Decoupled Association With Rate Splitting Multiple Access in UAV-Assisted Cellular Networks Using Multi-Agent Deep Reinforcement Learning

Jiequ Ji, Lin Cai, *Fellow, IEEE*, Kun Zhu, *Member, IEEE*, and Dusit Niyato, *Fellow, IEEE*

***Abstract*—In unmanned aerial vehicles (UAVs) assisted cellular networks, user association plays an important role in interference control and spectrum efficiency. In this paper, we study the performance of uplink-downlink decoupled (UDDe) user association in a multi-UAV assisted network in which each user can associate with different UAVs or the macro base station (MBS) for uplink (UL) and downlink (DL) transmissions. Since some popular data may be requested by multiple users, grouping these users and applying multicasting can significantly improve spectral efficiency. Unlike traditional linear precoding that treats interference entirely as noise, we propose a rate-splitting multiple access (RSMA) policy that employs rate splitting at the transmitter and successive interference cancellation (SIC) at the receiver. To be specific, the transmitted signal is split into a common part and a private part, and the interference is partially decoded and partially treated as noise. In this context, we formulate a joint optimization problem of UL-DL association and beamforming for maximizing the sum-rate of users in UL and that of multicast groups in DL under the constraints of UAV backhaul capacity and power budget. Since the formulated problem is non-convex with intricate states and an individual UAV may not know the rewards of other UAVs, we convert it into a robust partially observable Markov decision process (POMDP). Then we resort to multi-agent deep reinforcement learning (MADRL) that enables each UAV to learn and optimize its policy in a distributed manner. To achieve an optimal policy, we further propose an improved clip and count-based proximal policy optimization (PPO) algorithm to train actor and critic networks. Simulation results demonstrate the superiority of the proposed decoupled association strategy with RSMA and the MADRL learning algorithm.**

Manuscript received 14 December 2022; revised 6 February 2023; accepted 1 March 2023. Date of publication 15 March 2023; date of current version 5 February 2024. This work were supported in part by the National Natural Science Foundation of China under Grant 62071230, in part by the National Research Foundation (NRF), Singapore and Infocomm Media Development Authority under the Future Communications Research Development Programme (FCP) and DSO National Laboratories under the AI Singapore Programme under Grant AISG2-RP-2020-019, under Energy Research Test-Bed and Industry Partnership Funding Initiative, and part by the Energy Grid 2.0 programme under DesCartes and the Campus for Research Excellence and Technological Enterprise (CREATE) programme. Recommended for acceptance by C.S. Xin. *(Corresponding author: Jiequ Ji.)*

Jiequ Ji is with the College of Future Science and Engineering, Soochow University, Suzhou 215222, China (e-mail: jiequ@nuaa.edu.cn).

Lin Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6, Canada (e-mail: cai@ece.uvic.ca).

Kun Zhu is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zhukun@nuaa.edu.cn).

Dusit Niyato is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: dniyato@ntu.edu.sg).

Digital Object Identifier 10.1109/TMC.2023.3256404

***Index Terms*—Decoupled multiple association, multi-agent deep reinforcement learning, rate splitting, UAV-assisted networks.**

I. Introduction

UNMANNED aerial vehicle (UAV)-assisted cellular systems have attracted increasing interest for 5 G and beyond networks [1]. To support high data rate and extended wireless coverage, UAVs can be deployed as aerial base stations (ABSs) to assist the macro base station (MBS) in service provisioning. In particular, the traffic load of the macro-cell can be offloaded to multiple UAV-cells to alleviate the burden on the MBS [2]. However, since UAVs do not have any wired connectivity, the backhaul from the MBS to UAVs may become a bottleneck. In addition, in services such as video conferencing, the traffic is bidirectional so it is critical to ensure high quality of services for both uplink (UL) and downlink (DL). With limited backhaul and spectrum resources, uplink-downlink user association should be carefully designed for the performance improvement of UAV-assisted cellular networks.

Most existing works on user-UAV association assume that a user is associated with the same UAV in both UL and DL transmissions [3], [4], [5]. Although such a coupled association is effective in single-tier networks, it may not guarantee optimal performance in multi-UAV cellular networks due to the non-uniform traffic loads and variable transmit powers of different UAVs for DL and UL transmissions. For example, a user may obtain a higher UL rate if it associates to a nearby UAV rather than a far-away MBS. This is because the UL rate of a user is affected by its own transmit power and its distance to the UAV. However, the DL rate from the MBS may be higher since its high transmit power, high backhaul capacity, etc. To this end, the concept of UL-DL decoupled (UDDe) association is introduced, which enables each user to associate with different UAVs or the MBS for DL and UL [6]. With UDDe association in UAV-assisted cellular networks, if using traditional time-division duplex to avoid UL/DL mutual interference, strict time synchronization and scheduling among all UAVs, MBS and users are needed, but difficult to achieve. It is desirable to have the flexibility of full-duplex (FD) transmission to relax the strict requirements. Using advanced self-interference (SI) cancellation techniques, FD becomes feasible [7]. Therefore, it is possible that each user can use the same frequency-band for UL and DL transmissions with the same or different BSs simultaneously. In the DL, since

many users may request the same information at the same time, multicasting the same message to a group of users can improve spectral efficiency. However, decoupled user association and the co-existing of unicast and multicast make interference management much more complicated and challenging.

To mitigate interference, many techniques have been introduced. For example, [8] proposed a bidirectional scheduling transmission scheme to alleviate interference. However, DL-to-UL interference is typically much stronger than UL-to-UL inter-cell interference due to the strong transmit power of BSs and line-of-sight (LoS) links among UAVs. With the use of rate splitting precoding at transmitters and successive interference cancellation (SIC) at receivers, rate splitting multiple access (RSMA) has been emerging as a prospective policy to mitigate interference and improve spectral efficiency [9], [10], [11]. The idea of this policy is to split each message into a common part to be decoded at all receivers and a private part to be decoded only at the intended receiver. For a receiver, SIC is used so that the common part can be first decoded and canceled, then the remaining private part from other users is treated as noise. Particularly, the split of common and private signals can be flexibly adjusted to partially treat the interference as noise.

Inspired by the advantage of RSMA as a flexible NOMA, we propose an RS-based UDDe transmission mode for multi-UAV cellular networks in FD communication. Specifically, two users with different channel gains are paired to perform RS in UL, where the user with stronger channel splits its signal into two parts and transmits a superimposed encoded stream, while the weak-user transmits a single stream. On the other hand, in the DL, since many users are interested in the same data, they can form a multicast group and be served by multiple UAVs. Given the existence of multiple multicast groups, the message of each group is split into a common part and a private part. All the common parts are packed together and encoded into a common stream shared by all groups, while the private parts are encoded into private streams for each group independently.

In this context, we formulate a joint optimization problem of UL-DL association and beamforming design for maximizing the sum-rate of users in UL and that of multicast groups in DL, considering limited power budget and backhaul capacity. However, the resultant problem is a non-convex programming problem with a highly non-linear objective function and non-convex constraints. In addition, since RS introduces a large number of precoded streams, increasing the complexity of the network environment, our formulated problem is often hard to solve and may not converge using traditional algorithms.

Recently, deep reinforcement learning (DRL) has emerged as a powerful approach for solving high-complexity and non-convex problems, which has been widely used in UAV-assisted networks, i.e., trajectory design and channel control [12], [13], [14], [15]. [15] proposed a DRL-based anti-jamming framework to learn jamming channel selection in UAV-aided cellular networks. The objective of DRL is to learn decisions iteratively though interaction with a dynamic environment so as to maximize the cumulative reward. However, most DRL approaches for solving non-convex problems consider only single-agent learning frameworks, which are not appropriate for our problem. The reason is that the large number of beamforming decisions leads to a highly-complicated training process. In addition, since an individual UAV-agent may not have global knowledge for the rewards from other agents (i.e., due to estimation uncertainty), single-agent learning will become non-stability. Therefore, we model our problem as a robust partially observable Markov decision process (POMDP) to deal with the environment uncertainty and resort to multi-agent deep reinforcement learning (MADRL) so that each UAV selects its policy in a distributed manner. To encourage lethargic agents to actively explore and address the problem of serious deviations between new and old policies due to actions with negative advantages, we propose a new clip-and-count based proximal policy optimization (PPO) algorithm to solve our robust POMDP.

The main contributions of this paper can be summarized as follows:

- This paper studies the precoder design problem of achieving the maximum sum rate in DL and UL for a multi-UAV cellular network with decoupled user association. For UL and DL, we propose a beamforming policy based on RSMA. In DL, users are grouped into multiple multi-cast groups given the sameness of requested data. This is the first work to combine RSMA with UDDe association for multigroup multicast.
- We formulate a joint UL-DL association and beamforming design for maximizing the sum-rate of users in UL and multicast groups in DL subject to per-UAV transmit power and backhaul capacity constraints. Due to its non-convexity and reward uncertainty, the formulated problem is modeled as a robust POMDP. Specifically, each UAV is treated as an agent that can adjust its associated user and beamforming matrix using its local observations from the time-varying network environment.
- A distributed MADRL-based framework is developed to solve our robust POMDP problem and an improved clip-and-count based PPO algorithm is proposed to achieve a near-optimal policy. To be specific, we design an intrinsic reward to motivate exploration and a new clip distribution to tackle the deviations between old and current policies.
- Simulation results show that our proposed algorithm converges to an optimal policy up to 29.4% faster than the standard PPO algorithm. In addition, our proposed RSMA transmission scheme outperforms state-of-the-art transmission schemes in term of sum-rate.

The rest of the paper is organized as follows. In Section II, we discuss the related work. Section III introduces the system model and formulates a joint association and beamforming problem. Section IV models our problem as a robust POMDP and proposes a MADRL-based algorithm to solve. Simulation results are presented in Section V to show the performance of the proposed RSMS scheme. Section VI concludes our work.

II. Related Work

There have been a few studies on user association in UAV-assisted communication networks [16], [17], [18], [19]. In [16], the association selection and UAV deployment were jointly

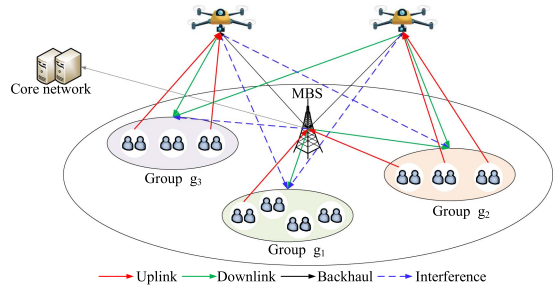

Fig. 1. An illustration for UL-DL decoupling association in a full-duplex wireless network consisting one MBS and multiple UAVs.

optimized to maximize the total rate of DL and UL. In [18], a learning-based UL-DL association scheme for rate fairness among users was proposed in dynamic multi-UAV systems. In addition, [19] numerically verified the feasibility of a decoupled association paradigm in UAV-assisted cellular networks. However, existing transmission techniques (e.g., non-orthogonal multiple access (NOMA) and space-division multiple access (SDMA)) used in these works are not efficient for scenarios with complex and diverse interference due to FD communications.

In contrast to NOMA that completely decodes interference and SDMA that treats interference as pure noise, RSMA can efficiently mitigate interference by making it partially decoded and partially treated as noise. There are a lot of studies on RSMA, which are categorized into downlink transmission [20], [21], [22], [23], [24] and uplink transmission [25], [26], [27]. A linear precoding rate splitting technique was considered in [20] for multi-user multi-antenna networks, and an optimal precoder with a guaranteed maximum weighted sum rate was derived. The minimum rate among users was maximized in [21] by jointly optimizing the message splitting, BS clustering and coordinate beamforming. [22] and [23] investigated the energy efficiency maximization problem for RSMA and NOMA schemes. In [24], the optimal rate allocation and power control were studied for maximizing the total rate of ground devices. Existing research efforts on RSMA mainly focus on the downlink rather than on the uplink. In [25], the outage performance was studied for uplink RSMA communications. In [26], a joint BS decoding order design and user power control algorithm was presented to maximize the total uplink rate. An efficient joint scheme of beamforming in the user side with rate splitting uplink NOMA was developed in [27] to improve the spectral efficiency. However, there is no work investigating the rate performance of combining RSMS transmission and UDDe association in full-duplex multi-UAV networks.

Notations: The following notations are used. $\mathcal{A}$ is a set, $\mathbf{A}$ is a matrix, a is a scalar, and $\mathbf{a}$ is a column vector. In addition, $\mathbb{C}^{M\times N}$ represents the complex space of dimension $M \times N$.

III. System Model

A. Network Model

As shown in Fig. 1, we consider the uplink and downlink of a cellular network comprised of multiple UAVs acting as aerial BSs and one MBS covering the entire target region. We use $\mathcal{M} = \{0, 1, \ldots, M\}$ and $\mathcal{U} = \{1, \ldots, U\}$ to denote the set of BSs (i.e., UAVs and MBS) and the set of users, respectively. All the UAVs are connected to the MBS by capacity-limited backhaul links that are orthogonal to each other and different from the radio links between BSs and users. The MBS has N_a transmit antennas and each UAV is equipped with N_b antennas. For UL, a user is associated with at most one BS. As such, we introduce binary variables $\{A^{\text{UL}}_{u,m}, m \in \mathcal{M}, u \in \mathcal{U}\}$ to indicate the user association states for UL, where $A^{\text{UL}}_{u,m} = 1$ when user u associates with BS m on UL and otherwise $A^{\text{UL}}_{u,m} = 0$. For DL, users interested in the same information are clustered into a multicast group and served cooperatively by multiple BSs. Note that a user requesting unicast services can be regarded as a multicast group with one user. We define the set of multicast groups as $\mathcal{N} = \{1, \ldots, N\}$. Then the set of users belonging to the n-th group is denoted by $\mathcal{G}_n$. Each user only belongs to at most one multicast group per transmission interval. Therefore, we have $\sum_{n\in\mathcal{N}} \mathcal{G}_n = \mathcal{U}$ and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ for $\forall i, j \in \mathcal{N}$ and $i \neq j$. We model the BS selection as $\{A^{\text{DL}}_{m,n}, m \in \mathcal{M}, n \in \mathcal{N}\}$, with $A^{\text{DL}}_{m,n} = 1$ means that the m-th BS is selected to serve the n-th multicast group and otherwise $A^{\text{DL}}_{m,n} = 0$.

We consider that both BSs and users perform in-band FD transmission to promote efficient spectrum reuse. However, FD transmission introduces self-interference (SI) between simultaneous UL and DL of each user or BS [28], which may lead to performance degradation. Thanks to recent breakthroughs in hardware design (e.g., digital baseband signal processing), SI can be reduced to near the noise level for low-power devices. Table I summarizes the main symbols used in this paper.

TABLE I
Notations and Definitions

Symbol	Definitions
$\mathcal{M}$, $\mathcal{U}$	Sets of BSs and users
N_a, N_b	Number of antennas of the MSB and each UAV
$A^{\text{UL}}_{u,m}$	User association variables for UL
$\mathcal{N}$, $\mathcal{F}$	Sets of multicast groups and user-pairs
$\mathcal{G}_n$	Set of users in the n-th group
$\mathbf{h}_{u,m}$, $\mathbf{h}_{m,n}$	Channel gain vectors
$P_{\max}$, $P^{\max}_m$	Peak power for each user and UAV m
$\Phi_{f,s}$, $\Phi_{f,w}$	Strong-user and weak-user in the f-th user-pair
q^{DL}_m, $A^{\text{DL}}_{m,n}$	BS selection for common and private streams in DL
$S_{f,s1}$, $S_{f,s2}$	Two strong signal streams
$S_{f,w}$	Weak signal stream
$W_{n,c}$, $W_{n,p}$	Common and private message streams
$P_{\Phi_{f,s}}$, P_m	Transmit power of strong-user $\Phi_{f,s}$ and BS m
σ^2	Additive white Gaussian noise for users
$\mathbf{w}_{m,c}$, $\mathbf{w}_{m,n}$	Beamforming for common and private streams
δ^2_m, $\delta^2_{z_n}$	SI cancellation capabilities of BS m and user z_n

B. Channel and Association Model

Denote the multiple-input multiple-output (MIMO) channel vectors from user u to BS m as $\mathbf{h}_{u,m} \in \mathbb{C}^{1\times N_k}$, from user u to user i as $\mathbf{h}_{u,i} \in \mathbb{C}^{1\times 1}$, from BS j to BS m as $\mathbf{h}_{j,m} \in \mathbb{C}^{N_k\times N_k}$, and from BS m to user z_n as $\mathbf{h}_{m,z_n} \in \mathbb{C}^{N_k\times 1}$ for any $j \in \mathcal{M}$

and $i \in \mathcal{U}$, where z_n is a user in the n-th multicast group and $k \in \{a, b\}$. These propagation channel vectors can be modeled by path-loss, shadowing and small-scale fading. On this basis, the received signal at BS m for UL is given by

$$y_m^{\mathrm{UL}} = \sum_{u\in\mathcal{U}} A_{u,m}^{\mathrm{UL}} \mathbf{h}_{u,m} \mathbf{x}_{u,m}^{\mathrm{UL}} + \underbrace{\sum_{j\in\mathcal{M}\setminus m} \sum_{n\in\mathcal{N}} A_{j,n}^{\mathrm{DL}} \mathbf{h}_{j,m} \mathbf{x}_{j,n}^{\mathrm{DL}}}_{\text{DL-to-UL interference}} + \underbrace{\sum_{j\in\mathcal{M}\setminus m} \sum_{u\in\mathcal{U}} A_{u,j}^{\mathrm{UL}} \mathbf{h}_{u,m} \mathbf{x}_{u,j}^{\mathrm{UL}}}_{\text{UL-to-UL interference}} + I_{\mathrm{DL}}^{\mathrm{self}} + n_m, \quad (1)$$

where $\mathbf{x}_{u,m}^{\mathrm{UL}} \in \mathbb{C}^1$ and $\mathbf{x}_{m,n}^{\mathrm{DL}} \in \mathbb{C}^{N_k}$ are the transmitted signals at user u towards BS m for UL and at BS m towards user u for DL, respectively; $I_{\mathrm{DL}}^{\mathrm{self}} = \mathbf{h}_m \mathbf{x}_{m,n}^{\mathrm{DL}}$ denotes the residual SI at BS m due to simultaneous UL and DL transmissions, where $\mathbf{h}_m$ is the SI channel that can be modeled as independent identically distributed Gaussian entries $\mathbf{h}_m \sim \mathcal{CN}(0, \delta_m^2)$ and $\frac{1}{\delta_m^2}$ is the SI cancellation capability for the BS[1]. $n_m \sim \mathcal{CN}(0, \sigma_m^2)$ is the additive white Gaussian noise (AWGN) at BS m. The transmit power for BS m and user u satisfies that $\sum_n^N \mathrm{tr}(P_{m,n}^{\mathrm{DL}}) \le P_m^{\max}$ and $\mathrm{tr}(P_{u,m}^{\mathrm{UL}}) \le P_{\max}$, where $P_{\max}$ and $P_m^{\max}$ are the peak power of each user and BS m, respectively. Similarly, the received signal of user z_n in DL is written as

$$y_{z_n}^{\mathrm{DL}} = \sum_{m\in\mathcal{M}} A_{m,n}^{\mathrm{DL}} \mathbf{h}_{m,z_n} \mathbf{x}_{m,z_n}^{\mathrm{DL}} + \underbrace{\sum_{m\in\mathcal{M}} \sum_{u\in\mathcal{U}\setminus z_n} A_{u,m}^{\mathrm{UL}} \mathbf{h}_{u,z_n} \mathbf{x}_{u,m}^{\mathrm{UL}}}_{\text{UL-to-DL interference}} + \underbrace{\sum_{m\in\mathcal{M}} \sum_{j\in\mathcal{N}\setminus n} A_{m,j}^{\mathrm{DL}} \mathbf{h}_{m,z_n} \mathbf{x}_{m,j}^{\mathrm{DL}}}_{\text{DL-to-DL interference}} + I_{\mathrm{UL}}^{\mathrm{self}} + n_{m,z_n}. \quad (2)$$

where $n_{m,z_n} \sim \mathcal{CN}(0, \sigma_{z_n}^2)$ denotes the AWGN at user z_n and $I_{\mathrm{UL}}^{\mathrm{self}} = \mathbf{h}_{z_n} \mathbf{x}_{z_n,m}^{\mathrm{UL}}$. Also, $\mathbf{h}_{z_n}$ follows $\mathcal{CN}(0, \delta_{z_n}^2)$.

C. Rate Splitting Transmission

Suppose that BSs can separate the signals of users through beamforming for UL and DL transmissions. In our model, we employ a linear precoding rate-splitting (RS) to mitigate inter-cell interference. The main idea of RS is to split the transmitted signal into common and private parts and enable the common signal to be decoded and removed from the original received signal by successive interference cancellation (SIC) to partially reduce interference. Note that the RS-enabled transmission is different in UL and DL, which is described in detail below.

1) RS-Uplink: To facilitate signal encoding and decoding operations, two users with different channel gains are paired to perform RS-enabled uplink NOMA [32], which are associated with the same BS. The channel gains for users are sorted in decreasing order, i.e., $\mathbf{h}_1 \ge \mathbf{h}_2 \ge, \ldots, \mathbf{h}_U$. In addition, the user pairing follows $(\mathbf{h}_1, \mathbf{h}_{\frac{U}{2}+1}), (\mathbf{h}_2, \mathbf{h}_{\frac{U}{2}+2}), \ldots, (\mathbf{h}_{\frac{U}{2}-1}, \mathbf{h}_U)$. We define $\mathcal{F} = \{1, \ldots, F\}$ as the set of user-pairs, which are pair-wise disjoint and $|\mathcal{F}| = \frac{|\mathcal{U}|}{2}$. The f-th user-pair is denoted by $\Phi_f = \{\Phi_{f,s}, \Phi_{f,w}\}$, where $\Phi_{f,s}$ and $\Phi_{f,w}$ are the strong-user with superior channel gain and the weak-user with inferior channel gain in the f-th user-pair, respectively. In each user-pair, only one user needs to split its message [33]. Specifically, the strong-user splits its message into two messages and sends $S_{f,s1}$ and $S_{f,s2}$ to its associated BS m with power $P_{f,s1}$ and P_{s2}, while the weak-user only sends a single stream $S_{f,w}$ to the same BS m with power $P_{f,w}$. As shown in Fig. 2, RS is performed via assigning two different powers to these two split parts. By using SIC, the m-th BS decodes the signals received from the f-th user-pair in the order of $S_{f,s1} \to S_{f,w} \to S_{f,s2}$.

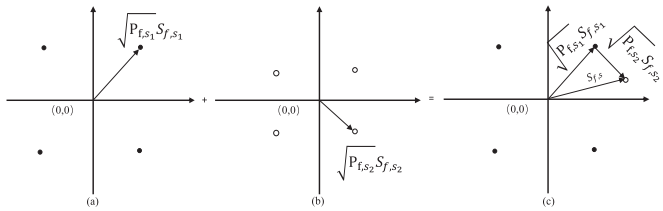

Fig. 2. An example of RS: (a) and (b) are the signal constellations of messages S_{f,s_1} and S_{f,s_2}; (c) is the signal constellation of transmitted signal $S_{f,s}$.

2) RS-Downlink: Since users interested in the same message are in the same multicast group for DL, the message of the n-th multicast group is split into a common message $W_{n,c}$ and a private message $W_{n,p}$ for $n \in \mathcal{N}$ based on the RS policy, i.e., $W_n \to \{W_{n,c}, W_{n,p}\}$. All the common messages from N multicast groups are then packed into a concatenated message $W_c \to \{W_{n,c}\}_{n\in\mathcal{N}}$ and encoded into a single common stream $\{W_{1,c}, W_{2,c}, \ldots, W_{N,c}\} \to S_c$. Meanwhile, private messages are individually encoded as separate private streams for each multicast group $W_p \to \{S_1, S_2, \ldots, S_N\}$. Therefore, the message stream vector $\mathbf{S} = [S_c, S_1, S_2, \ldots, S_N]^T \in \mathbb{C}^{(N+1)\times 1}$ is precoded by using the precoder matrix $\mathbf{P} = [\mathbf{p}_c, \mathbf{p}_1, \ldots, \mathbf{p}_N]$, where $\mathbf{p}_n \in \mathbb{C}^{N_k \times 1}$ and $\mathbf{p}_c \in \mathbb{C}^{N_k \times 1}$ are the precoder vectors for the m-th group's private stream and the common stream, respectively. At the beginning of decoding, the common stream S_c is decoded at each user by treating all the private streams as noise. After S_c has been decoded, each group of users decode their desired private stream by removing S_c from the received signal via SIC while treating the other private streams as noise.

D. Received Signal and Interference

Based on the above RS transmission model, we describe the received signal and interference for UL and DL.

1) Uplink: Let $\mathbf{h}_{f,m}^s \in \mathbb{C}^{1\times N_k}$ and $\mathbf{h}_{f,m}^w \in \mathbb{C}^{1\times N_k}$ denote the channel vectors from users $\Phi_{f,s}$ and $\Phi_{f,w}$ in the f-th user-pair to BS m, respectively. In addition, we define a binary variable $A_{f,m}^{\mathrm{UL}}$, which indicates that users $\Phi_{f,s}$ and $\Phi_{f,w}$ in the f-th user-pair are associated with BS m for UL if $A_{f,m}^{\mathrm{UL}} = 1$ and otherwise $A_{f,m}^{\mathrm{UL}} = 0$. For DL, we use $\mathbf{q} \in \{0, 1\}^{M\times U}$ to denote the BS selection vector, where $q_m^{\mathrm{DL}} = 1$ indicates that the m-th BS is selected to transmit the common stream and otherwise $q_m^{\mathrm{DL}} = 0$.

1.Different interference mitigation methods such as antenna cancellation and balun cancellation were proposed to alleviate the loss caused by SI [29], [30], [31].

As a results, the received signal at BS m is given by

$$y_m^{\text{UL}} = \sum_{f\in\mathcal{F}} A_{f,m}^{\text{UL}}[\mathbf{h}_{f,m}^s(\sqrt{P_{f,s1}}S_{f,s1}+\sqrt{P_{f,s2}}S_{f,s2}) + \mathbf{h}_{f,m}^w\sqrt{P_{f,w}}S_{f,w}] + I_{\text{UL-UL}} + I_{\text{DL-UL}} + I_{\text{self}}^{\text{DL}} + n_m, \tag{3}$$

where $S_{f,j}$ for $f\in\mathcal{F}$ and $j\in\{s,w\}$ satisfies $\mathbb{E}\{S_{f,j}S_{f,j}^H\}=\mathbf{I}$; $I_{\text{UL-UL}}$ and $I_{\text{DL-UL}}$ denote the interference from other users and BSs, respectively; and $I_{\text{self}}^{\text{DL}}$ is the residual SI at BS m. Let $\mathcal{J}$ denote the set of user-pairs except those associated with the m-th BS, where $|J| = F - \sum_{f\in\mathcal{F}} A_{f,m}^{\text{UL}}, \forall m\in\mathcal{M}$. Therefore, $I_{\text{UL-UL}}$ is written as

$$I_{\text{UL-UL}} = \sum_{m'\in\mathcal{M}\setminus m}\sum_{j\in\mathcal{J}} A_{j,m'}^{\text{UL}}\Big[\mathbf{h}_{j,m}^w\sqrt{P_{j,w}}S_{j,w} + \mathbf{h}_{j,m}^s\Big(\sqrt{P_{j,s1}}S_{j,s1}+\sqrt{P_{j,s2}}S_{j,s2}\Big)\Big]. \tag{4}$$

Moreover, $I_{\text{DL-UL}}$ is expressed as

$$I_{\text{DL-UL}} = \underbrace{\sum_{n\in\mathcal{N}}\sum_{m'\in\mathcal{M}\setminus m} A_{m',n}^{\text{DL}}\mathbf{h}_{m',m}\sqrt{P_{m',n}}S_n}_{\text{Private stream interference}} + \underbrace{\sum_{m'\in\mathcal{M}\setminus m} q_{m'}^{\text{DL}}\sqrt{P_{m',c}}S_c\mathbf{h}_{m',m}}_{\text{Common stream interference}}. \tag{5}$$

Finally, $I_{\text{self}}^{\text{DL}}$ is written as

$$I_{\text{self}}^{\text{DL}} = \sum_{n\in\mathcal{N}} A_{m,n}^{\text{DL}}\sqrt{P_{m,n}}S_n\mathbf{h}_m + q_m^{\text{DL}}\sqrt{P_{m,c}}S_c\mathbf{h}_m. \tag{6}$$

Upon receiving these encoded symbols, each BS and user-pair constructs their transmit signals by employing a superposition of linear precoded streams with beamforming weight matrices $\mathbf{W}_m = \{\mathbf{w}_{m,c}, \mathbf{w}_{m,1}, \mathbf{w}_{m,2}, \ldots, \mathbf{w}_{m,N}\} \in \mathbb{C}^{N_k\times 1}$ and $\mathbf{W}_f = \{\mathbf{w}_{f,s1}, \mathbf{w}_{f,s2}, \mathbf{w}_{f,w}\} \in \mathbb{C}^{1\times N_k}$, respectively. Since each transmitted symbol stream has unit energy, the transmit power of strong-user $\Phi_{f,s}$ and BS m for $\forall f\in\mathcal{F}$ and $\forall m\in\mathcal{M}$ is the energy cost of the beamforming. We thus have

$$P_{\Phi_{f,s}} = \|\mathbf{w}_{f,s_1}\|^2 + \|\mathbf{w}_{f,s_2}\|^2, \tag{7a}$$

$$P_m = \sum_{n\in\mathcal{N}} \|\mathbf{w}_{m,n}\|^2 + \|\mathbf{w}_{m,c}\|^2. \tag{7b}$$

Thus, $I_{\text{DL-UL}}$ and $I_{\text{self}}^{\text{DL}}$ are rewritten as follows:

$$I_{\text{DL-UL}} = \sum_{n\in\mathcal{N}}\sum_{m'\in\mathcal{M}\setminus m} A_{m',n}^{\text{DL}}\mathbf{h}_{m',m}\mathbf{w}_{m',n}S_n + \sum_{m'\in\mathcal{M}} q_{m'}^{\text{DL}}\mathbf{h}_{m',m}\mathbf{w}_{m',c}S_c, \tag{8}$$

$$I_{\text{self}}^{\text{DL}} = \sum_{n\in\mathcal{N}} A_{m,n}^{\text{DL}}\mathbf{h}_m\mathbf{w}_{m,n}S_n + q_m^{\text{DL}}\mathbf{w}_{m,c}\mathbf{h}_mS_c. \tag{9}$$

Based on the above interference analysis, the output signal of the f-th user-pair at BS m is expressed as

$$y_{f,m}^{\text{UL}} = \mathbf{h}_{f,m}^s(\mathbf{w}_{f,s_1}+\mathbf{w}_{f,s_2}) + \mathbf{h}_{f,m}^w\mathbf{w}_{f,w} + \sum_{i=1,i\neq f}^{F}(\mathbf{h}_{i,m}^s(\mathbf{w}_{i,s_1}+\mathbf{w}_{i,s_2}) + \mathbf{h}_{i,m}^w\mathbf{w}_{i,w}) + I_{\text{DL-UL}} + I_{\text{self}}^{\text{DL}} + n_m. \tag{10}$$

Since the decoding order of $S_{f,s1}\to S_{f,w}\to S_{f,s2}$ is used to decode the received signal, the output signal-to-interference-plus-noise ratio (SINR) for decoding $S_{f,s1}$ is written as

$$\gamma_{f,m}^{s1} = \frac{|\mathbf{h}_{f,m}^s\mathbf{w}_{f,s_1}|^2}{|\mathbf{h}_{f,m}^s\mathbf{w}_{f,s_2}|^2+|\mathbf{h}_{f,m}^w\mathbf{w}_{f,w}|^2+I_i+I_{\text{DL-UL}}+I_{\text{self}}^{\text{DL}}+\sigma_m^2} \tag{11}$$

where $I_i = \sum_{i=1,i\neq f}^{F}(|\mathbf{h}_{i,m}^s|^2(|\mathbf{w}_{i,s_1}|^2+|\mathbf{w}_{i,s_2}|^2)+|\mathbf{h}_{i,m}^w\mathbf{w}_{i,w}|^2)$ denotes the total interference from other user-pairs. Similarly, the SINRs of decoding $S_{f,w}$ and $S_{f,s2}$ are expressed as

$$\gamma_{f,m}^{w} = \frac{|\mathbf{h}_{f,m}^w\mathbf{w}_{f,w}|^2}{|\mathbf{h}_{f,m}^s\mathbf{w}_{f,s_2}|^2+I_i+I_{\text{DL-UL}}+I_{\text{self}}^{\text{DL}}+\sigma_m^2}, \tag{12}$$

$$\gamma_{f,m}^{s2} = \frac{|\mathbf{h}_{f,m}^s\mathbf{w}_{f,s_2}|^2}{I_i+I_{\text{DL-UL}}+I_{\text{self}}^{\text{DL}}+\sigma_m^2}. \tag{13}$$

Therefore, the received rates for streams $S_{f,s1}$, $S_{f,s2}$ and $S_{f,w}$ are expressed as

$$\begin{aligned} R_{f,m}^{s1} &= \log_2(1+\gamma_{f,m}^{s1}),\\ R_{f,m}^{s2} &= \log_2(1+\gamma_{f,m}^{s2}),\\ R_{f,m}^{w} &= \log_2(1+\gamma_{f,m}^{w}). \end{aligned} \tag{14}$$

Then the sum rate from both strong and weak users in the f- th user-pair is given by $R_{f,m} = R_{f,m}^{s1} + R_{f,m}^{s2} + R_{f,m}^{w}$.

2) Downlink: For DL, the transmitted signal of BS m is given by

$$\mathbf{s}_m = \mathbf{w}_{m,c}S_c + \sum_{n=1}^{N}\mathbf{w}_{m,n}S_n. \tag{15}$$

In addition, the z_n-th user in the n-th multicast group receives superimposed signals from multiple BSs, such as its intended common and private signals or interference signals from other multicast groups $n'\neq n$ or UL streams from other users. User z_n is assumed to be a strong user in the k-th user-pair for UL. The received signal of the z_n-th user is written as

$$y_{z_n} = \underbrace{\sum_{m\in\mathcal{M}} q_m^{\text{DL}}\mathbf{h}_{m,z_n}\mathbf{w}_{m,c}S_c}_{\text{Desired common signal}} + \underbrace{\sum_{m\in\mathcal{M}} A_{m,z_n}^{\text{DL}}\mathbf{h}_{m,z_n}\mathbf{w}_{m,z_n}S_n}_{\text{Desired private signal}} + \underbrace{\sum_{j\in\mathcal{N}\setminus n}\sum_{m\in\mathcal{M}} A_{m,j}^{\text{DL}}\mathbf{h}_{m,z_n}\mathbf{w}_{m,j}S_j}_{\text{Private interference}} + I_{\text{UL-DL}} + I_{\text{self}}^{\text{UL}} + \sigma_{z_n}^2 \tag{16}$$

for $\forall z_n \in \mathcal{G}_n$ and $\forall n \in \mathcal{N}$, where $I_{\text{self}}^{\text{UL}} = \mathbf{h}_{k,z_n}^{s}(\mathbf{w}_{z_n,s_1}S_{z_n,s_1} + \mathbf{w}_{z_n,s_2}S_{z_n,s_2}) + \mathbf{h}_{k,z_n}^{w}\mathbf{w}_{k,w}S_{k,w}$ is the residual SI at user z_n, $\mathbf{h}_{k,z_n}^{s}$ and $\mathbf{h}_{k,z_n}^{w}$ follow $\mathcal{CN}(0,\delta_{z_n}^2)$. Also $I_{\text{UL-DL}}$ denotes the interference from other user-pairs, which can be written as

$$I_{\text{UL-DL}} = \sum_{f\in\mathcal{F}\setminus k}[\mathbf{h}_{f,z_n}^{s}(\mathbf{w}_{f,s_1}S_{f,s_1} + \mathbf{w}_{f,s_2}S_{f,s_2}) + \mathbf{h}_{f,z_n}^{w}\mathbf{w}_{f,w}S_{f,w}]. \tag{17}$$

By using a linear precoded RS, each user recovers its desired message stream from the received signal based on a two-step decoding. In the first step, the common stream S_c is decoded at each user by treating all private streams as noise. The SINR for decoding common stream S_c at user z_n is given by

$$\gamma_{z_n,c} = \frac{\sum_{m\in\mathcal{M}} q_m^{\text{DL}}|\mathbf{h}_{m,z_n}\mathbf{w}_{m,c}|^2}{I_{z_n,c} + I_{\text{UL-DL}} + I_{\text{self}}^{\text{UL}} + \sigma_{z_n}^2}, \tag{18}$$

where $I_{z_n,c} = \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}} A_{m,n}^{\text{DL}}|\mathbf{h}_{m,z_n}\mathbf{w}_{m,n}|^2$ denotes the interference caused by all the private streams. The achievable rate of decoding S_c at user z_n is written as

$$R_{z_n,c} = \log_2(1+\gamma_{z_n,c}). \tag{19}$$

After S_c is successfully decoded, it will be cancelled from the original received signal y_{z_n} by means of SIC. Meanwhile, each group of users $\mathcal{G}_n$ can decode their intended private stream S_n by treating the irrelevant private streams as noise. The SINR for decoding private stream S_n at user z_n is given by

$$\gamma_{z_n,p} = \frac{\sum_{m\in\mathcal{M}} A_{m,n}^{\text{DL}}|\mathbf{h}_{m,z_n}\mathbf{w}_{m,z_n}|^2}{I_{z_n,p} + I_{\text{UL-DL}} + I_{\text{self}}^{\text{UL}} + \sigma_{z_n}^2}, \tag{20}$$

where $I_{z_n,p} = \sum_{j\in\mathcal{N}\setminus n}\sum_{m\in\mathcal{M}} A_{m,j}^{\text{DL}}|\mathbf{h}_{m,z_n}\mathbf{w}_{m,j}|^2$ denotes the interference caused by the private streams of other groups. The achievable rate of decoding S_n at user z_n is expressed as

$$R_{z_n,p} = \log_2(1+\gamma_{z_n,p}). \tag{21}$$

Due to the fact that each user should be able to successfully decode the common part first, the achievable transmit rate with the common stream shall not exceed R_c, which is written as

$$R_c = \min_{\forall z_n\in\mathcal{G}_n,\forall n\in\mathcal{N}} \log_2(1+\gamma_{z_n,c}). \tag{22}$$

Since R_c is shared among all multicast groups, we have

$$R_c = \sum\nolimits_{n=1}^{N} C_n, \tag{23}$$

where C_n is the common rate allocated to the n-th multicast group, which is defined as

$$C_n = \frac{\mathbb{L}(W_{n,c})}{\sum_{n=1}^{N}\mathbb{L}(W_{n,c})}R_{z_n,c}, \tag{24}$$

where $0 \le \frac{\mathbb{L}(W_{n,c})}{\sum_{n=1}^{N}\mathbb{L}(W_{n,c})} \le 1$ is the splitting ratio and $\mathbb{L}(\cdot)$ is the length of message. In the n-th multicast group, the private stream S_n shall be decoded by all users in $\mathcal{G}_n$. Thus, the private rate R_n of the n-th multicast group is determined by its worst user, which is given by

$$R_n = \min_{\forall z_n\in\mathcal{G}_n} \log(1+\gamma_{z_n,p}),\ \forall n\in\mathcal{N}. \tag{25}$$

To this end, the total rate of the n-th multicast group is written as $\widetilde{R}_n = C_n + R_n$.

Due to the wireless backhaul link is orthogonal to the radio access link, there is no interference among them. Accordingly, the achievable backhaul rate of UAV m is given by

$$R_m^{\text{back}} = \log_2\left(1 + \frac{|\mathbf{h}_{0,m}\mathbf{w}_{0,m}|^2}{\sigma_m^2}\right). \tag{26}$$

E. Problem Formulation

Motivated the aforementioned analysis, we formulate a joint optimization problem of common rate allocation, beamforming design, and decoupled association. The objective is to maximize the sum rate of user-pairs in UL and that of multicast groups in DL while ensuring user fairness within each group subject to the constraints of user-BS transmit power and UAV backhaul capacity. Mathematically, this problem is written as

$$\text{P0:}\ \max_{\mathbf{A},\mathbf{q},\mathbf{C},\mathbf{w}_m,\mathbf{w}_f} \sum_{f\in\mathcal{F}}\sum_{m\in\mathcal{M}} A_{f,m}^{\text{UL}}R_{f,m} + \sum_{n\in\mathcal{N}}\widetilde{R}_n \tag{27a}$$

$$\text{s. t.}\quad C_n \ge 0, \forall n\in\mathcal{N}, \tag{27b}$$

$$\sum_{n=1}^{N} C_n \le R_c, \tag{27c}$$

$$P_{f,s1} + P_{f,s2} \le P_{\max}, P_{f,w} \le P_{\max}, \forall f\in\mathcal{F}, \tag{27d}$$

$$P_{m,c} + \sum_{n=1}^{N} P_{m,n} \le P_m^{\max}, \forall m\in\mathcal{M}, \tag{27e}$$

$$q_m^{\text{DL}} \in \{0,1\}, \forall m\in\mathcal{M}, \tag{27f}$$

$$(1-q_m^{\text{DL}})\mathbf{w}_{m,c} = 0, \forall n\in\mathcal{N}, \tag{27g}$$

$$A_{m,n}^{\text{DL}} \in \{0,1\}, \forall n\in\mathcal{N}, \forall m\in\mathcal{M}, \tag{27h}$$

$$(1-A_{m,n}^{\text{DL}})\mathbf{w}_{m,n} = 0, \forall m\in\mathcal{M}, \forall n\in\mathcal{N}, \tag{27i}$$

$$A_{f,m}^{\text{UL}} \in \{0,1\}, \sum_{m=1}^{M} A_{f,m}^{\text{UL}} \le 1, \forall f\in\mathcal{F}, \tag{27j}$$

$$A_{f,m}^{\text{UL}}(1-\mathbf{w}_{f,j}) = 0, \forall f\in\mathcal{F}, j\in\{s_1,s_2,w\}, \tag{27k}$$

$$\sum_{f\in\mathcal{F}} A_{f,m}^{\text{UL}}R_{f,m} + \sum_{n\in\mathcal{N}} q_m^{\text{DL}}C_n + \sum_{n\in\mathcal{N}} A_{m,n}^{\text{DL}}R_n \le R_m^{\text{back}}, m\in\mathcal{M}\setminus 0, \tag{27l}$$

where $\mathbf{C} = \{C_1,\ldots,C_N\}$ is the common rate allocation vector. Constraint (27b) guarantees that the assigned common rate of each group is non-negative, while constraint (27c) ensures that the received common rate of all groups cannot exceed the achievable common rate of any group. Constraints (27d) and (27e) describe the transmit power limitations for users and BSs. Constraints (27f)–(27i) mean that the beamforming vector is zero if the corresponding BS is not selected. Constraint (27j) ensures that each user-pair only associates with one BS for UL. Constraint (27k) means that the beamforming vector is zero if the user-pair does not associate with the BS. Constraint (27l)

restricts the number of users associated to each UAV for UL and DL to avoid the backhaul overload.

IV. Learning-Based Association and Beamforming

A. Methodology

The formulated optimization problem P0 is non-convex and challenging to get a global optimal solution. Meanwhile, some binary decision variables make it more hard to solve. Existing studies simplify such a non-convex problem as several different subproblems and then alternately optimize the variables of each subproblem during each iteration until convergence. Such an approach makes this non-convex problem easy to solve at the cost of optimality. In addition, the computational complexity greatly increases with the introduction of beamforming design and decoupling association. Deep reinforcement learning (DRL) has received considerable attention due to its ability to transform intractable optimization problems into maximizing cumulative rewards through reward design. There are a few studies on using DRL-based centralized solutions to solve high complexity problems in multi-agent scenarios. However, such solutions may lead to poor fault tolerance and flexibility as the network scale becomes larger. Fortunately, multi-agent DRL (MADRL) is able to provide a distributed solution for a multi-agent problem, where each agent makes decisions based on its own local information, which thus can keep the state space and action space from increasing with the size of the network. However, an individual agent may not be able to get complete and accurate knowledge from the training model, which is also known as model uncertainty. Inspired by the aforementioned facts, we develop a robust MADRL-based framework to solve our problem in a distributed manner.

To be specific, we model the original problem P0 as a robust partially observable Markov Decision Process (POMDP). The network settings in this paper are treated as the environment and each UAV is treated as a controller (i.e., agent) that learns and updates its experience from the environment based on a distributed MADRL framework and until reaching an optimal policy. To overcome the instability in MADRL-based learning systems due to model uncertainty, we propose a new clip and count-based Proximal Policy Optimization (PPO) algorithm to facilitate agents to continuously train neutral networks.

B. Preliminaries of POMDP

Since the target of problem P0 is to maximize the sum rate of user-pairs on UL and that of multicast groups on DL in a time-varying full-duplex decoupled system, we formulate it as a POMDP, which is denoted by

$$\Omega = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P_0, \mathcal{R}, \gamma, \mathcal{P} \rangle, \tag{28}$$

where $\mathcal{S}$ denotes the set of states describing the environment; $\mathcal{O}$ and $\mathcal{A}$ are the observation and action spaces, respectively. $\mathcal{R}$ is the reward function that maps the network state and the joint actions of agents to rewards; P_0 denotes the initial environment state distribution function; $\gamma \in [0,1]$ denotes the discount factor related with future rewards; and $\mathcal{P}$ is a state transition function. In particular, $P_{s_t,s_{t+1}}(a_t)$ is the probability that state s_t enters a new state s_{t+1} after executing action a_t. At decision time slot t, each agent k gets a local observation from its state $s_k(t) \in \mathcal{S}$ and takes an action $a_k(t) \in \mathcal{A}$ with a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$. Then it obtains a reward and the environment moves to the next state $s_k(t+1)$ according to the probability $P(s_k(t+1)|s_k(t), a_k(t))$. For our problem, these elements are described in detail below.

1) State and Observation Space: We use s_t to denote the state at time slot t, which reveals the current conditions of each user and UAV and contains four parameters, namely, rate of decoding common stream R_c, private rate of the n-th multicast group R_n, backhaul rate of each UAV R_m^{back}, sum-rate of users in the f-th user-pair $R_{f,m}$, which can be defined as

$$s_t = \{R_m^{\text{back}}, R_{f,m}, R_n, R_c\}, \forall n \in \mathcal{N}, f \in \mathcal{F}, m \in \mathcal{M} \backslash 0. \tag{29}$$

The state space is then denoted as $\mathcal{S} = \{s_t | t = 1, \ldots, T\}$. At time slot t, each agent observes its own state information so as to make an efficient decision. However, the rate between each user and UAV is determined by link information such as inter-cell interference and channel gain, which can only be observed locally and not known to other user-UAV pairs. According to the rate expression, the observation of agent m is given by

$$\begin{aligned} o_t^m = \{&R_m^{\text{back}}, R_{f,m}, R_n, R_c, \gamma_{f,m}^{s1}, \gamma_{f,m}^{s2}, \gamma_{f,m}^{w}, \\ &\gamma_{z_n,c}, \gamma_{z_n,p}\}, \forall n \in \mathcal{N}, f \in \mathcal{F}, m \in \mathcal{M} \backslash 0. \end{aligned} \tag{30}$$

Thus the set of the observation space is given by $\mathcal{O} = \{o_t^m | t = 1, \ldots, T, m \in \mathcal{M} \backslash 0\}$.

2) Action Space: At time slot t, each UAV is responsible for associating with suitable users and determining the beamforming as well as the power and common rate allocation. To this end, the action of the m-th UAV is written as

$$\begin{aligned} a_t^m = \{&P_{m,n}, P_{f,m}, \mathbf{w}_{f,m}, \mathbf{w}_{m,n}, P_{f,m}, \mathbf{w}_{m,c}, C_n, q_m^{\text{DL}}, \\ &A_{f,m}^{\text{UL}}, A_{m,n}^{\text{DL}}\}, \forall n \in \mathcal{N}, f \in \mathcal{F}, m \in \mathcal{M} \backslash 0. \end{aligned} \tag{31}$$

We define $\mathcal{A} = \{a_t^m | t = 1, \ldots, T, m \in \mathcal{M} \backslash 0\}$ as the set of the action space.

3) Reward Design: In a DRL-based framework, each agent aims at exploring a policy that maximizes its expected reward from the environment every decision time slot. Therefore, our formulated difficult-to-optimize objective can be simplified as maximizing the expected cumulative reward through effective reward design.

The objective of problem P0 is to maximize the total rate of multicast-groups and user-pairs in both DL and UL. In general, the cumulative reward corresponds to the objective function. However, it should be considered that constraints in P0 are not satisfied during the training phase when designing the reward function. In order to avoid backhaul capacity overload, we can introduce a penalty term to the original objective function. In particular, a specific reward function is defined as follows:

$$r(t) = R(t) - \omega_1 \chi_{\text{back}}(t) - \omega_2 \chi_{\text{power}}(t), \tag{32}$$

where the first term $R(t) = \sum_{f \in \mathcal{F}} \sum_{m \in \mathcal{M}} A_{f,m}^{\text{UL}}(t) R_{f,m}(t) + \sum_{n \in \mathcal{N}} \widetilde{R}_n(t)$ denotes the immediate data rate and the latter two

terms are penalty functions on the overloaded backhaul capacity (27l) and the excess transmit power (27e). Moreover, the weights ω_1 and ω_2 are positive constants used to evaluate the importance of constraints. $\chi_{\text{back}}(t)$ and $\chi_{\text{power}}(t)$ are binary indicators, where $\chi_{\text{back}}(t) = 0$ means that the UAV backhaul capacity is satisfied at time slot t and $\chi_{\text{back}}(t) = 1$ otherwise. Similarly, $\chi_{\text{power}}(t) = 0$ if the transmit power is satisfied at time slot t and $\chi_{\text{power}}(t) = 0$ otherwise. Thus, we have $r(t) = \{r_t^m, m \in \mathcal{M} \setminus 0\}$.

Each agent continuously observes the environment and its interaction process can be represented via a Markov chain $\zeta = \boldsymbol{s}_1, \boldsymbol{a}_1, \boldsymbol{r}_1, \boldsymbol{s}_2, \boldsymbol{a}_2, \boldsymbol{r}_2, \ldots, \boldsymbol{s}_T, \boldsymbol{a}_T, \boldsymbol{r}_T$. Thus, the probability of each interaction process is given by

$$P(\zeta) = \boldsymbol{s}_1 \prod_{t=1}^{T} \pi(\boldsymbol{s}_1, \boldsymbol{a}_1) P(\boldsymbol{s}_{t+1} | \boldsymbol{a}_t, \boldsymbol{s}_t), \tag{33}$$

where $\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = P(\boldsymbol{a}_t | \boldsymbol{s}_t)$ is a stochastic policy. Then the state transition probability from $\boldsymbol{s}$ to $\boldsymbol{s}' \in \mathcal{S}' \subseteq \mathcal{S}$ after taking action $\boldsymbol{a}$ is expressed as

$$P(\boldsymbol{s}_{t+1} \in \mathcal{S}') = \int_{\mathcal{S}'} \Psi(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') d\boldsymbol{s}', \tag{34}$$

where $\Psi(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$ denotes the transition function [34]. Starting from state $\boldsymbol{s}$, each agent can evaluate and improve its policy by maximizing the state-value function $V^{\pi}(\boldsymbol{s})$ and action-value function $Q^{\pi}(\boldsymbol{s}, \boldsymbol{a})$, which can be written as

$$V^{\pi}(\boldsymbol{s}) = \mathbb{E}_{\zeta \sim P(\boldsymbol{s}_1)} \left(\sum_{t=1}^{T} \gamma^{t-1} r(t) | \zeta_{\boldsymbol{s}_1} = \boldsymbol{s} \right), \tag{35}$$

$$Q^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}' \sim P(\boldsymbol{s}' | \boldsymbol{s}, \boldsymbol{a})} (r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma V^{\pi}(\boldsymbol{s}')), \tag{36}$$

where $Q^{\pi}(\boldsymbol{s}, \boldsymbol{a})$ is also the expected cumulative reward and γ is the discount factor reflecting the weight of future rewards. Considering that each agent's objective is to search a policy π that takes an action $\boldsymbol{a}$ at state $\boldsymbol{s}$ so as to maximize the expected discounted reward, the objective function of POMDP is given by

$$J(\pi) = \mathbb{E}_{\boldsymbol{s}} Pr^{\pi}(\boldsymbol{s}) \sum \pi(\boldsymbol{s}, \boldsymbol{a}) A^{\pi}(\boldsymbol{s}, \boldsymbol{a}), \tag{37}$$

where $Pr^{\pi}(\boldsymbol{s})$ is the probability distribution of selecting policy π under state $\boldsymbol{s}$. $A^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = Q^{\pi}(\boldsymbol{s}, \boldsymbol{a}) - V^{\pi}(\boldsymbol{s})$ is an advantage function that evaluates how good a specific action is compared to other available actions.

C. Robust POMDP

It is obvious that the objective of POMDP is to maximize the expected cumulative reward, which depends on the behavior of all agents. In practice, an individual agent may not be able to get complete and accurate information from the environment, such as transition probability function (34) and reward function (32). Specifically, each UAV selects an individual action with-out fully understanding the rewards and joint transitions of other UAVs. In this case, poor system performance may be experienced in practice. To tackle this issue, the trained policy needs to be robust to possible uncertainties of POMDP [35]. In particular, we transform the original problem into a robust POMDP, which is described as follows:

$$\tilde{\Omega} = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, \tilde{P}_{\boldsymbol{s}}, P_0, \tilde{r}_{\boldsymbol{s}}, \gamma \rangle, \tag{38}$$

where $\tilde{P}_{\boldsymbol{s}}$ and $\tilde{r}_{\boldsymbol{s}}$ are the uncertainty sets of possible transition probability functions and expected rewards at state $\boldsymbol{s}$, respectively. The behavior of an individual agent (i.e., *natural* agent are indexed by 0) is used to characterize uncertainty, which is mutually resistant to the behavior of all other agents. Hence, the set of policies is given by

$$\boldsymbol{\pi}_{\theta^0} = \{\pi_{\theta^{0,m}} | m \in \mathcal{M} \setminus 0\}, \tag{39}$$

where $\theta^0 = (\theta^{0,1}, \theta^{0,2}, \ldots, \theta^{0,M})$, which indicates that different agents have varying uncertainty sets. Furthermore, the joint policies for all individual and natural agents are parameterized by $\theta = (\theta^0, \theta^1, \ldots, \theta^M)$. For our robust POMDP, the state-action and state-value functions are respectively expressed as

$$\tilde{Q}^{\pi}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}' \sim \tilde{P}(\boldsymbol{s}' | \boldsymbol{s}, \boldsymbol{a})} (\tilde{r}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma V^{\pi}(\boldsymbol{s}')), \tag{40}$$

$$\tilde{V}^{\pi}(\boldsymbol{s}) = \mathbb{E}_{\zeta \sim \tilde{P}(\boldsymbol{s}_1)} \left(\sum_{t=1}^{T} \gamma^{t-1} \tilde{r}(t) | \zeta_{\boldsymbol{s}_1} = \boldsymbol{s} \right). \tag{41}$$

We define the the natural-agent objective function $J(\pi_{\theta^{0,m}})$ and the individual-agent objective function $J(\pi_{\theta^m})$ as follows:

$$J(\pi_{\theta^{0,m}}) = \mathbb{E}_{\boldsymbol{s}} Pr^{\pi_{\theta^{0,m}}}(\boldsymbol{s}) \sum \pi(\theta^{0,m}), \tag{42}$$

$$J(\pi_{\theta^m}) = \mathbb{E}_{\boldsymbol{s}} Pr^{\pi_{\theta^m}}(\boldsymbol{s}) \sum \pi(\boldsymbol{s}, \boldsymbol{a}) \tilde{A}^{\pi_{\theta^m}}(\boldsymbol{s}, \boldsymbol{a}), \tag{43}$$

where $\tilde{A}^{\pi_{\theta^m}}(\boldsymbol{s}, \boldsymbol{a}) = \tilde{Q}^{\pi_{\theta^m}}(\boldsymbol{s}, \boldsymbol{a}) - \tilde{V}^{\pi_{\theta^m}}(\boldsymbol{s})$ is the advantage function. Next, we develop a distributed MADRL framework to learn the optimal policy for each agent.

D. Distributed Multi-Agent DRL

Note that finding the optimal policy for our robust POMDP using a simple RL-based method is challenging due to its large and complex state space. In order to overcome this challenge, we consider a MADRL framework with local states and define deep neural networks (DNNs) as function approximators. To be specific, each UAV acts as an agent that interacts with the network environment and learns its experience independently, which can be used to optimize the joint policy of beamforming allocation and decoupled association. This indicates that each agent may need to explore the optimal policy without complete information about all agents. As shown in Fig. 3, our proposed MADRL framework adopts centralized training and distributed execution to address model uncertainty due to individual agent training. Specifically, each agent has an actor-network and a critic-network, where the actor network makes decisions based on its local observations while the critic network evaluates the output of the actor-network.

1) Centralized Training: In this phase, each UAV-agent has to learn the association with users, control the transmit power and common stream rate, and determine the beamforming. In addition, experience replay techniques are used to increase the training stability for the optimal policy. The state transition

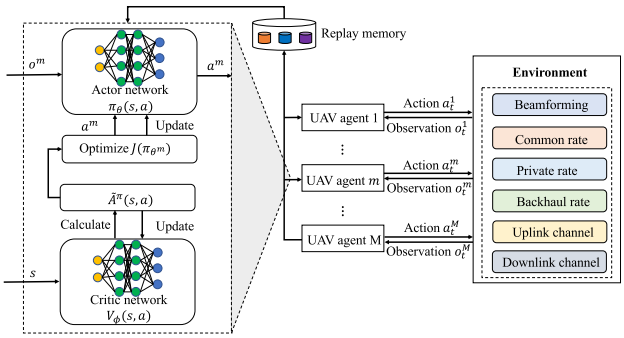

Fig. 3. Illustration of multi-agent DRL framework for full-duplex networks.

samples of each agent are stored into a replay buffer with size $\mathcal{B}$, which consists of the tuple $\{\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s}'\}$ and is implemented on the MBS. In the learning phase, the neural network is updated via randomly sampling mini-batch experiences from the replay buffer, which breaks the correlation between sequential samples and alleviates the learning oscillation.

At the beginning of each episode, each UAV-agent observes the state $s_t = \{R_m^{\text{back}}, R_{f,m}, R_n, R_c\}, \forall n \in \mathcal{N}, f \in \mathcal{F}, m \in \mathcal{M} \backslash 0$ and the received information is then stored in the replay buffer. Next, the actor network of agent m takes its local observations $\mathcal{O} = \{\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s}'\}$ from the replay buffer as the input and then outputs the policy probability distribution. In other words, the actor network is responsible for generating a sequence of actions through optimizing $J(\pi_{\theta^m})$, i.e., the objective function of the robust POMDP defined in (43). Based on this purpose, the actor network will generate the following policy

$$\pi_{\boldsymbol{\theta}^A}(\boldsymbol{s}, \boldsymbol{a}) = \frac{1}{\sqrt{2\pi}\hat{\varrho}(\boldsymbol{s})} \exp\left(-\frac{\boldsymbol{a} - \hat{\mu}(\boldsymbol{s})}{2\hat{\varrho}(\boldsymbol{s})^2}\right), \tag{44}$$

where $\boldsymbol{\theta}^A$ denotes the parameters of actor networks; $\hat{\mu}(\boldsymbol{s})$ and $\hat{\varrho}(\boldsymbol{s})$ denote the mean and standard deviation for the generated actions, which are expressed respectively as

$$\hat{\varrho}(\boldsymbol{s}) = f_{\hat{\varrho}}(\boldsymbol{\theta}^A \boldsymbol{s}^\top + \boldsymbol{\kappa}), \tag{45}$$

$$\hat{\mu}(\boldsymbol{s}) = f_{\hat{\mu}}(\boldsymbol{\theta}^A \boldsymbol{s}^\top + \boldsymbol{\kappa}), \tag{46}$$

where $\boldsymbol{\kappa}$ denotes the bias vector; $f_{\hat{\mu}}$ and $f_{\hat{\varrho}}$ are the activation functions of the output layer and the hidden layer of the actor network. The critic network is responsible for computing the centralized advantage function $\tilde{A}^{\pi_{\theta^m}}(\boldsymbol{s}, \boldsymbol{a})$, which can be used to guide the gradient of the actor to move toward the direction with low cost. Moreover, the advantage function is constantly updated as the training progresses.

Since the MBS has a significant computational advantage over UAVs in the network, the training of our MADRL frame-work can be conducted centrally on the MBS in an offline way. After sufficient training, the resulting training model is directly utilized in the distributed execution phase.

2) Distributed Execution: In this phase, each UAV employs a trained actor network to generate the corresponding action sequences with its own observations in each learning step. As a results, each UAV is able to adjust its common stream and transmit power allocation as well as beamforming to provide better services for associated users. Although the actions of all UAVs may be updated simultaneously, it is also possible for an individual UAV to have no knowledge of the actions taken by other UAVs.

Based on the above analysis, the MADRL approach for joint decoupled association and beamforming and common rate allocation is summarized as Algorithm 1. At the beginning, the actor-critic network, the parameter settings for our multi-UAV assisted cellular network and the replay memory are initialized. Each training episode is set to have T time slots. At time slot t, agent $m \in \mathcal{M}$ observes the state o_t^m to receive the common and private stream rates and the rate from the user-pair to the UAV through importance sampling. Note that only the actor network works in this step. Then the sequence of states is fed into the corresponding actor-network to calculate the actions that receive the reward $r_m(t+1)$. Finally, each agent stores the transition tuples $\{o_t^m, a_t^m, r_m(t+1), o_{t+1}^m\}$ into the replay memory and then exploits the proposed Algorithm 2 to train the actor-critic network.

E. Training With Clip and Count-Based PPO

It is clear that the action space defined in (31) includes both discrete and continuous variables. Although conventional DRL algorithms (i.e., policy-based learning or value-based learning) can provide corresponding policies for actions that are either all continuous or discrete, they can not tackle the hybrid action space. To overcome this issue, a basic policy gradient approach is introduced, namely, trust region policy optimization (TRPO) [36]. Then the objective function is rewritten as

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{\boldsymbol{s}} P^{\pi_{\theta_{\text{old}}}} \sum_{\boldsymbol{a}} \pi_{\theta_{\text{old}}}(\boldsymbol{s}, \boldsymbol{a}) \frac{\pi_\theta(\boldsymbol{s}, \boldsymbol{a})}{\pi_{\theta_{\text{old}}}(\boldsymbol{s}, \boldsymbol{a})} A(\boldsymbol{s}, \boldsymbol{a}), \\ &= \mathbb{E}_{\boldsymbol{s} \sim P^{\pi_{\theta_{\text{old}}}}, \boldsymbol{a} \sim \pi_{\theta_{\text{old}}}} \frac{\pi_\theta(\boldsymbol{s}, \boldsymbol{a})}{\pi_{\theta_{\text{old}}}(\boldsymbol{s}, \boldsymbol{a})} A(\boldsymbol{s}, \boldsymbol{a}), \end{aligned} \tag{47}$$

where $\pi_{\theta_{\text{old}}}$ and π_θ are the old and current politics, respectively. The Kullback-Leibler (KL) divergence is used to restrict the step of the policy update in (47) to ensure the training stability of TRPO. We have

$$\mathbb{E}_{\boldsymbol{s} \sim P^{\pi_{\theta_{\text{old}}}}} [D_{KL}(\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{s}) \| \pi_\theta(\cdot|\boldsymbol{s}))] \leq \vartheta, \tag{48}$$

where $D_{KL}(\cdot)$ is the KL divergence function. ϑ is a constant ensures that there is no significant difference between the new and old policies. However, since the second-order optimization of TRPO is inadequate, it is time-consuming to train [37]. As a results, we develop a clip-and-count based Proximal Policy Optimization (PPO) algorithm to train actor-critic networks, which uses a clipping function to ensure that undesirable actions do not corrupt its training. The probability ratio between the current and old policies is given by

$$\Upsilon(\theta) = \frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta_{\text{old}}}(\boldsymbol{a}|\boldsymbol{s})}, \tag{49}$$

where θ denotes the policy parameter, which is updated based on the following loss function,

$$L(\boldsymbol{s}, \boldsymbol{a}, \theta_{\text{old}}, \theta) = \mathbb{E}\left[\min\left(\Upsilon(\theta)\tilde{A}_{\pi_{\theta_{\text{old}}}}(\boldsymbol{s}, \boldsymbol{a}),\right.\right.$$

Algorithm 1: MADRL-Based Association and Beamforming.

Initialize: the actor-critic network; the network parameter settings; the replay memory.
Input: Observation space $\mathcal{O}$; action space $\mathcal{A}$; number of episodes N_{ept}; discount factor γ; network update period T; minibatch size D.
Output: Optimal action sequences on user-UAV association, beamforming and transmit power allocation.
for each episode **do**
  while UAVs are located in the range of the MBS **do**
    Update the rates of user-pairs to UAVs; common and private stream rates; backhaul rate of each UAV
    Obtain an initial state $\boldsymbol{s}_1$
    for $t = 1, 2, \ldots, T$ **do**
      for each UAV-agent **do**
        Observe o_t^m and select action a_t^m through importance sampling the density function
      end for
      Receive a reward r_t^m and transit the next state s_{t+1} for the current action and state
      Each agent executes action $\boldsymbol{a}_t$ and interacts with the environment for receiving their reward $r_{(t+1)}$
      for each agent m **do**
        Calculate the reward function $r_m(t+1)$
        Store the tuple $\{o_t^m, a_t^m, r_m(t+1), o_{t+1}^m\}$ in the experience replay memory
      end for
    end for
  end while
  for each agent m **do**
    Use Algorithm 2 to train actor and critic networks of each agent
  end for
end for

$$\text{clip}(\Upsilon(\theta), 1-\epsilon, 1+\epsilon)\tilde{A}_{\pi_{\theta_{\text{old}}}}\Big)\Big], \quad (50)$$

where $\tilde{A}_{\pi_{\theta_{\text{old}}}}(\boldsymbol{s}, \boldsymbol{a})$ denotes the estimated advantage function; ϵ denotes the threshold and the function $\text{clip}(\Upsilon(\theta), 1-\epsilon, 1+\epsilon)$ is used to indicate that the reward will be canceled if $\Upsilon(\theta)$ is outside $[1+\epsilon, 1-\epsilon]$. However, a fixed clip threshold ϵ will result in poor feasibility of the standard PPO [37]. In order to tackle this issue, ϵ is designed to follow the normal distribution $\epsilon \sim \mathcal{CN}(\hat{\mu}, \hat{\varrho}^2)$, where $\hat{\varrho}$ and $\hat{\mu}$ are the standard deviation and expected value, respectively. Such modification allows agents to be limited to a larger exploration range during training based on the differences between the current and old policies. Since the difference gradually decreases with the training, we add a small in-scope limit to each agent to speed up convergence.

To deal with the instability risk brought by the uncertainty of the model, this paper introduces extrinsic and intrinsic rewards. The former is a discount reward $\tilde{R}(t) = \gamma^{t-1} r(t)$, which is defined in (35). The latter is used to motivate agents to expand their exploration before receiving any extrinsic rewards, which is denoted by

$$\hat{R}(t) = \begin{cases} \hat{\lambda}\frac{1}{C_t}, & \text{if the count } C_t > 0, \\ 0, & \text{otherwise}, \end{cases} \quad (51)$$

where $\hat{\lambda} \in [0, 1]$ is a constant and C_t denotes the total number of counts allocated to the beamforming $\mathbf{w}_m(t)$ corresponding to action $\boldsymbol{a}_t$ before time slot t. There are more counts used to allocate the same beamforming, the less intrinsic reward can be obtained. To this end, each UAV-agent tends to explore the same beamforming with as few counts as possible for a larger cumulative reward, which enhances its exploration capability. Finally, we employ an improved PPO in which the parameters of actor and critic networks are shared. In addition, we add a mean-squared-error term to the value-estimation function (47) to facilitate full exploration. On this basic, the mean-squared-error loss of the critic network is denoted by

$$L(\boldsymbol{s}, \boldsymbol{a}, \phi_{\text{old}}, \phi) = (V_{\phi_{\text{old}}}(\boldsymbol{s}, \boldsymbol{a}) - (\tilde{R}(t) + \hat{R}(t)))^2, \quad (52)$$

where ϕ denotes the value function parameter. Our proposed clip and count-based PPO is summarized as Algorithm 2.

Algorithm 2: Clip and Count-Based PPO.

Input: Initialized policy parameters $\boldsymbol{\theta}_0$ and value function parameters ϕ_0.
for m = 0,1,2,... **do**
  Collect $\{\boldsymbol{s}_m, \boldsymbol{a}_m, \boldsymbol{r}_m\}$ for $m \in \mathcal{M}$
  Calculate extrinsic reward $\tilde{R}(t)$
  Estimate advantage function $A^{\pi_k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ based on the current value function V_{ϕ_m}
  Set $\epsilon \sim \mathcal{CN}(\hat{\mu}, \hat{\varrho}^2)$
  Update the policy parameter using ϵ and (50):
  $\boldsymbol{\theta}_{m+1} = \arg\max_\theta \frac{1}{|\mathcal{D}_m|N} \sum_{\tau \in \mathcal{D}_m} \sum_{t=0}^{T} \min(\Upsilon(\boldsymbol{\theta}) \times \tilde{A}_{\pi_{\theta_m}}(\boldsymbol{s}_t, \boldsymbol{a}_t), g(\epsilon, \tilde{A}_{\pi_{\theta_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t)))$, where $g(\cdot)$ denotes the stochastic gradient policy
  Count C_t and calculate intrinsic reward $\hat{R}(t)$
  Update the value function parameter by C_t and (52):
  $\phi_{m+1} = \arg\min_\phi \frac{1}{|\mathcal{D}_m|T} \sum_{\tau \in \mathcal{D}_m} \sum_{t=0}^{T} (V_\phi(s_t) - (\tilde{R}(t) + \hat{R}(t)))^2$
end for

F. Practical Implementation

1) Communication Signal Between UAVs and Users: In our distributed system, each UAV is responsible for making optimal decisions about decoupling association and beamforming allocation by interacting with the environment. Consequently, we focus on the transmission signals between users and UAVs rather than between MBS and UAVs. In practice, only a small amount of information is needed when calculating the signals between UAVs and users, such as inter-cell interference and channel gain. Particularly, each UAV uses its control channel to interact with the transmitted signals in UL and DL [38].

2) Computational Complexity and Scalability: Our developed MADRL is an actor-critic algorithm and uses centralized

training and distributed execution in each learning episode. In the training phase, the actor network of each agent inputs local observations and then makes an action. As each actor and critic network has three fully connected hidden layers, the computational complexity of centralized training is $\mathcal{O}(\sum_{i=1}^{I} n_i \cdot n_{i-1})$, where n_i denotes the number of neurons in hidden layer i[39].

However, the computational overhead of Algorithm 1 mainly comes from each critic network evaluating the actions of all agents, i.e., decoupling association, beamforming and common rate allocation. According to (31) and (43), the computational complexity of taking action and evaluating the output of each actor network is calculated as $\mathcal{O}(U \cdot M \cdot T)$.

In terms of implementation, this MADRL framework can be easily scaled up as the number of UAVs increases. This is because we use only one experience pool for storing historical experience, and increasing the number of UAVs only requires expanding the size of this experience pool.

V. Simulation Results

This section presents extensive simulation results to demonstrate the performance of our proposed RS-based transmission scheme. We first introduce the simulation setting and network architecture. We then compare the algorithm in this paper with several baselines and analyze experimental results.

A. Simulation Setup

We consider a full-duplex system with $U = 30$ users that are randomly and uniformly distributed within an area of $2 \times 2\,\text{km}^2$. Then 4 UAVs are deployed in fixed positions to assist the MBS to provide services for users.[2] The MBS has $N_a = 6$ antennas, while each UAV has $N_b = 4$ antennas. The maximum transmit powers of the MBS and each UAV and user are set as 43 dBm, 33 dBm and 23 dBm, respectively. The receiver noise power is set as $\sigma^2 = -120$ dBm. The self-interference cancellation capabilities of BSs and users are set as $\frac{1}{\delta_m^2} = \frac{1}{\delta_{z_n}^2} = 100$ dB.

We use a wireless model similar to [41] for BS-to-user links. Therefore, the channel between antenna $\mathcal{N}_a = 1, \ldots, N_a$ of the MBS and user u is expressed as

$$\mathbf{h}_{0,u}^{n_a} = \hbar_{0,u}^{n_a} \sqrt{G_0 \beta d_{0,u}^{-\alpha} \xi_{0,u}}, \tag{53}$$

where $\hbar_{0,u}^{n_a} \sim \mathcal{CN}(0,1)$ is the Rayleigh fading coefficient; G_0 is the MBS antenna gain; $\xi_{0,u}$ is the shadowing coefficient; and $\beta d_{0,u}^{-\alpha}$ accounts for the path-loss effect, which is given by

$$\ell_{0,u}(\text{dB}) = 128.1 + 37.6 \log_{10}(d_{0,u}), \tag{54}$$

where $d_{0,u}$ is the distance between the MSB and user u. The path losses of user-to-user links are given as

$$\ell_{u,u'}(\text{dB}) = 98.4 + 20 \log_{10}(d_{u,u'}). \tag{55}$$

The UAV-to-user wireless channel is dominated by the probabilistic line-of-sight (LoS) and non-line-of-sight (NLoS) links.

2.Similar to [40], users are randomly clustered into 5 multicast groups for downlink transmission

TABLE II
Numerical Calculation Parameter Settings

Description	Symbol	Value
Speed of light	v_c	$3 * 10^8$
Carrier frequency	f_c	2 GHz
Shadowing factor	χ_{LoS}, χ_{NLoS}	6dB, 20dB
Environmental factor	c_1, c_2	11.9, 0.13
Additional path loss factor	η	20 dB
Path loss exponent	α	2
BS antenna gain	G_m	5 dBi
User antenna gain	G_u	0 dBi [43]
Shadowing BS-to-user	$\xi_{m,u}$	10 dB
Shadowing user-to-user	$\xi_{u,u'}$	12 dB
Shadowing BS-to-BS	$\xi_{m,m'}$	6 dB

Thus, the path-loss from UAV m to user u is given by

$$\ell_{m,u} = P_{m,u}^{\text{LoS}} \ell_{m,u}^{\text{LoS}} + P_{m,u}^{\text{NLoS}} \ell_{m,u}^{\text{NLoS}}, \tag{56}$$

where the probabilities of LoS and NLoS links are denoted as $P_{m,u}^{\text{LoS}} = \frac{1}{1+c_1 \exp(-c_2(\theta_{mu}-c_1))}$ and $P_{m,u}^{\text{NLoS}} = 1 - P_{m,u}^{\text{LoS}}$, respectively. Also, c_1 and c_2 are environment-related constants (e.g., rural and dense urban) and $\theta_{mu} = \frac{180}{\pi} \arcsin(\frac{H}{d_{m,u}})$ is the elevation angle. In addition, $\ell_{m,u}^{\text{LoS}} = 20 \log(\frac{4\pi f_c d_{m,u}}{v_c}) + \chi_{\text{LoS}}$ and $\ell_{m,u}^{\text{NLoS}} = 20 \log(\frac{4\pi f_c d_{m,u}}{v_c}) + \chi_{\text{NLoS}}$ are the LoS and NLoS path losses between user u and UAV m, respectively, where v_c is the light speed; f_c denotes the carrier frequency; $d_{m,u}$ is the distance; χ_{LoS} and χ_{NLoS} are shadowing factors. The channel coefficient $\mathbf{h}_{m,u}^{n_b}$, $n_b \in \mathcal{N}_b = 1, \ldots, N_b$ for the UAV-to-user link is described similarly to (53). The UAV-to-UAV channel is dominated by LoS link. Hence, the channel coefficient $\mathbf{h}_{m,m'}^{n_b}$ from UAV m to UAV m' is given by $\mathbf{h}_{m,m'}^{n_b} = \rho d_{m,m'}^{-\alpha}$, where $\rho = -60$ dB is the channel gain at the reference distance $d = 1$ m. As shown in Table II, the parameters related to wireless communication are set according to 3GPP standard [42].

B. Network Architecture

This simulation is performed on a server with an NVIDIA GTX 2080 GPU. The proposed MADRL-based joint optimization algorithm is composed of two neural networks, namely, actor network and critic network, which are trained based on a Python 3.6 platform with PyTorch. In addition, each actor and critic network is built with three hidden layers. All the three hidden layers have an equal number of neurons, i.e., $e = 64$. Each hidden layer neural network is activated based on the rectified linear unit (ReLU) function $f_{\text{ReLU}}(x) = \max\{0, 1\}$. The parameters of each actor and critic network are updated through an Adam optimizer with a learning rate of 0.001. The clip parameter is set to $\epsilon = 0.2$. We set the discount factor used to calculate the expected reward as $\gamma = 0.999$. Two neural networks are trained every $N_{\text{ept}} = 15000$ episodes, while the number of time slots in an episode is set to $T = 250$. The weights ω_1 and ω_2 are set as 40 and 60, respectively. The size of experience replay buffer is set as $\mathcal{B} = 50000$.

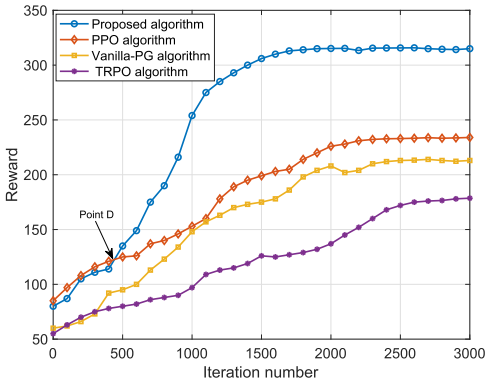

Fig. 4. Convergence of MADRL-based training with different algorithms.

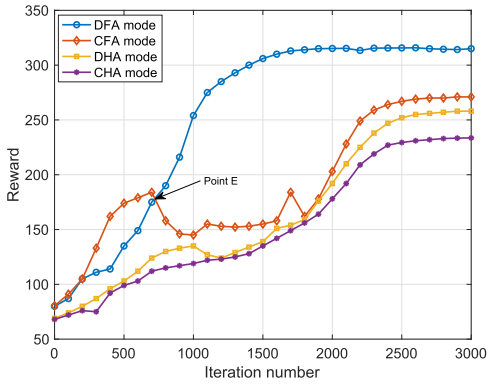

Fig. 5. System performance under different user association modes.

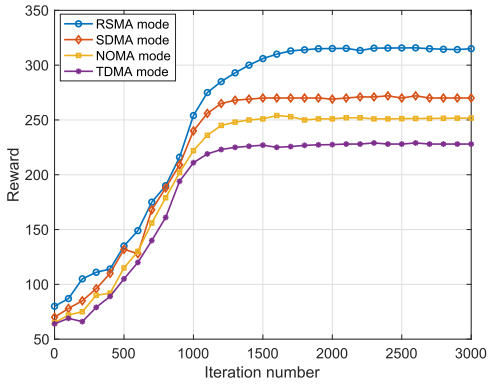

Fig. 6. System performance under different transmissions modes.

C. Result Analysis

1) Comparison of Different Learning Algorithms: To evaluate the effectiveness of the MADRL-based learning framework with the improved PPO algorithm, we consider the following four policy gradient-based RL algorithms:

- Vanilla policy gradient (Vanilla-PG) [44]: It is trained in an on-policy way and a stochastic gradient ascent is used to approximate a high-return policy.
- Trust region policy optimization (TRPO) [36]: It uses an off-policy training manner and the KL divergence is used to control the policy update step for each iteration.
- Standard PPO [37]: It is trained in an off-policy manner and simplifies TRPO based on a clip function.
- Proposed improved PPO: It is trained in an off-policy way and a new clip distribution is proposed to cope with the constraints between old and current policies.

Fig. 4 shows the cumulative reward versus iteration number for the above four algorithms. It is clear that a monotonically increasing reward can be obtained by training the actor-critic network using our proposed algorithm. Comparing the curves of the four algorithms, it is not difficult to find our proposed training algorithm converges after about 1600 iterations. This means that our developed intrinsic reward and clip distribution can efficiently train each actor and critic network. In addition, the cumulative reward of the proposed algorithm is lower than that of the standard PPO algorithm before point D, while it is always maximum after this point. This is due to the increased computational complexity for calculating the intrinsic reward after revising the procedure of the standard PPO. On the other hand, the revised performance gain gradually makes up for the loss of complex computations with the number of iterations.

2) Comparison of Different Association Modes: To evaluate the performance of our proposed DFA association mode, four association modes are considered and listed in the following:

- CHA: For both uplink and downlink, one user associates with the same UAV using time-division half-duplex.
- DHA: For both uplink and downlink, one user associates with two UAVs using time-division half-duplex.
- CFA: One user associates with the same UAV for simultaneous uplink and downlink.
- DFA: One user associates with two different UAVs for simultaneous uplink and downlink.

In Fig. 5, we compare the reward achieved by the above four association modes. It can be observed from Fig. 5 that the reward of DHA is higher than that of CHA, while DFA is not superior to CFA until point E. The reason is that the additional interference generated by the decoupled mode reduces the transmission rate. As the iteration proceeds, the rate gain from the decoupled association is sufficient to compensate for the reduction due to the additional interference. In addition, DFA (CFA) achieves significant higher rewards compared to DHA (CHA). This indicates that the user-UAV association with full-duplex outperforms the one with half-duplex. This is because although the inherent self-interference in simultaneous uplink and downlink reduces the data rate, the total transmission time is halved. This means that correlation costs can be reduced, e.g., by leasing radio resources for associations, which fully compensates for the rate reduction caused by self-interference.

3) Comparison of Different Transmission Modes: To evaluate the performance of our proposed RSMA association mode, we consider the following three baseline transmission modes: SDMA [45]; NOMA [46] and TDMA [26].[3] Fig. 6 depicts the convergence behaviour of our proposed algorithm for different

3.The channel allocation for SDMA, power allocation for NOMA, and time allocation for TDMA are solved by our proposed algorithm.

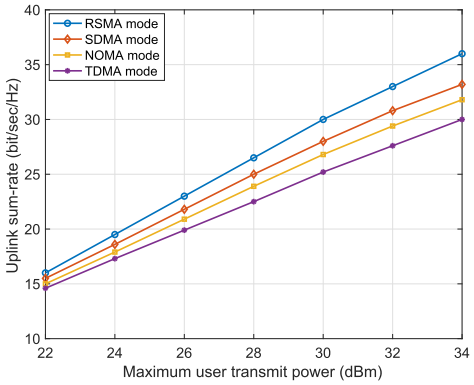

Fig. 7. Uplink sum-rate versus maximum user transmit power.

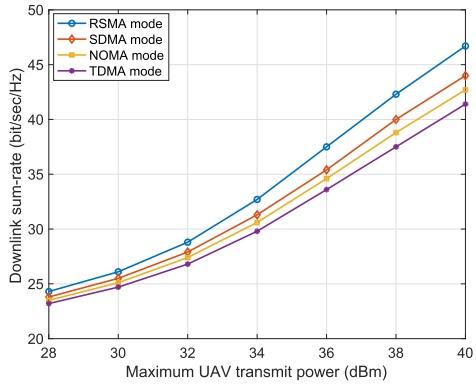

Fig. 8. Downlink sum-rate versus maximum UAV transmit power.

transmission modes. As expected, the rewards achieved by the four transmission modes increase quickly with the number of iterations and eventually converge. It can be observed that the convergence speed of our proposed RSMA mode is slightly slower than that of the other three transmission modes. This is because our proposed RSMA mode results in a large number of common streams, which increases the training complexity of Algorithm 1. However, our proposed RSMA mode achieves the highest reward and increases up to 20.7% compared to the SDMA mode. Such results are able to make up for the loss of computational effort due to rate-splitting precoding.

4) Different Transmit Power: Fig. 7 plots the uplink sum-rate achieved by the various transmission modes versus maximum user transmit power $P_{\max}$. It is clear that the sum-rates of all multiple access modes linearly increase with the maximum transmit power of each user. The reason is that the sum-rate is a logarithmic function of the user transmit power. In addition, our proposed RSMA mode can increase up to 7.14%, 12.3% and 19.6% sum-rate compared to SDMA, NOMA and TDMA for $P_{\max} = 30$ dBm, respectively. The reason is that our proposed RSMA mode can adjust the splitting power of two messages for each strong-user so as to control the interference decoding thus optimizing the sum-rate of all users, while there is no power splitting in other three multiple access modes. As $P_{\max}$ increases, the proposed RSMA mode always achieves the highest sum-rate, while the TDMA mode has the worst sum-rate. In Fig. 8, we depict the downlink sum-rate achieved by the various association modes versus maximum UAV transmit power $P_m^{\max}$. The trend of curves in Fig. 8 is similar with Fig. 7. For $P_m^{\max} = 34$ dBm, our proposed RSMA mode can achieve sum-rate of up to 2.94%, 7.69% and 12.9% higher than those of SDMA, NOMA and TDMA, respectively.

Fig. 9 plots the sum-rate achieved by the various association modes versus maximum UAV transmit power $P_m^{\max}$. It is clear that the sum-rate increases for the four association modes as $P_m^{\max}$ becomes large. For $P_m^{\max} \leq 31$ dBm, the curves of DFA and CFA modes are very close to each other, while the former achieves a little higher sum-rate. Similarly, the curves of DHA and CHA modes are close to each other when $P_m^{\max} \leq 34$ dBm. This is because when $P_m^{\max}$ is small, each user may associate to the same node for UL and DL transmissions. In addition, the

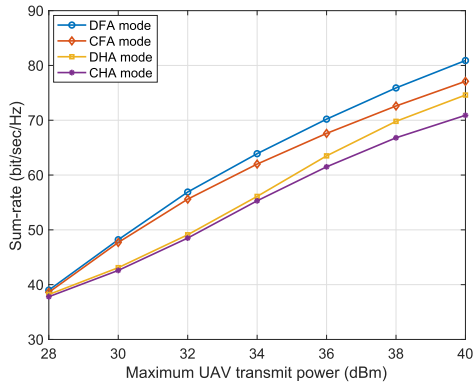

Fig. 9. System sum-rate versus maximum UAV transmit power.

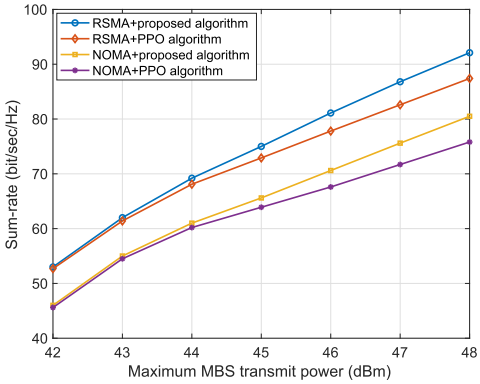

Fig. 10. System sum-rate versus maximum MBS transmit power.

rate gaps between DFA and CFA and between DHA and CHA become large with increasing $P_m^{\max}$. The reason is that as $P_m^{\max}$ increases, a user may achieve better UL rate by associating to a nearby UAV and the same user may receive higher DL rate from other multiple high-power UAVs. Such results show the superiority of decoupled uplink and downlink associations.

Fig. 10 plots the sum-rate achieved by the various transmission modes versus maximum MBS transmit power $P_0^{\max}$. As shown in (26), the maximum backhaul rate is proportional to the maximum MBS transmission power. With limited backhaul rate, our proposed RSMA transmission mode can significantly

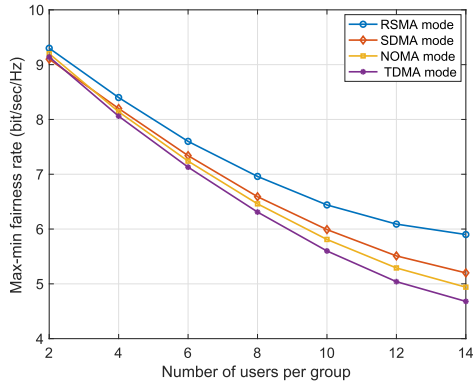

Fig. 11. Max-min fairness rate versus number of users per group.

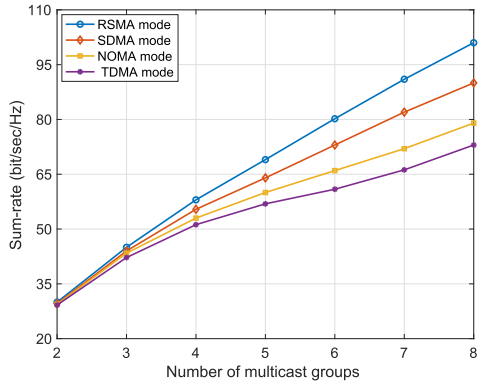

Fig. 12. System sum-rate versus number of multicast groups.

improve the sum-rate. By mitigating the inter-cell interference more efficiently using RS, the performance improvement of our proposed RSMA mode is more obvious as $P_0^{\max}$ increases. When $P_0^{\max} \leq 44$ dBm, the curves of our proposed algorithm and the PPO-based algorithm are close to each other, while the former has a higher sum-rate. However, the rate gap between the two algorithms increases as $P_0^{\max}$ grows. The reason is that the PPO algorithm may converge a near-global optimal policy when the MBS transmit power is almost expanded.

5) Increased Number of Users Per Group: Fig. 11 plots the max-min fairness rate (MMFR) among users in each multicast group for DL versus the number of users per group. It can be observed that the MMFR decreases for the four transmission modes as the number of users per group grows. The reason is that each group has only one precoder for its private stream, so users within a group need to share this precoder even though they all have different channels. Therefore, the user with the worst SINR will then affect its group rate dramatically. Despite this performance degradation, our proposed RSMA mode is still able to provide gains of up to 13.4%, 19.4% and 26% compared to SDMA, NOMA and TDMA for the number of users per group is equal to 14, respectively. The reason is that our proposed RSMA mode enables the receiver to decode the interference partially, while the other three modes treat the interference as noise and neglect their specific characteristics.

6) Increased Number of Multicast Groups: Fig. 12 plots the sum-rate achieved by the various transmission modes versus the number of multicast groups. All the curves in this figure increase with the number of multicast groups. In addition, our proposed RSMA mode can still achieve the highest sum-rate thanks to its ability to alleviate inter-cell interference. However, this also comes with the cost of a large number of common streams, which increases the computational complexity of the proposed algorithm. The performance improvement of RSMA is not significant when fewer multicast groups are scheduled. The reason is that the other three transmission modes are able to neutralize the intergroup interference by carefully steering the precoding. The rate gaps between our proposed RSMA mode and other transmission modes gradually increase when scheduling more groups. The reason for this gap is that when the number of groups exceeds the number of UAV antennas, multiple access modes without RS may saturate. However, our proposed RSMA mode avoids the rate saturation phenomenon. Such results further indicate that rate splitting has a significant impact on the sum-rate improvement.

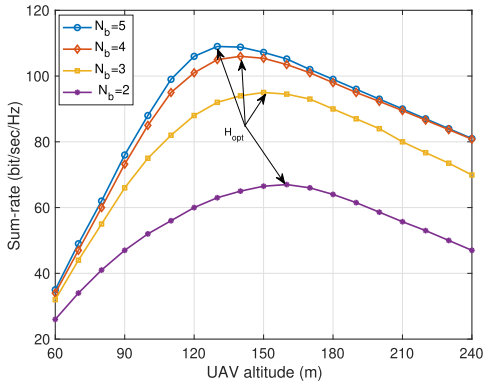

Fig. 13. Sum-rate versus UAV altitude for different number of antennas.

7) Impact of UAV Height on Sum-Rate: In Fig. 13, we show the sum-rate versus UAV altitude H for different number of antennas N_b. In this simulation, we only change the altitude of each UAV. It can be observed that the sum-rate first increases rapidly and then decreases gently for all different number of antennas as H increases. The optimal altitude values leading to a maximum sum-rate are around 130 m, 140 m, 150 m and 160 m for $N_b = 5$, 4, 3 and 2 antennas, respectively. The sum-rate increases as the number of antennas increases, while the curves for the 5 and 4 antenna systems are very close to each other. This means that using more than four antennas to increase the sum-rate is not a viable option. This is caused by UAV power limitations and limited backhaul rates. In addition, increasing the number of antennas helps UAVs to improve the procedure of data stream transmission by carefully designing the beamforming vector. In the regime of $110 \leq H \leq 180$ m, the sum-rate is guaranteed to be greater than 80 bit/sec/Hz for $N_b \geq 3$. This is because higher altitude leads to low channel gains, while lower height cannot guarantee high beanmforming gains.

VI. Conclusion

In this paper, we studied the performance of UDDe association in a full-duplex multi-UAV network. Based on the fact that the decoupled UL-DL association can bring the network new types of interference, we proposed a RSMA policy to mitigate inter-cell interference and formulated a sum-rate maximization problem. To achieve this objective, we jointly optimize the user association with beamforming and message splitting under the constraints of transmit power and backhaul capacity. Due to the resulting problem is non-convex and there exist model uncertainty for an individual agent, we modeled our problem as a robust POMDP and proposed a distributed MADRL-based framework. To motivate agents to continually explore and deal with significant policy deviations due to negative advantaged actions, we proposed a clip and count-based PPO algorithm to solve POMDP. Simulation results shown that our proposed algorithm outperforms traditional learning algorithms in terms of reward and convergence. In addition, our proposed RSMA-based decoupled association scheme achieved significant rate gains over other multi-access schemes. In terms of future work, the proposed idea can be future extended by considering UAV mobility with the resource allocation to improve the sum-rate of UL and DL in full-duplex multi-UAV networks.

References

[1] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, Apr. 2019.

[2] Y. Wu, Y. He, L. Qian, J. Huang, and X. Shen, "Optimal resource allocations for mobile data offloading via dual-connectivity," *IEEE Trans. Mobile Comput.*, vol. 17, no. 10, pp. 2349–2365, Oct. 2018.

[3] H. El 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.

[4] Y. Sun, T. Wang, and S. Wang, "Location optimization and user association for unmanned aerial vehicles assisted mobile networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10056–10065, Oct. 2019.

[5] M. Sami and J. N. Daigle, "User association and power control for UAV-enabled cellular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 267–270, Mar. 2020.

[6] F. Boccardi et al., "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 110–117, Mar. 2016.

[7] M. J. Youssef, J. Farah, C. A. Nour, and C. Douillard, "Full-duplex and Backhaul-constrained UAV-enabled networks using NOMA," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9667–9681, Sep. 2020.

[8] A. M. Fouladgar, O. Simeone, O. Sahin, P. Popovski, and S. Shamai, "Joint interference alignment and bi-directional scheduling for MIMO two-way multi-link networks," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 4126–4131.

[9] Y. Li, W. Ni, H. Tian, M. Hua, and S. Fan, "Rate splitting multiple access for joint communication and sensing systems with unmanned aerial vehicles," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 37–42.

[10] Z. Lin, M. Lin, T. Cola, J. Wang, W. Zhu, and J. Cheng, "Supporting IoT with rate-splitting multiple access in satellite and aerial-integrated networks," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11123–11134, Jul. 2021.

[11] A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, "Rate splitting multiple access in C-RAN: A scalable and robust design," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5727–5743, Sep. 2021.

[12] V. Saxena, J. Jalden, 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, Dec. 2019.

[13] G. Faraci, C. Grasso, and G. Schembra, "Design of a 5G network slice extension with MEC UAVs managed with reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2356–2371, Oct. 2020.

[14] Y. Hsu and R. Gau, "Reinforcement learning-based collision avoidance and optimal trajectory planning in UAV communication networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 1, pp. 306–320, Jan. 2022.

[15] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020.

[16] X. Xi, X. Cao, P. Yang, J. Chen, T. Quek, and D. Wu, "Joint user association and UAV location optimization for UAV-aided communications," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1688–1691, Dec. 2019.

[17] 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, Sep. 2020.

[18] Y. Wang, Y. P. Hong, and W. Chen, "Trajectory learning, clustering and user association for dynamically connectable UAV base stations," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 4, pp. 1091–1105, Dec. 2020.

[19] 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.

[20] Y. Mao, B. Clerckx, and V. K. Li, "Rate-splitting for multi-user multi-antenna wireless information and power transfer," in *Proc. IEEE Signal Process. Adv. Wireless Commun.*, 2019, pp. 1–5.

[21] J. Zhou, Y. Sun, and R. Chen, "Rate splitting multiple access for multi-group multicast beamforming in cache-enabled C-RAN," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12758–12770, Dec. 2021.

[22] A. Rahmati, Y. Yapici, I. Guvenc, and A. Bhuyan, "Energy efficiency of RSMA and NOMA in cellular-connected mm wave UAV networks," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.

[23] Y. Mao, B. Clerckx, and V. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[24] Z. Yang, M. Chen, W. Saad, and B. M. Shikh, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5988–6002, Sep. 2021.

[25] H. Liu, T. Tsiftsis, K. Kim, K. Kwak, and V. Poor, "Rate splitting for uplink NOMA with enhanced fairness and outage performance," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4657–4670, Jul. 2020.

[26] Z. Yang, M. Chen, W. Saad, and W. Xu, "Sum-rate maximization of uplink rate splitting multiple access (RSMA) communication," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2596–2609, Jul. 2022.

[27] H. Kong, M. Lin, Z. Wang, J. Wang, W. Zhu, and J. Wang, "Performance analysis for rate splitting uplink NOMA transmission in high throughput satellite systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 816–820, Apr. 2022.

[28] C. Liu and H. Hu, "Full-duplex heterogeneous networks with decoupled user association: Rate analysis and traffic scheduling," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2084–2100, Mar. 2019.

[29] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.

[30] B. P. Day, A. R. Margetts, D. W. Bliss, and P. Schniter, "Full-duplex MIMO relaying: Achievable rates under limited dynamic range," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 8, pp. 1541–1553, Sep. 2012.

[31] M. Duarte and A. Sabharwal, "Full-duplex wireless communications using off-the-shelf radios: Feasibility and first results," in *Proc. IEEE Conf. Signals Syst. Comput.*, 2010, pp. 1558–1562.

[32] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[33] M. Z. Hassan, M. J. Hossain, J. Cheng, and V. C. Leung, "Device-clustering and rate-splitting enabled device-to-device cooperation framework in fog radio access network," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1482–1501, Sep. 2021.

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

[35] Z. Kaiqing, S. Tao, T. Yunzhe, G. Sahika, M. Sunil, and B. Tamer, "Robust multi-agent reinforcement learning with model uncertainty," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1–20.

[36] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.

[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. Accessed: Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.06347
[38] F. A. Ian, G. David, and C. Elias, "The evolution to 4G cellular systems: LTE-advanced," *Phys. Commun.*, vol. 3, no. 7, pp. 217–244, Aug. 2010.
[39] G. Ian, B. Yoshua, and C. Aaron, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[40] C. Zheng, L. Jemin, Q. Tony, and K. Marios, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.
[41] A. Alameer and A. Sezgin, "Joint beamforming and network topology optimization of green cloud radio access networks," in *Proc. Int. Symp. Turbo Codes*, 2016, pp. 375–379.
[42] Study on Enhanced LTE Support for Aerial Vehicles, document 3GPP TR 36.777, Dec. 2017.
[43] X. Su, L. Li, and P. Zhang, "Rate splitting based asymmetric uplink-downlink cooperative transmission in dynamic TDD MIMO small cell networks," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.
[44] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1999, pp. 1057–1063.
[45] Q. Yang, H. Wang, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: Performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, Oct. 2017.
[46] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349–353, Mar. 2020.

Jiequ Ji received the PhD degree from the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2021. From 2018 to 2020, she was a research assistant with Nanyang Technological University, Singapore, with Prof. Dusit Niyato. She is currently a post-doctoral research fellow with the Department of Electrical and Computer Engineering, University of Victoria, Canada. Her research interests include UAV-enabled wireless communications, wireless content caching, resource allocation in 5G and beyond, mobile edge computing, and physical layer security.

Lin Cai (Fellow, IEEE) received the MASc and PhD degrees (awarded Outstanding Achievement in Graduate Studies) in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2002 and 2005, respectively. Since 2005, she has been with the Department of Electrical and Computer Engineering, University of Victoria, and she is currently a professor. She is an NSERC E.W.R. Steacie Memorial fellow, an Engineering Institute of Canada (EIC) fellow. In 2020, she was elected as a member of the Royal Society of Canada's College of New Scholars, Artists and Scientists, and a 2020 "Star in Computer Networking and Communications" by N2Women. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic and the Internet of Things. She has co-founded and chaired the IEEE Victoria Section Vehicular Technology and Communications Joint Societies Chapter. She has been elected to serve the IEEE Vehicular Technology Society Board of Governors, and served its VP Mobile Radio. She has been a voting board member of IEEE Women in Engineering. She has served as an associate editor-in-chief of *IEEE Transactions on Vehicular Technology*, a member of the Steering Committee of *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Big Data* and *IEEE Transactions on Cloud Computing*, an associate editor of the *IEEE Internet of Things Journal*, *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Communications*, etc., and as the distinguished lecturer of the IEEE VTS Society and the IEEE Communications Society.

Kun Zhu (Member, IEEE) received the PhD degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2012. He was a research fellow with the Wireless Communications Networks and Services Research Group, University of Manitoba, Canada, from 2012 to 2015. He is currently a professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. He is also a Jiangsu specially appointed professor. His research interests include resource allocation in 5G, wireless virtualization, and self-organizing networks. He has published more than fifty technical papers and has served as TPC for several conferences. He won several research awards including IEEE WCNC 2019 Best paper awards, ACM China rising star chapter award.

Dusit Niyato (Fellow, IEEE) received the PhD degree in electrical and computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2008. He is currently a professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has published more than 400 technical articles in the area of wireless and mobile computing. He received the Best Young Researcher Award of the IEEE Communications Society Asia Pacifica and the 2011 IEEE Communications Society Fred W. Ellersick Prize Paper Award. He is also serving as a senior editor of the *IEEE Wireless Communication Letters*, an area editor of *IEEE Transactions on Wireless Communications* and *IEEE Communications Surveys and Tutorials*, an editor of *IEEE Transactions on Communications*, and an associate editor of *IEEE Transactions on Mobile Computing*. He was a distinguished lecturer of the IEEE Communications Society from 2016 to 2017. He was named a highly cited researcher in computer science.