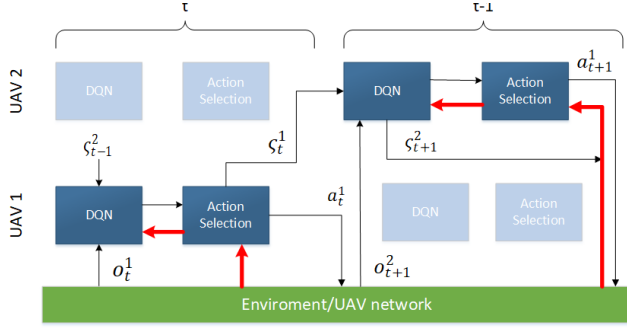
Fig. 3. The proposed reinforced inter-agent learning scheme

TABLE 1
Key Simulation Parameters

Notations	Description	Value
f_c	carrier frequency	2 GHz
α	Path loss exponent	2
M	number of the UAVs	$2-9$
U	number of the GUs	$10-50$
ϖ	learning rate	0.01
ξ	Discount factor	0.8
η_1	excessive coefficient for LoS	3 dB
η_2	excessive coefficient for NLoS	23 dB
P_m^{max}	maximum transmit power	23 dBm

$|\Omega|+|\mathcal{A}|$ outputs are required for the network and the action selection requires maximizing over $\mathcal{A}$ and Ω, but not $\Omega \times \mathcal{A}$.

We use modified DQN to train Q_a^m and Q_ς^m. The following two essential modifications are made to the DQN to guarantee the performance. First, as multiple UAVs' simultaneous learning can mislead the experience and render it obsolete, the experience replay is disabled to avoid non-stationarity. Second, to take into consideration of the partial observability, the actions $\mathbf{a}$ and ς of each UAV are feed in as the inputs of the next time slot. In Fig. 3, the information flows between UAVs and the network are presented together with how the action selector can process the Q-values to find proper actions. As shown, in order to choose environment action a^m and communication action ς^m, all Q-values are passed to the action selector. For the selected actions, the gradients (red arrows in the figure) are calculated using DQN, and flow only through one single UAV's Q-network. Although the considered setting allows a centralized learning, as the each UAV is treated independently, the overall process is not a centralized learning procedure. In addition, all the UAVs are equally treated during the proposed decentralized execution process.

The proposed scheme can be extended to improve the centralized learning by parameter sharing among the UAVs. Such an extension only needs to learn one network and then used by all UAVs. However, because each UAV still has different observation, the UAVs can still behave differently and thus go to different hidden states. Moreover, each UAV obtains own index as input which allows them to specialize. The DQN is able to ease the learning process of a common policy while permitting the specialization. Sharing the parameters among all the UAVs also significantly decreases the amount of parameters that needs to be learned, which can also hasten the speed of learning. By sharing the parameters, the UAVs learn two Q-functions $Q_a(o_t^m, \varsigma_{t-1}^{m'}, h_{t-1}^m, a_{t-1}^m, \varsigma_t^{m'}, m, a_t^m)$ and $Q_\varsigma()$, for $\mathbf{a}$ and ς, respectively, where a_{t-1}^m and ς_{t-1}^m are the last action inputs and $\varsigma_t^{m'}$ are messages from other UAVs. During the execution process, each UAV uses own copy of the learned network, chooses own actions, evolves into own hidden state, and communicates with the others via the signalling channel.

6 SIMULATION RESULTS AND DISCUSSIONS

In this section, simulations are conducted to verify the advantages of the proposed single agent (CDRL) and multi-
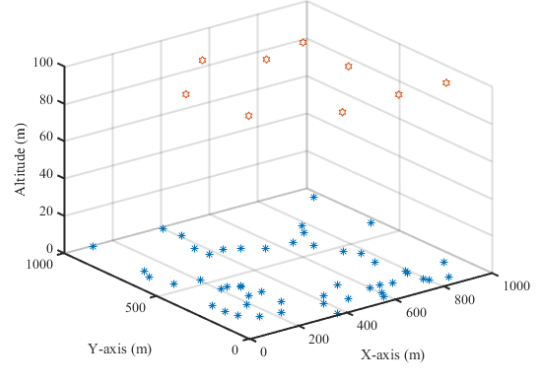

Fig. 4. Locations of UAVs and GUs in a 3-D snapshot

agent DRL-based (MADRL) resource allocation schemes for multi-UAV networks. The setup of whole networks are mainly based on the parameters in [16], [25]. Some of the key notations for communications can be found Table 1. The initial locations of the UAVs are randomized. The maximum transmit power of each UAV is the same. Based on this setting, the system utility, 3-D trajectory design and UAV-GU association are analyzed.

The 3-D and 2-D snapshots of the UAVs' locations and their associated GUs resulting from the proposed scheme are presented in Figs. 4 and 5. In both figures, 50 GUs are uniformly located and 9 UAVs are deployed to provide services. In Fig. 5, the 2D locations of UAV are marked in number. In this case, all GUs are able to connected with the UAVs and receive data from the associated UAVs by using the proposed scheme. The 3-D locations/trajectory of the UAVs and the UAV-GU association results are obtained based on the locations of the GUs and its minimum data rate requirement.

In Fig. 6, the optimized trajectories of the UAVs are illustrated. In Fig. 6a, we plot the trajectory of four UAVs by using the proposed MADRL scheme, while in Fig. 6b, the trajectory of one UAV is obtained by using the proposed CDRL scheme. It is observed that for the case of four UAVs, most of the users can be served by the UAVs. However, due to the limited battery capability, there are still some of users cannot be served by the UAVs. It can also be found that four UAVs can cooperate with each others through the proposed multi-agent learning scheme, and the users can be associated with individual UAV accordingly. As for the case of single UAV, due to the limited battery capability, the UAV has to come back after serving some of the users. Thus, only
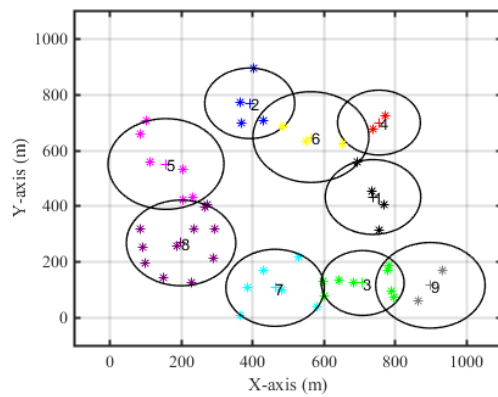
Fig. 5. Association of UAVs and GUs in a 2D snapshot

some of the users can be associated with the UAV.

In Fig. 7, we present the total utility versus the number of episodes with different number of UAVs when considering MADRL. As shown in the figure, our presented scheme shows a fast convergence speed for all of the cases. Besides, increasing the number of UAVs can lead to the increase of system utility. In Fig. 8, we present the total utility versus the number of episodes with different number of UAVs when considering CDRL. We can obtain similar performance as presented in Fig. 7. Nevertheless, for CDRL, when the number of UAVs becomes larger, it takes a bit longer time to converge. This may due to the fact that the CDRL needs to collect relative information in a centralized manner, which cost more time.

In Fig. 9 and in Fig. 10, we compare the throughput and utility performance of traditional Q-learning scheme, the proposed CDRL and the proposed MADRL. As we can observe from Fig. 9, as the number of UAVs increases, the total throughput of all these three schemes become larger. This is mainly due to the fact that the increase of the number of UAVs results in a better service coverage, and can provide better data services to the GUs. Similar situation can be observed from Fig. 10 when we investigate the utility performance. In addition, we can also find that both of the proposed schemes outperform the traditional Q-learning scheme, the centralized scheme obtain the best performance. This is mainly due to the fact that when the central controller can obtain all the relevant information, such as CSI and position of UAV, it can carry out more accurate decision via deep learning schemes. Nevertheless, the MADRL has a close performance to the CDRL, which demonstrates its effectiveness.

We have compared the proposed CDRL with two commonly-used baseline methods, "Benchmark" and "TRRA". The "Benchmark" is the random UAV deployment scheme where the whole area is equally separated to a number of parts according to the number of UAVs. Then each UAV has its responded area, and then randomly flies within each area and serve the GUs. The "TRRA" refers to the traditional RRA scheme, where the power allocation is according to the waterfilling scheme and the association ignores the minimum data requirement. From Fig. 11, it is found that the system utilities of all three schemes increase

with the number of UAVs. This is due to the fact that a larger number of UAVs can ensure more GUs being served with required data rate. Moreover, when the number of UAVs is sufficiently large, it turns out that there are less GUs who cannot be served and the increase of system utility becomes slow. It can also be observed the proposed scheme can obtain the best performance among all three, which shows the importance of adopting DRL and the development of power allocation and UAV association schemes.

7   CONCLUSION

In this work, to establish a smart and autonomous multi-UAV wireless communication system, novel DRL-based trajectory design and resource allocation schemes are presented. In the considered system, the UAVs act as aerial Base Stations and provide ubiquitous coverages. Specifically, aiming at maximizing the defined system utility over all served GUs, a joint design of trajectory, user association and power allocation problem is presented. To address the formulated problem, we first propose a machine learning-based algorithm which comprises of reinforcement learning and deep learning to learn the optimal policy of all the UAVs. Then, we also present a multi-agent deep reinforcement learning scheme for decentralized implementation without knowing a priori knowledge of the dynamics of networks. Extensive simulation studies are conducted to demonstrate advantages of the proposed schemes are demonstrated. Future work is to improve the multi-UAV system performance via energy efficiency and delay optimization in the proposed framework.

REFERENCES

[1] Z. Zhou, J. Feng, C. Zhang, Z. Chang, Y. Zhang and K. Huq, "SAGE-CELL: Software-defined space-air-ground integrated moving cells," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 92-99, Aug. 2018

[2] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: recent advances and future trends", *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2241-2263, 2019.

[3] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569-572, Dec. 2014.

[4] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[5] L. Panait, and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 3, pp. 387-434, Nov. 2005.

[6] X. Wang, H. Zhang, Y. Tian, and V. C. M. Leung, "Modeling and analysis of aerial base station-assisted cellular networks in finite areas under LoS and NLoS propagation", *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6985-7000, Oct. 2018.

[7] C. She, C. Liu, T. Q. S. Quek, C. Yang, and Y. Li, "Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems", *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3768-3781, May 2019.

[8] C. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach", *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059-2070, 2018.

[9] Z. Chang, W. Guo, X. Guo and T. Ristaniemi, "Machine Learning-based Resource Allocation for Multi-UAV Communications System ," *IEEE ICC workshop*, June 2020.

[10] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective", *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2580-2604, Mar. 2019.
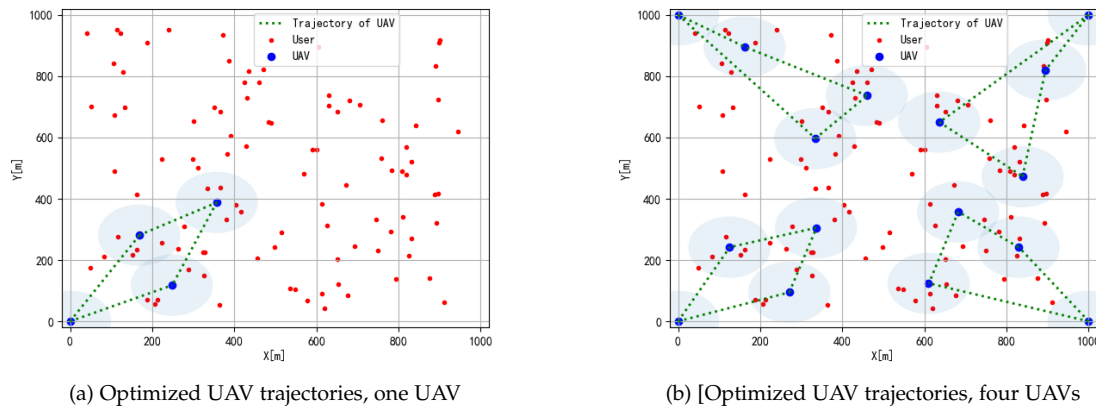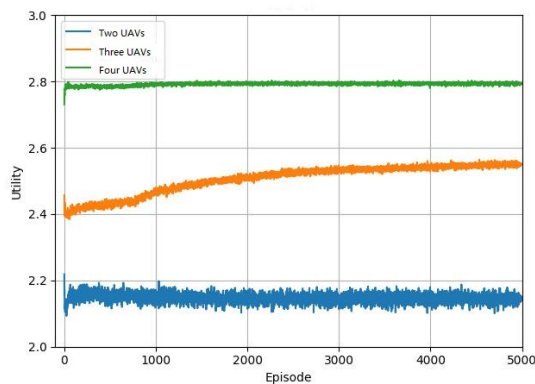
(a) Optimized UAV trajectories, one UAV

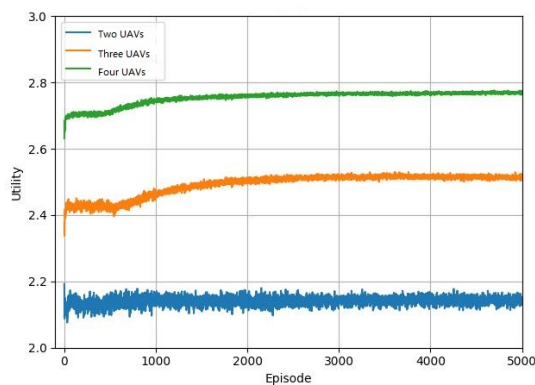(b) [Optimized UAV trajectories, four UAVs

Fig. 6. Optimized UAV trajectories

Fig. 7. Total utility versus the number of episodes, MADRL
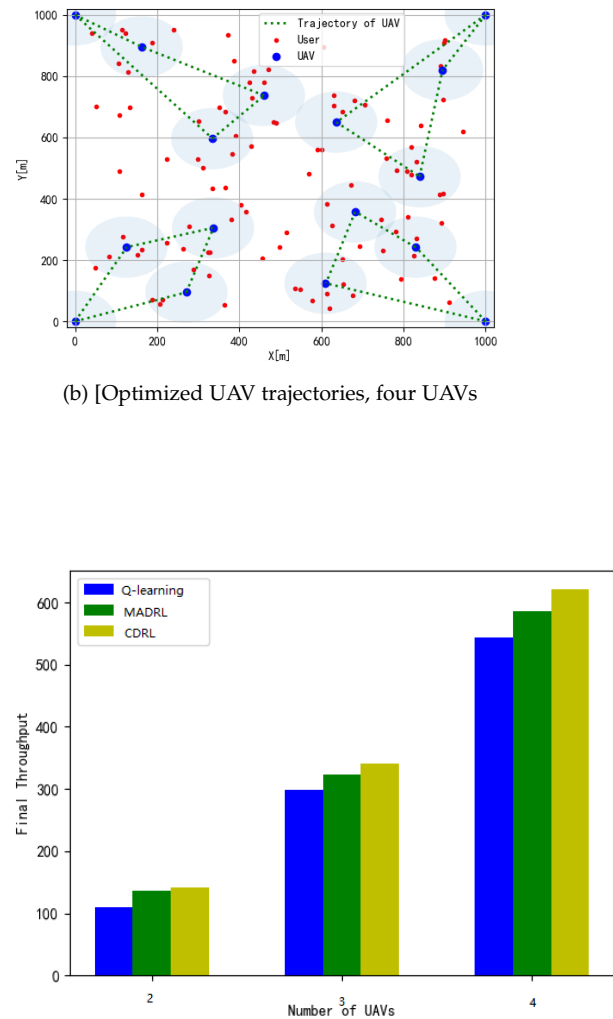

Fig. 9. The impact of the number of UAVs on system throughput

Fig. 8. Total utility versus the number of episodes, CDRL

[11] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks", *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1346-1359, 2019.

[12] Y. Cai, F. Cui, Q. Shi, M. Zhao, and G. Y. Li, "Dual-UAV-enabled secure communications: Joint trajectory design and user scheduling," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 1972-1985, Sept. 2018.

[13] J. Zhang, Y. Zeng and R. Zhang, "UAV-Enabled radio access network: multi-mode communication and trajectory design," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5269-5284, 15 Oct.15, 2018.

[14] Q. Wu, Y. Zeng and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109-2121, March 2018.

[15] H. Wang, G. Ren, J. Chen, G. Ding, and Y. Yang, "Unmanned aerial vehicle-aided communications: Joint transmit power and trajectory optimization," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 522-525, Aug. 2018.

[16] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016.

[17] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7574-7589, Nov. 2017.

[18] Z. Yang et al., "Joint altitude, beamwidth, location, and bandwidth optimization for UAV-enabled communications," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1716-1719, Aug. 2018.

[19] X. Yuan, Y. Hu, D. Li and A. Schmeink, "Novel optimal trajectory design in UAV-assisted networks: A mechanical equivalence-based strategy," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3524-3541, Nov. 2021ïïjŇ

[20] T. Chu, J. Wang, L. Codeca, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, to appear.

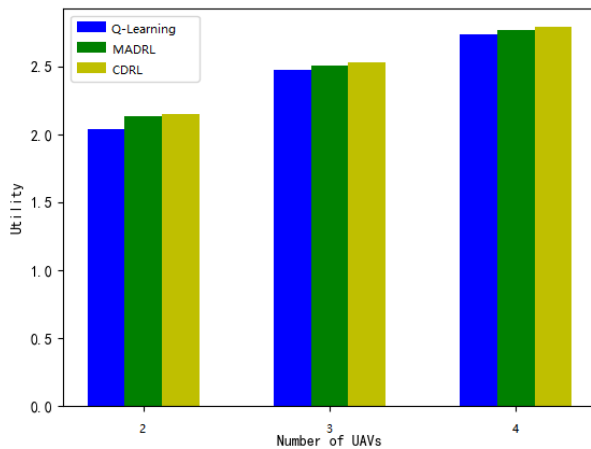[21] E. A. O. Diallo, A. Sugiyama, and T. Sugawara, "Learning to coor-

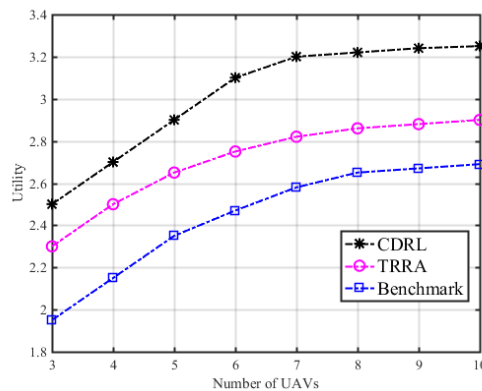Fig. 10. The impact of the number of UAVs on system utility

Fig. 11. The impact of the number of UAVs on system utility

[29] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *EEE Transactions on Vehicular Technology* vol. 67, no. 5, pp. 4539-4551, May 2018.
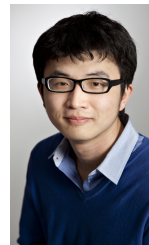[30] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Transactions on Wireless Communications* vol. 18, no. 1, pp. 310-323, Jan. 2019.
[31] J. Hu, H. Zhang, L. Song, R. Schober and H. V. Poor, "Cooperative Internet of UAVs: Distributed Trajectory Design by Multi-Agent Deep Reinforcement Learning," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6807-6821, Nov. 2020.
[32] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam and L. Hanzo, "Multi-Agent Deep Reinforcement Learning-Based Trajectory Planning for Multi-UAV Assisted Mobile Edge Computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 73-84, March 2021.
[33] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PLoS ONE*, vol. 12, no. 4, 2017.
[34] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," *Association for the Advancement of Artificial Intelligence Fall Symp. Series*, 2015.
[22] J. K. M. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," *2017 International Conference on Autonomous Agents and Multiagent Systems*, pp 66-83, 2017.
[23] J. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning." *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.
[24] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3996-4012, Sept. 2016.
[25] X. Li, J. Li, Y. Liu, Z. Ding, A. and Nallanathan, "Residual Transceiver Hardware Impairments on Cooperative NOMA Networks",Âǎ*IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 680-695, Jan. 2020.
[26] X. Li, M. Zhao, M. Zeng, S. Mumtaz, V.G. Menon, Z. Ding, and O. A. Dobre, "Hardware Impaired Ambient Backscatter NOMA System: Reliability and Security,"Âǎ*IEEE Transactions Communications*, vol. 69, no. 4, pp. 2723-2736, Apr. 2021.
[27] X. Huang, S. Leng, S. Maharjan and Y. Zhang, "Multi-Agent Deep Reinforcement Learning for Computation Offloading and Interference Coordination in Small Cell Networks", *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp.9282-9293, Sept. 2021.
[28] F. Yao and L. Jia, "A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks," *IEEE Wireless Communications Letters*, in press.

dinate with deep reinforcement learning in doubles pong game," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, 2017, pp. 14-19.

Zheng Chang (S'10-M'13-SM'17) received the B.Eng. degree from Jilin University, Changchun, China in 2007, M.Sc. (Tech.) degree from Helsinki University of Technology (Now Aalto University), Espoo, Finland in 2009 and Ph.D degree from the University of Jyväskylä, Jyväskylä, Finland in 2013. Since 2008, he has held various research positions at Helsinki University of Technology, University of Jyväskylä and Magister Solutions Ltd in Finland. He has been awarded by the Ulla Tuominen Foundation, the Nokia Foundation and the Riitta and Jorma J. Takanen Foundation for his research excellence. He has been awarded as 2018 IEEE Communications Society best young researcher for Europe, Middle East and Africa Region.

He has published over 140 papers in Journals and Conferences, and received received best paper awards from IEEE TCGCC and APCC in 2017. He serves as an editor of IEEE Wireless Communications Letters, Springer Wireless Networks and International Journal of Distributed Sensor Networks, and a guest editor for IEEE Network, IEEE Wireless Communications, IEEE Communications Magazine, IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics, etc. He was the exemplary reviewer of IEEE Wireless Communication Letters in 2018. He has participated in organizing workshops and special sessions in Globecom'19, WCNC'18-22, SPAWC'19 and ISWCS'18. He also serves as Symposium Chair for ICC'20 and Publicity Chair for INFOCOM'22. His research interests include IoT, cloud/edge computing, security and privacy, vehicular networks, and green communications.

Hengwei Deng is a Master Student at UESTC. His research interests include machine learning, UAV, cloud computing and mobile computing.

Li You (Senior Member, IEEE) received the B.E. and M.E. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree from Southeast University, Nanjing, in 2016, all in electrical engineering.

From 2014 to 2015, he conducted Visiting Research at the Center for Pervasive Communications and Computing, University of California Irvine, Irvine, CA, USA. Since 2016, he has been with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests lie in the general areas of communications, signal processing, and information theory, with the current emphasis on massive MIMO communications.

Dr. You received the National Excellent Doctoral Dissertation Award from the China Institute of Communications (CIC) in 2017, the Young Elite Scientists Sponsorship Program (2019–2021) by the China Association for Science and Technology (CAST), and the URSI Young Scientist Award in 2021.

Sahil Garg received the Ph.D. degree from the Thapar Institute of Engineering and Technology, Patiala, India, in 2018. He is currently a Research Associate at Resilient Machine Learning Institute (ReMI) co-located with ĂĽcole de Technologie SupĂĽrieure (ĂĽTS), MontrĂĽal. Prior to this, he worked as a Postdoctoral Research Fellow at ĂĽTS, Montreal and MITACS Researcher at Ericsson, Montreal. He has many research contributions in the area of Machine Learning, Big Data Analytics, Knowledge Discovery, Cloud Computing, Internet of Things, Software Defined Networking, and Vehicular Ad-hoc Networks. He has over 80 publications in high ranked Journals and Conferences, including 50+ top-tier journal papers and 30+ reputed conference articles. He was the recipient of the prestigious Visvesvaraya PhD fellowship from the Ministry of Electronics & Information Technology under Government of India (2016-2018). He has been awarded the 2021 IEEE Systems Journal Best Paper Award; the 2020 IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researcher) and the IEEE ICC best paper award in 2018 at Kansas City, Missouri. He is currently a Managing Editor of Springer's Human-centric Computing and Information Sciences (HCIS) journal; and an Associate Editor of IEEE Network Magazine, IEEE Transactions on Intelligent Transportation Systems, Elsevier's Applied Soft Computing (ASoC), and Wiley's International Journal of Communication Systems (IJCS). In addition, he also serves as the Workshops and Symposia Officer for the IEEE ComSoc ETI on Aerial Communications. He guest edited a number of special issues in top-cited journals including IEEE T-ITS, IEEE TII, IEEE TNSE, IEEE IoT Journal, IEEE Network Magazine, FGCS, Computer Networks, NCAA, etc. He also served as the TPC Co-Chair/Publicity Co-chair/Special Sessions Chair/Publication Chair for several conferences. He also served as the workshop co-chair for different workshops in IEEE /ACM conferences including IEEE Infocom, IEEE Globecom, ACM MobiCom, etc. Moreover, he is also serving as a Symposium Chair for Aerial Communications track in IEEE ICC 2022 to be held at Seoul, South Korea. He is a member of IEEE, IEEE Communications Society, IEEE Industrial Electronics Society, IEEE Software Defined Networks Community, IEEE Smart Grid Community, ACM, and IAENG.

Geyong Min received the B.Sc. degree in computer science from Huazhong University of Science and Technology, China, in 1995, and the Ph.D. degree in computing science from the University of Glasgow, U.K., in 2003. He is currently a Professor of high-performance computing and networking with the Department of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, U.K. His research interests include future internet, computer networks, wireless communications, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering. He has served as a Guest Editor for numerous international journals, such as ACM Transactions on Intelligent Systems and Technology, IEEE Transactions on Sustainable Computing, and ACM Transactions on Embedded Computing.

Georges Kaddoum (Member, IEEE) received the BachelorâĂŹs degree in electrical engineering from the ĂĽcole Nationale SupĂĽrieure de Techniques AvancĂĽes (ENSTA Bretagne), Brest, France, in 2004 and the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the UniversitĂĽ de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005 and the Ph.D. degree (with honors) in signal processing and telecommunications from the National Institute of Applied Sciences (INSA), University of Toulouse, Toulouse, France, in 2009. He is currently an Associate Professor and Tier 2 Canada Research Chair with the ĂĽcole de Technologie SupĂĽrieure (ĂĽTS), UniversitĂĽ du QuĂĽbec, MontrĂĽal, Canada. In 2014, he was awarded the ĂĽTS Research Chair in physical-layer security for wireless networks. Since 2010, he has been a Scientific Consultant in the field of space and wireless telecommunications for several US and Canadian companies. He has published over 200+ journal and conference papers and has two pending patents. His recent research activities cover mobile communication systems, modulations, security, and space communications, and navigation. Dr. Kaddoum received the Best Papers Awards at the 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications (WIMOB), with three coauthors, and at the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), with four co-authors. Moreover, he received IEEE Transactions on Communications Exemplary Reviewer Award for the years 2015, 2017, 2019. In addition, he received the research excellence award of the UniversitĂĽ du QuĂĽbec in the year 2018. In the year 2019, he received the research excellence award from the ĂĽTS in recognition of his outstanding research outcomes. Prof. Kaddoum is currently serving as an Associate Editor for IEEE Transactions on Information Forensics and Security, and IEEE Communications Letters.