Multi-Agent Deep Reinforcement Learning for Full-Duplex Multi-UAV Networks

Chen Dai*†‡, Kun Zhu*†, and Ekram Hossain‡
* Nanjing University of Aeronautics and Astronautics
† Collaborative Innovation Center of Novel Software Technology and Industrialization
‡ University of Manitoba, Canada
Emails: {daichen, zhukun}@nuaa.edu.cn, Ekram.Hossain@umanitoba.ca

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771902

***Abstract*—We study the joint decoupled uplink (UL)-downlink (DL) association and trajectory design problem for full-duplex multi-UAV networks. A joint optimization problem is formulated aiming to maximize the sum-rate of user equipments (UEs) in both UL and DL. Since the formulated problem is non-convex and with sophisticated states, a multi-agent deep reinforcement learning (MADRL) approach is employed for enabling each agent (i.e., UAV) to select policy in a distributed manner. Moreover, in order to obtain the optimal policy, a clip-and-count based proximal policy optimization (PPO) algorithm is proposed to train actor-critic neural networks. In particular, a modified clip distribution is designed to deal with the hard restrictions between current and old policies, and an intrinsic reward is introduced to enhance the exploration capability. Simulation results demonstrate the significant performance improvement of our proposed schemes when compared to the benchmarks.**

***Index Terms*—Unmanned aerial vehicle (UAV), full-duplex, decoupled user association, trajectory design, multi-agent deep reinforcement learning (MADRL), proximal policy optimization (PPO).**

I. Introduction

With high maneuverability and low operating expenses, unmanned aerial vehicles (UAVs) are able to provide high-quality services, especially for temporarily wireless coverage (e.g., rescue operations) [1], [2]. If line-of-sight (LoS) links can be established between UAVs and user equipments (UEs), the available network throughput and coverage can be improved. Also, in-band full-duplex (FD) communications, which allow simultaneous transmission and reception of wireless signals in the same frequency band, can be introduced to further improve the performance of UAV networks [3], thus improving the system spectral efficiency.

Due to the UAV mobility with varying speeds, the transmission links between UAVs and UEs can vary dynamically, which can result in dynamic network topologies. However, most of the existing studies on UE-UAV association assume that a UE associates to the same UAV for uplink (UL) and downlink (DL) transmissions [4]. This association mode may not be optimal for multi-UAV networks with mobility and varying network topology. Therefore, we consider the idea of DL-UL decoupling (DUDe) for the UE-UAV association [5], which allows a UE to associate with different UAVs in DL and UL. A simulation study in [5] revealed that UL-DL decoupling mechanism brought a significant performance improvement for two-tier cellular networks. Similarly, the authors in [6] studied how to enable a UE to associate with a ground BS in UL and a UAV in DL for a MmWave UAV network. However, this work did not consider the dyanmics of UL-to-DL and DL-to-UL interferences when DL-UL decoupling and trajectory design for the UAVs are jointly considered.

In this work, we investigate decoupled UL-DL user association for a full-duplex multi-UAV network jointly with the trajectory design of the UAVs. A joint optimization problem for DUDe association, power allocation and trajectory design is formulated. Considering the nonlinear objective function and non-convex constraints, we utilize the framework of distributed multi-agent deep reinforcement learning (MADRL) to solve the joint problem. Moreover, in order to obtain the optimal strategy, we propose a clip-and-count based PPO algorithm to train actor-critic neural networks. Simulation results demonstrate the significant performance improvement of our proposed schemes when compared to the benchmarks.

The remainder of this paper is structured as follows. The system model is described in Section II, followed by the joint optimization problem formulation. In Section III, the MADRL approach is investigated, and a PPO algorithm is developed. Numerical results are presented in Section IV. Finally, conclusions are drawn in Section V.

II. System Model, Assumptions, and Problem Formulation

A. Network Model

We consider a heterogeneous network that consists one macro BS (MBS) on the ground and multiple UAVs in the sky as aerial BSs. The sets of UEs and BSs (including the MBS and UAVs) are denoted as $\mathcal{N} = \{1, 2, 3, ..., N\}$ and $\mathcal{M} = \{0, 1, 2, 3, ..., M\}$, respectively. The MBS is indexed by 0 and the UAVs by $1, 2, ..., M$. As depicted in Fig. 1, each UE can be associated with different BSs in UL and DL. The UAVs connect to the MBS via wireless backhaul links, which are operated on different frequency bands than the links of BS-UE association. The sequential association decision evolves over continuous time slots, and each time slot t has a constant duration ΔT. We assume that the flight-time of the UAVs can be well estimated based on the amount of onboard energy.

The location of the MBS is assumed to be in the origin, and the UAVs and UEs are randomly located. Let $l_n(t) = (x_n(t), y_n(t))$ and $l_m(t) = (x_m(t), y_m(t), h(t))$ denote the

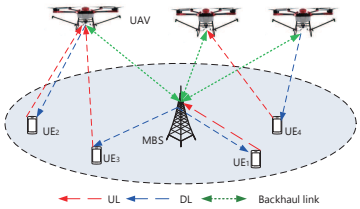

Fig. 1. A graphical illustration for decoupled UL-DL association in a full-duplex network consisting one MBS and multiple UAVs. For clarity, four possible association of BS-UE in the UL and DL are shown for UE_1, UE_2, UE_3 and UE_4. In UL or DL, each UE is allocated with one channel of bandwidth B during a time slot.

coordinates of UE n and UAV m at time t, respectively. For clarity of derivations, we assume that all UAVs fly at a fixed altitude h. The height of the MBS's antenna can be negligible compared with the altitude of the UAVs. The Euclidean distance of UE n to UAV m at time t, $d_{mn}(t)$, is given by

$$d_{mn}(t) = \|(x_m(t) - x_n(t))^2 + (y_m(t) - y_n(t))^2 + h^2\|_2. \quad (1)$$

B. Channel Model

The air-to-ground communication link is viewed as either line-of-sight (LoS) or non-line-of-sight (NLoS), which highly depends on the network environment. Following [7], the LoS probability of a link between UE n to UAV m at height h in different network environments, is given by

$$P^{LoS}(t) = [1 + c_1 \exp(-c_2(\vartheta_{mn}(t) - c_1))]^{-1}, \quad (2)$$

where c_1 and c_2 are environment-related constants (for rural, urban, etc.) [**?**]. $\vartheta_{mn}(t) = \frac{180}{\pi} \arcsin(\frac{h}{d_{mn}(t)})$ is the elevation angle. The probability of NLoS link is $P^{NLoS}(t) = 1 - P^{LoS}(t)$.

The LoS and NLoS path losses between UAV m and UE n can be expressed as [8]

$$PL_{mn}^{LoS}(t) = \alpha_{mn}(t)\beta^{LoS}, \quad (3)$$

$$PL_{mn}^{NLoS}(t) = \alpha_{mn}(t)\beta^{NLoS}, \quad (4)$$

where β^{LoS} and β^{NLoS} are two different attenuation factors to differentiate LoS and NLoS links. $\alpha_{mn}(t) = (\frac{4\pi f_c}{c} d_{mn}(t))^2$ is a quantified power gain according to the Friis transmission equation [7], where f_c is the carrier frequency and c is the speed of light. Therefore, the average path loss for a communication link can be expressed as

$$PL_{mn}(t) = P^{LoS}(t)PL_{mn}^{LoS}(t) + P^{NLoS}(t)PL_{mn}^{NLoS}(t). \quad (5)$$

C. User Association and Interference Model

1) Decoupled UL-DL Association: We define a binary variable $b_{mn}^{(\cdot)}$ to indicate whether UE n associates to BS m at a certain channel, which is expressed as

$$b_{mn}^{(\cdot)} = \begin{cases} 1, & \text{if UE } n \text{ is served by BS } m \text{ at a certain channel,} \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $(\cdot) = U$ denotes the uplink and $(\cdot) = D$ denotes the downlink. Note that, $\sum_{m=0}^{M} b_{mn}^{U} = 1$ and $\sum_{m=0}^{M} b_{mn}^{D} = 1$, $\forall n \in \mathcal{N}$, are required to guarantee that each UE can associate with only one BS either in UL or DL.

2) Interference Analysis for DUDe in FD Mode: In this case, the UL and DL association decisions are independent. We particularly consider UE n associating with BS μ in the UL, but BS v in the DL ($\mu, v \in \mathcal{M}$).

- Interferences received at BS μ associated by UE n in UL:

$$I_{\mu}^{U} = I_{\mu'\mu}^{D} + I_{n'\mu}^{U} + I_{\mu}^{self}, \quad (7)$$

where

$$I_{\mu'\mu}^{D} = \sum_{\substack{n' \in \mathcal{N}\setminus\{n\} \\ \mu' \in \mathcal{M}\setminus\{\mu\}}} b_{n'\mu'}^{D} \cdot p_{\mu'} \cdot PL_{\mu'\mu}, \quad (8)$$

$$I_{n'\mu}^{U} = \sum_{\substack{n' \in \mathcal{N}\setminus\{n\} \\ \mu' \in \mathcal{M}\setminus\{\mu\}}} b_{n'\mu'}^{U} \cdot p_{n'} \cdot PL_{n'\mu}. \quad (9)$$

Besides, I_{μ}^{self} denotes the residual self-interference of BS μ due to simultaneous UL and DL data transmissions, which can be calculated by $I_{\mu}^{self} = \frac{p_\mu}{\delta}$, where p_μ is its transmit power, and δ is the self-interference cancellation (SIC) exponent. For ease of analysis, δ is assumed to be a constant value in this paper [9].

- Interferences received at UE n associated to BS v in DL:

$$I_{n}^{D} = I_{v'n}^{D} + I_{n'n}^{U} + I_{n}^{self}, \quad (10)$$

where

$$I_{v'n}^{D} = \sum_{\substack{n' \in \mathcal{N}\setminus\{n\} \\ v' \in \mathcal{M}\setminus\{v\}}} b_{n'm'}^{D} \cdot p_{v'} \cdot PL_{v'n}, \quad (11)$$

$$I_{n'n}^{U} = \sum_{\substack{n' \in \mathcal{N}\setminus\{n\} \\ v' \in \mathcal{M}\setminus\{v\}}} b_{n'v'}^{U} \cdot p_{n'} \cdot PL_{n'n}. \quad (12)$$

D. Transmission Rate

The network performance is measured by the data transmission rate between the UE and the BS. In particular, for a given BS m, the data transmission rate for UE n at time t can be calculated as:

$$\Phi_{mn}^{(\cdot)}(t) = b_{mn}^{(\cdot)}(t) B \log_2 (1 + \text{SINR}_{mn}^{(\cdot)}(t)), \quad (13)$$

where $\text{SINR}_{mn}^{(\cdot)}(t)$ is the signal to interference plus noise ratio (SINR). For decoupled UL-DL association, the SINR can be calculated by $\text{SINR}_{mn}^{U}(t) = \frac{p_n(t)PL_{mn}(t)}{I_m^U + \sigma^2}$ and $\text{SINR}_{mn}^{D}(t) = \frac{p_m(t)PL_{mn}(t)}{I_n^D + \sigma^2}$, where σ^2 is the additive noise power, which is assumed to be identical for all UEs.

For backhauling, orthogonal frequency channels are employed to avoid self-interference in the side of UAV [8]. Hence, the backhaul rate of UAV m can be written as

$$\Phi_{m}^{back}(t) = B^{back} \log_2 (1 + \frac{p_o(t)PL_{0m}(t)}{\sigma^2}), \quad (14)$$

where B^{back} is the backhaul channel bandwidth. Note that the

number of UEs that a UAV can associate with is constrained by the achieved backhaul rate of the UAV.

E. Problem Formulation

Given the above system model, we formulate a joint decoupled association and trajectory design problem. The objective is to maximize the average data rate of the UEs in a period T by appropriately associating with BSs in UL and DL, allocating power and designing UAVs' trajectories. Accordingly, an optimization problem is formulated as follows:

$$\max_{\boldsymbol{b},\boldsymbol{L},\boldsymbol{p}} \quad \frac{1}{T}\Big(\sum_{t=1}^{T}\sum_{n=1}^{N}\Phi_{mn}^{U}(t)+\sum_{t=1}^{T}\sum_{n=1}^{N}\Phi_{mn}^{D}(t)\Big) \tag{15a}$$

$$\text{s.t.} \quad \sum_{m=0}^{M} b_{mn}^{U}(t)=1, \sum_{m=0}^{M} b_{mn}^{D}(t)=1, \forall n, \tag{15b}$$

$$\sum_{n=1}^{N}\Phi_{mn}^{U}(t)+\sum_{n=1}^{N}\Phi_{mn}^{D}(t)\le\Phi_{m}^{back}(t), m\in\mathcal{M}\setminus\{0\}, \tag{15c}$$

$$\|l_m(t)-l_{m'}(t)\|_2\ge d_{min}, m\ne m'\in\mathcal{M}\setminus\{0\}, \tag{15d}$$

$$\|l_m(t+1)-l_m(t)\|_2\le v_{max}\triangle t, \forall t\in\mathcal{T}, \tag{15e}$$

$$l_m(1)=l_m(T), \tag{15f}$$

$$\frac{1}{T}\sum_{t=1}^{T}p_m(t)\le\bar{p}, \tag{15g}$$

$$p_m(t)\le p_m^{max}, p_n(t)\le p_n^{max}, \tag{15h}$$

where $\boldsymbol{b}=\{(b_{mn}^{U}(t),b_{mn}^{D}(t))|m\in\mathcal{M},n\in\mathcal{N}\}$, $\boldsymbol{L}=\{l_m(t)|m\in\mathcal{M}\}$, and $\boldsymbol{p}=\{p_m(t)|m\in\mathcal{M}\}$. In this optimization problem, constraint (15b) ensures that each UE can associate with only one BS in either UL or DL. (15c) limits the maximum number of UEs that can associate with each UAV (due to the limited backhaul rate). (15d) ensures the minimum distance between any two UAVs for flying safely. (15e) limits the maximum fly velocity of each UAV in single time interval. (15f) ensures that the UAV can fly back to its original position along the trajectory after a period T. (15g) ensures that the power of each UAV is below the average power $\bar{p}$, and (15h) guarantees that the transmit power of each UE will not exceed the power rating p^{max} in a period T.

III. Multi-Agent DRL Based Decoupled Association and Trajectory Design

The solution to the above joint problem will give the policy of decoupled UE-BS association and UAV trajectory in a period. Due to the non-convexity of the objective function and the coupling between the user association and the UAVs' locations in the expression of $\Phi_{mn}^{(\cdot)}(t)$, (15) is a non-convex optimization problem. Deep reinforcement learning, which enables the agents to obtain the optimal policy by constantly interacting with environment, is considered as a promising technique to solve such problems in a sub-optimal manner. Nonetheless, most of the DRL-based methods for solving such problems only consider single agent systems. This may not be efficient when the number of network nodes increases. Also, leveraging the experiences of the nodes at the DRL agent will require exchanging of large amount of network information [10]. This motivates us to design a distributed multi-agent DRL-based approach to obtain an efficient solution.

A. Basic Model

We specify each UAV as an agent to learn and update experiences from the environment. Since the location of each UAV may be influenced by the current state that is observed locally, we first convert the joint optimization problem as a Partially Observable Markov Decision Process (POMDP) problem, where each agent independently chooses its action without having complete environment information.

In the following, we define the POMDP as a tuple $(\mathcal{S},\mathcal{A},R,P_0,P,\mathcal{O},Pr,\gamma)$, where $\mathcal{S}$, $\mathcal{A}$, and R are environment state set, action set and reward function, respectively. P is the transition function mapping the relationship from state s_t to s_{t+1}, and P_0 is the initial environment state distribution function. Since the environment is partially observable, the agent obtains an observation o_t from $\mathcal{O}$ instead of directly getting s_t from $\mathcal{S}$. Pr is the conditional probability of the observation. Finally, let $\gamma\in[0,1]$ be the discount factor. The detailed definitions for the POMDP are given as follows.

State and Observation Space: The state s_t consists of four elements: the attainable data transmission rate $\Phi_{mn}^{(\cdot)}(t)$, UAV location $l_m(t)$, fly velocity $v_m(t)$ and backhaul rate $\Phi_m^{back}(t)$ of all UAVs, which can be defined as

$$\boldsymbol{s}_t=\{\Phi_{mn}^{(\cdot)}(t),l_m(t),v_m(t),\Phi_m^{back}(t)\}, m\in\mathcal{M}\setminus\{0\}. \tag{16}$$

Thus, the state space can be expressed as $\mathcal{S}=\{\boldsymbol{s}_t\mid t=1,2,...,T\}$.

For the state space, calculating the transmission rate for each UAV-UE association pair involves link information (i.e., channel gain and intercell interference), which can only be observed locally and is not known by other association pairs. That is, for each UAV agent, only the link information of the UEs connected to itself can be observed. Therefore, combining (13) and the calculation of SINR, the observation space of agent m can be summarized as

$$o_t^m=\{\mathrm{SINR}_{mn}^{(\cdot)}(t),l_m(t),v_m(t),\Phi_m^{back}(t)\}, m\in\mathcal{M}\setminus\{0\}. \tag{17}$$

Accordingly, the observation space set is $\mathcal{O}=\{o_t^m\mid t=1,2,...,T,m\in\mathcal{M}\setminus\{0\}\}$.

Action Space: As described in (15), the UAV needs to associate with UEs and it determines the next position (i.e., flight direction and velocity) and power allocation. Therefore, for each UAV m at time slot t, its action can be defined as

$$\boldsymbol{a}_t=\{\boldsymbol{b}_m(t),\varsigma_m(t),v_m(t),\boldsymbol{p}_m(t)\}, \tag{18}$$

where $\boldsymbol{b}_m(t)=\{b_{mn}^{(\cdot)}\mid n\in\mathcal{N}\}$ denote the binary association variables, flight direction $\varsigma_m(t)\in[0,2\pi]$, flight velocity $v_m(t)\in[0,v_{max}]$, and $\boldsymbol{p}_m(t)=\{p_{mn}\mid n\in\mathcal{N}\}$ is the transmit power to each UE. Accordingly, the action space $\mathcal{A}=\{\boldsymbol{a}_t\mid t=1,2,...,T\}$.

Transition Probability: After executing actions, current state $\boldsymbol{s}_t$ will transfer to next one $\boldsymbol{s}_{t+1}$ with the probability $P(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)$.

Reward Design: In addition to state transition, the environment can feed back an immediate reward to the agent. In (15a), our objective is to maximize the attainable data rate in a long run. Naturally, the reward should be set as the data rate provided by a UAV in both UL and DL a time slot, i.e.,

$$R_t^m = \sum_{n=1}^{N} \Phi_{mn}^U(t) + \sum_{n=1}^{N} \Phi_{mn}^D(t), m \in \mathcal{M} \setminus \{0\}. \quad (19)$$

Note that each learning agent is capable of evaluating and improving the policy starting from state $\boldsymbol{s}$ by maximizing the state value function $V^\pi(\boldsymbol{s})$ and state-action value function $Q^\pi(\boldsymbol{s}, \boldsymbol{a})$, which are the expected reward of policy π, i.e.,

$$V^\pi(\boldsymbol{s}) = \mathbb{E}_{\zeta \sim P(\boldsymbol{s}_1)}(\sum_{t=1}^{T} \gamma^{t-1} R_t | \zeta_{\boldsymbol{s}_1} = \boldsymbol{s}), \quad (20)$$

$$Q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s'} \sim P(\boldsymbol{s'}|\boldsymbol{s},\boldsymbol{a})}(R(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'}) + \gamma^{V^\pi(\boldsymbol{s'})}. \quad (21)$$

Considering that the goal of an agent is to maximize the value of the expected cumulative discounted reward for a given policy π, the objective function of the POMDP can be expressed as

$$J(\pi) = \mathbb{E}_{\boldsymbol{s}} Pr^\pi(\boldsymbol{s}) \sum \pi(\boldsymbol{s}, \boldsymbol{a}) A^\pi(\boldsymbol{s}, \boldsymbol{a}), \quad (22)$$

where $A^\pi(\boldsymbol{s}, \boldsymbol{a}) = Q^\pi(\boldsymbol{s}, \boldsymbol{a}) - V^\pi(\boldsymbol{s})$ represents the advantage function, which evaluates the advantage of the value generated by a specific action compared to that by a general action.

B. Distributed Multi-Agent DRL Approach

In this section, a distributed multi-agent DRL with local states is investigated to solve the joint optimization problem, and deep neural network (DNN), as a function approximator, is used to cope with the challenges on large state and action spaces. In the initialization phase, agent m is employed with an Actor-Critic (AC) DNN [11] that takes the observation $\mathcal{O}_t^m$ as input and outputs the probability distribution of policy. Specifically, the Actor DNN is based on a policy gradient (PG) method to generate action sequences from a Gaussian distribution from given states, i.e.,

$$\pi_{\boldsymbol{\theta}^A}(\boldsymbol{s}, \boldsymbol{a}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}(\boldsymbol{s})} \exp(-\frac{\boldsymbol{a} - \hat{\mu}(\boldsymbol{s})}{2\hat{\sigma}(\boldsymbol{s})^2}), \quad (23)$$

where $\boldsymbol{\theta}^A$ denotes the parameters of Actor DNN, and $\hat{\sigma}(\boldsymbol{s})$ and $\hat{\mu}(\boldsymbol{s})$ are the standard deviation and the mean of the generated actions, which can be respectively expressed as $\hat{\sigma}(\boldsymbol{s}) = f_{\hat{\sigma}}(\boldsymbol{\theta}^A \boldsymbol{s}^\top + \boldsymbol{\kappa})$ and $\hat{\mu}(\boldsymbol{s}) = f_{\hat{\mu}}(\boldsymbol{\theta}^A \boldsymbol{s}^\top + \boldsymbol{\kappa})$, where $\boldsymbol{\kappa}$ is the bias vector. As an auxiliary module, the critic DNN uses a value-based method to generate temporal difference (TD) errors with discount γ, which can be used to guide the gradient of the actor to move toward the direction with a low cost.

In the distributed implementation phase, agent m first selects an action a_t^m according to the probability density function and then forms a joint action set $\dashv_t$ to interact with environment. As summarized in **Algorithm 1**, by importance sampling, the agent observes the state to obtain transmission rate, UAVs' locations and fly velocities, and backhaul rate. Then the state sequences are fed into Actor DNN to calculate actions for receiving the reward R_{t+1}^m. Finally, each agent stores the transition tuples $(o_t^m, a_t^m, R_{t+1}^m, done)$ in the replay memory dequeue and then adopts the proposed **Algorithm 2** to train the AC network.

Algorithm 1 MADRL based DUDe and Trajectory Design in Full-duplex Multi-UAV Networks

```
Initialize Actor-Critic networks for each agent
for each iteration do
    Run environment simulator.
    while UAVs are within the range of MBS do
        Initialize environment and receive an initial state s_1.
        for t = 1, ... , T do
            for each UAV agent m do
                Observe o_t^m and choose action a_t^m through importance sampling the density function.
            end for
            Get reward R_t^m and new environment state s_{t+1}.
            All agents perform an action a_t based on R_t^m and interact with environment for receiving reward R_{t+1}.
            for each UAV agent m do
                Calculate R_{t+1}^m.
                Store (o_t^m, a_t^m, R_{t+1}^m, done) into the replay memory.
            end for
        end for
    end while
    for each UAV agent m do
        Using Algorithm 2 to train AC DNN for each agent.
    end for
end for
```

C. Training With a Clip-and-Count-Based PPO

Note that in action (18), variable $\boldsymbol{b}_m(t)$ is discrete and the others are continuous. Although the policy gradient methods (e.g., deep deterministic policy gradient) can effectively obtain policy for the continuous space, it cannot cope with the discrete one. To this end, we develop a clip-and-count-based PPO algorithm to train the AC network (**Algorithm 2**), which uses a clipped surrogate ratio function to simplify the formulation [12]. For this, we first denote the probability ratio between the new and old policies by $\Upsilon(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\boldsymbol{\theta}_{old}}(\boldsymbol{a}|\boldsymbol{s})}$, where $\boldsymbol{\theta}$ is policy parameter. Accordingly, the loss function of the actor network can be expressed as

$$L(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\theta}_{old}, \boldsymbol{\theta}) = \min[\Upsilon(\boldsymbol{\theta}) \hat{A}_{\pi_{\boldsymbol{\theta}_{old}}}(\boldsymbol{s}, \boldsymbol{a}), clip(\Upsilon(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon) \hat{A}_{\pi_{\boldsymbol{\theta}_{old}}}], \quad (24)$$

where ϵ is a parameter for limiting the range of $\Upsilon(\boldsymbol{\theta})$.

Nonetheless, the original PPO may offer less flexibility and suffer from the risk of performance instability due to the use

Algorithm 2 Clip-and-Count-Based PPO

1: Input: initialize policy parameters $\boldsymbol{\theta}_0$ and value function parameters $\boldsymbol{\phi}_0$.
2: **for** $k = 0, 1, 2, ...$ **do**
3: Collect set of trajectories $\{\boldsymbol{s}_1, \boldsymbol{a}_1, \boldsymbol{s}_2, \boldsymbol{a}_2, ..., \boldsymbol{s}_k, \boldsymbol{a}_k\}$ by running policy $\pi_k = \pi(\boldsymbol{\theta}_k)$ in the environment.
4: Compute extrinsic reward $\hat{R}_t$.
5: Compute estimated advantage function $\hat{A}_t$ based on the current value function $V_{\boldsymbol{\phi}_k}$.
6: Set $\epsilon \sim N(\tilde{\mu}, \tilde{\sigma}^2)$.
7: Update the policy parameter using ϵ and (24):

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min(\Upsilon(\boldsymbol{\theta}) \cdot \hat{A}_{\pi_{\boldsymbol{\theta}_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t), g(\epsilon, \hat{A}_{\pi_{\boldsymbol{\theta}_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t))),$$

where $g()$ means stochastic gradient strategy.
8: Count C_t and compute intrinsic reward $\tilde{R}_t$.
9: Update the value function parameters using C_t and (26):

$$\boldsymbol{\phi}_{k+1} = \arg\min_{\boldsymbol{\phi}} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} (V_{\phi}(s_t) - (\hat{R}_t + \tilde{R}_t))^2.$$

10: **end for**

of the fix clip parameter ϵ. To deal with this issue, we set its value to a Normal distribution $\epsilon \sim N(\tilde{\mu}, \tilde{\sigma}^2)$, where $\tilde{\mu}$ is its expected value and $\tilde{\sigma}$ is its standard deviation. The meaning behind the modification is that, at the beginning of training, the surrogate is limited in a larger range since there exists a difference between old and current policies. As the training proceeds, the difference gradually decreases so that the limit in a smaller range should be added to the surrogate for accelerating convergence.

In addition to that, two reward functions are defined to further improve the training performance of Algorithm 2, i.e., extrinsic and intrinsic rewards. The former is the discounted reward $\hat{R}_t = \gamma^{t-1} r_t$, which has been shown in (20). The later refers to a bonus for serving marginal UEs by motivating UAVs to explore in a wide area, i.e.,

$$\tilde{R}_t = \begin{cases} \tilde{\lambda} \frac{1}{C_t}, & \text{if the count } C_t > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

where $0 \leq \tilde{\lambda} \leq 1$ is a coefficient, and C_t counts the number of times that the position l_t specified by the action $\boldsymbol{a}_t$ has been hovered before time slot t. It can be concluded that the more times the UAV hovers, the smaller the bonus. Therefore, each UAV tends to explore less hovering positions for achieving greater overall reward, and its exploration capability is then enhanced. Finally, we apply the PPO on the network architecture with shared parameters for both policy (actor) and value (critic) functions, the objective function is hence augmented with a mean-squared error term on the value estimation to encourage sufficient exploration. By the counting method, the loss function of the critic network can be written as

$$L(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{\phi}_{old}, \boldsymbol{\phi}) = (V_{\boldsymbol{\phi}_{old}}(\boldsymbol{s}, \boldsymbol{a}) - (\hat{R}_t + \tilde{R}_t))^2, \tag{26}$$

where $\boldsymbol{\phi}$ is the value function parameter.

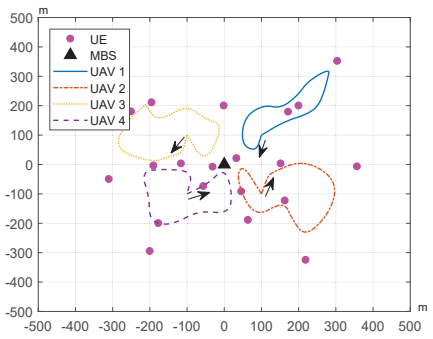

Fig. 2. Network layout and flight trajectories of four UAV.

IV. Simulation Results

For simulations, we consider a full-duplex multi-UAV network with one MBS, 4 UAVs and 20 UEs, which are randomly distributed within the coverage area of the MBS with radius of 500 m. All UAVs fly at an altitude of 200 m, and the hover plane is meshed in units of 10 m^2. The maximum transmit powers of MBS, UAVs, and UEs are set to be 43 dBm, 33 dBm, and 30 dBm, respectively. Besides, all BSs operate on 20 shared subchannels, and the carrier frequency $f_c = 2$ GHz. We assume $\delta = 60$ dB for SIC, and noise level $\sigma^2 = -60$ dBm. For the LoS, the factors $c_1 = 12.081$, $c_2 = 0.11395$, $\beta^{LoS} = 1.44544$ and $\beta^{NLoS} = 199.526$ [8]. The coefficient $\tilde{\lambda}$ in (25) is set to 1.

The actor and critic network for each agent both have 2 fully connected hidden layers, all containing 50 neurons. The learning rate of actor and critic network are both 0.0001. Besides, the discount factor is 0.999 and clip parameter distribution is set as $\epsilon \sim N(1, 0.3)$.

We first validate the effectiveness of the MADRL based approach for the considered network, then the performance of DFA mode is evaluated using the proposed approach. Finally, the superiorities of the clip-and-count-based PPO algorithm is demonstrated by comparing with several existing methods.

As shown in Fig. 2, the initial positions of four UAVs are set to $(100, 100)$, $(100, -100)$, $(-100, 100)$ and$(-100, -100)$, all at an altitude of 200 m. For clarity, four arrows are used to indicate the initial flight direction. As can be seen, in such settings, our approach ensures that the UAVs do not collide and can fly back to their initial positions after performing a series of position selections. This validates the effectiveness of the proposed MADRL approach.

For performance evaluation, the proposed DUDe in FD mode (DFA mode) is compared with three benchmarks using the proposed MADRL approach. All four association modes are described as follows:

- CHA: UE associates to the same UAV in UL and DL, which are separated by time division duplexing.
- CFA: UE associates to the same UAV in UL and DL, which can be used simultaneously.
- DHA: UE associates to different UAVs in UL and DL, which are separated by time division duplexing.

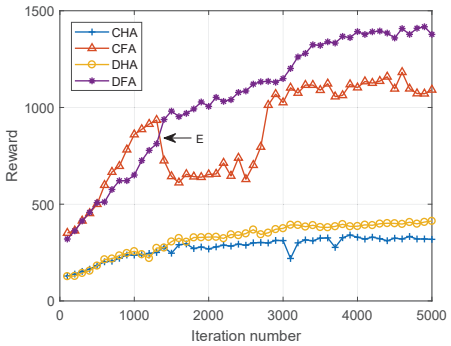

Fig. 3. System performance under different user association modes.

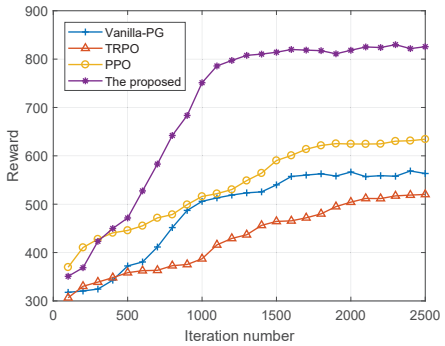

Fig. 4. Performance of the proposed MADRL training with different PG-based algorithms.

- DFA [Proposed]: UE associates to different UAVs in UL and DL, which can be used simultaneously.

As can be seen from Fig. 3, DFA outperforms the other three modes from point E onwards. In addition, in order to further evaluate the advantage of employing the decoupled UL-DL over the coupled one, we need to compare DFA with CFA or DHA with CHA. As shown in Fig. 3, the rewards in DHA are always higher than those in CHA. Note that DFA is not superior to CFA until point E. This is because, the extra interference caused by the decoupled association will bring down the transmission rate. As the iteration proceeds, with decoupled association, the rate gain becomes sufficient to compensate for the loss due to the extra interference.

Moreover, in order to verify the performance of the approach training with the proposed clip-and-count-based PPO algorithm, we compare the proposed algorithm with three other policy gradient-based methods. Fig. 4 unfolds a clear comparison among Vanilla-PG [13], trust region policy optimization (TRPO) [14], the original PPO [12], and the proposed clip-and-count based PPO in terms of the obtained reward. We can clearly see that among these algorithms, the proposed algorithm converges comparatively faster after 1500 iterations.

V. Conclusion

We have considered decoupled UL-DL association for a full-duplex multi-UAV network, and formulated an optimization problem of joint DUDe and trajectory design in order to maximize the sum-rate in the network. Considering that the formulated problem is non-convex, we have developed a distributed multi-agent deep reinforcement learning-based decoupled user association and trajectory design approach. To obtain a nearly optimal strategy, we have designed a Clip-and-Count based PPO algorithm to train the actor-critic neural network. For performance evaluation, we have compared the proposed association mode with other benchmarks. The results have shown the superiority of our proposed approach in terms of the attainable reward.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (Grant No. 62071230), Natural Science Foundation of Jiangsu Province (Grant No. BK20211567), and China Scholarship Council (Grant No. 202006830129). Kun Zhu is the corresponding author.

References

[1] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surv. Tuts.*, vol. 21, no. 3, pp. 2334–2360, 2019.

[2] B. Fan, L. Jiang, Y. Chen, Y. Zhang, and Y. Wu, "UAV assisted traffic offloading in air ground integrated networks with mixed user traffic," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–11, 2021. Early Access.

[3] M. Youssef, J. Farah, C. A. Nour, and C. Douillard, "Full-duplex and backhaul-constrained UAV-enabled networks using NOMA," *IEEE Trans. Veh. Technol.*, pp. 1–14, 2020. Early Access.

[4] H. E. Hammouti, M. Benjillali, B. Shihada, and M. Alouini, "Learn-as-you-fly: A distributed algorithm for joint 3D placement and user association in multi-UAVs networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5831–5844, Dec. 2019.

[5] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive approach design for 5G networks," in *Proc. IEEE GLOBECOM*, 2014, pp. 1798–1803.

[6] C. Liu, K. Ho, and J. Wu, "MmWave UAV networks with multi-cell association: Performance limit and optimization," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2814–2831, Dec. 2019.

[7] 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.

[8] C. Qiu, Z. Wei, X. Yuan, Z. Feng, and P. Zhang, "Multiple UAV-mounted base station placement and user association with joint fronthaul and backhaul optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5864–5877, Sept. 2020.

[9] Y. Shen, Z. Pan, N. Liu, X. You, and F. Zhu, "Performance analysis for full-duplex UAV legitimate surveillance system," in *Proc. IEEE ICC Workshops*, 2020, pp. 1–6.

[10] V. Saxena, J. Jaldén, and H. Klessig, "Optimal UAV base station trajectories using flow-level models for reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1101–1112, 2019.

[11] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern*, vol. 42, no. 3, pp. 1291–1307, Nov. 2012.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv: 1707.06347*, 2017.

[13] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. NIPS*, 2000, pp. 1057–1063.

[14] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. ICML*, 2015, pp. 1889–1897.