Highlights

Enhancing User Fairness in UAV-Assisted RSMA Networks : A Proximal Policy Optimization Approach

Donghyeon Hur, Donghyun Lee, Manh Cuong Ho, Wonjong Noh, Sungrae Cho

- A UAV-assisted RSMA system is proposed to enhance user fairness in 6G networks.
- The optimization jointly considers UAV trajectory, precoding, and common rate allocation.
- The nonconvex problem is formulated as a Markov decision process and solved via PPO.
- Simulation results show that the proposed approach provides 23.31%, 48.59%, 50.87%, 63.36%, 79.13%, and 145.67% enhanced minimum rates over DDPG, SAC, TRPO, REINFORCE algorithm, Greedy algorithm, and Random strategies, respectively.

Enhancing User Fairness in UAV-Assisted RSMA Networks : A Proximal Policy Optimization Approach

Donghyeon Hur[a], Donghyun Lee[a], Manh Cuong Ho[a], Wonjong Noh[b,*] and Sungrae Cho[a,*]

[a]School of Computer and Engineering, Chung-Ang University, Seoul, 06974, Republic of Korea
[b]School of Software, Hallym University, Chuncheon, 24252, Republic of Korea

ARTICLE INFO

Keywords:
Deep reinforcement learning
Proximal policy optimization
Rate-splitting multiple access
Unmanned aerial vehicles

ABSTRACT

Rate-splitting multiple access (RSMA) and unmanned aerial vehicles (UAVs) are emerging as key technologies for enhancing connectivity and resource efficiency in future 6G networks. This study investigated a UAV-assisted RSMA downlink communication system and developed a novel framework that jointly optimizes the UAV's trajectory, precoding matrix, and common rate to maximize the minimum achievable user rate. The proposed system was formulated as a Markov decision process (MDP) and solved using deep reinforcement learning (DRL), specifically using the proximal policy optimization (PPO) algorithm. This learning-based approach enables UAVs to adapt dynamically to the environment without relying on prior channel state information (CSI), allowing for efficient resource allocation and interference management in complex and dynamic wireless scenarios. Furthermore, a precoding design based on a uniform rectangular array (URA) was employed to enhance directional transmission and spatial multiplexing. In simulations, the proposed method significantly outperformed existing benchmarks, achieving an average minimum rate improvement of 23.31% over Deep Deterministic Policy Gradient (DDPG), 48.59% over Soft Actor-Critic (SAC), 50.87% over Trust Region Policy Optimization (TRPO), 63.36% over REINFORCE algorithm, 79.13% over Greedy algorithm, and 145.67% over Random strategies, respectively. These results confirm the potential of UAV-aided RSMA networks in next-generation wireless environments.

1. Introduction

UAV-enabled wireless networks have been actively investigated owing to their potential for improving quality of service (QoS) in diverse use cases, ranging from emergency response and smart agriculture to military surveillance and temporary event coverage. Their high mobility allows them to adapt to user distributions, minimize communication distances, and circumvent obstacles that degrade signal quality[1, 2]. Additionally, UAVs can operate in three-dimensional space, enabling agile repositioning in real time to satisfy user-centric or network-wide performance objectives [3, 4]. However, the inherent dynamics of UAV platforms introduce new challenges to resource allocation, interference management, and spectrum efficiency in multi-user communication settings.

Advanced multiple-access (MA) techniques are required to address these issues. Spatial division MA (SDMA) leverages spatial separation through precoding to enable simultaneous transmission to multiple users. However, its effectiveness is limited in interference-heavy or high-mobility environments where residual inter-user interference is treated as noise. Non-orthogonal MA (NOMA) improves spectral efficiency by overlapping users in the power domain. However, it places heavy computational burden on receivers owing to successive interference cancellation (SIC), and its performance degrades significantly under imperfect channel state information (CSI) [5, 6]. Recently, rate-splitting MA (RSMA) was introduced as a more generalized and robust framework encompassing SDMA and NOMA as special cases [7]. By splitting messages into common parts for all users and private parts for individual users, RSMA enables flexible interference management. Users can decode part of the interference and treat the rest as noise, leading to enhanced robustness against CSI uncertainty and improved user fairness [8]. A comprehensive survey of the fundamentals and evolution of RSMA, including its theoretical foundation and practical deployments across various wireless scenarios, can be found in [9]. This survey highlighted RSMA's versatility in addressing challenges such as user heterogeneity, imperfect CSI, and dynamic channel conditions. These features make RSMA particularly attractive for UAV-based communication networks, wherein channel variations and user mobility are inherent. Several studies have demonstrated the efficacy of RSMA in UAV-assisted downlink communication in terms of throughput and energy efficiency [10, 11, 12, 13].

However, the following requirements must be met for RSMA to achieve its full potential in such networks: 1) Real-time and accurate control of RSMA-UAV systems: Conventional methods often require perfect channel knowledge and are not well-suited to online adaptation in dynamic network environments. To overcome these limitations, recent efforts have turned toward deep reinforcement learning (DRL), which has shown promise in solving high-dimensional and sequential decision-making problems without the explicit modeling of system dynamics. 2) Advanced precoding designs that go beyond conventional uniform linear arrays (ULAs): Uniform rectangular arrays (URAs) offer improved spatial selectivity in both the azimuth and elevation planes,

*Corresponding author
dhhur@uclab.re.kr (D. Hur); dhlee@uclab.re.kr (D. Lee); hmcuong@uclab.re.kr (M.C. Ho); wonjong.noh@hallym.ac.kr (W. Noh); srcho@cau.ac.kr (S. Cho)

enabling 3D directional precoding that is essential for airborne platforms with high degrees of maneuverability. Notably, URA-equipped UAVs can effectively suppress inter-user interference while targeting ground users more precisely [14].

This study investigated the DRL-based real-time joint control of RSMA and UAV with URA, aiming at optimal user fairness.

1.1. Related Works

Recent studies have focused on fairness-oriented RSMA transmission strategies. For example, Dizdar *et al.* [15] proposed a low-complexity algorithm with iterative precoder design that achieves maximum fairness in overloaded MIMO systems, particularly when the number of users exceeds the number of transmitting antennas. Their method effectively balances complexity and performance by avoiding heavy convex optimization. Similarly, Lee and Shin [16] investigated fairness-aware precoding under imperfect CSI, with robust precoder design based on bounded CSI error models. Xu *et al.* [17] extended the analysis to short-packet transmission scenarios using finite blocklength theory, highlighting the importance of fairness in low-latency RSMA systems. However, these studies mainly focused on quasi-static or low-mobility environments with limited dynamics.

Previously, UAVs have been investigated because of their flexibility, LoS capability, and capability for rapid deployment in wireless systems [1, 2]. Applying RSMA to UAV-assisted networks improves throughput and energy efficiency through the efficient handling of interference in mobile scenarios. For example, Jaafar *et al.* [12] conducted an analytical study on the downlink rate performance in RSMA-based UAV systems, demonstrating its superiority over NOMA and OMA under Rician fading. Xiao *et al.* [18] proposed a traffic-aware, energy-efficient RSMA resource allocation algorithm using successive convex approximation (SCA), balancing latency and energy consumption. Feng *et al.* [19] jointly optimized the UAV trajectory and RSMA resource allocation using an alternating optimization algorithm that separates the spatial and physical layers. Although effective, these approaches rely on convex or heuristic optimization methods that can be computationally expensive and less responsive to real-time or high-mobility environments.

On the other hand, online learning algorithms suffer from slow convergence due to their reliance on real-time feedback and gradual policy updates. This limitation makes them unsuitable for dynamic wireless networks [20]. Additionally, Game-theoretic methods require accurate modeling of user utilities and rely on equilibrium computations. These processes are computationally expensive and assume full knowledge of the environment. Such assumptions do not hold in UAV-based scenarios [21]. To address these limitations, DRL has emerged as a scalable alternative to adaptive UAV communication control. Guan *et al.* [22] developed a multiagent PPO framework for cooperative UAV trajectory optimization in emergency scenarios, enabling UAVs to adaptively serve ground users without centralized control. Zhan *et al.* [23] introduced an enhanced PPO-based algorithm for the distributed control of multiple UAVs using a ray framework, achieving improved scalability and performance in coordinated UAV missions. However, these studies largely treated trajectory planning as an isolated control problem, and did not consider the joint optimization of the trajectory with RSMA-related variables such as precoding and common rate allocation.

Another underexplored aspect is the use of advanced antenna structures in UAV-RSMA systems. Most previous studies considered ULAs, which limit the beam control to a single spatial domain. By contrast, URA supports high-resolution 3D precoding in both the azimuth and elevation dimensions [24]. This capability is critical in UAV platforms that operate in three-dimensional space and require precise directional control for efficient service coverage.

1.2. Motivations and Contributions

Despite the extensive work on RSMA-UAV systems, to the best of the authors' knowledge, this is the first study integrating UAV mobility control, RSMA transmission, URA-based precoding, and PPO-based optimization simultaneously into a unified end-to-end learning framework. The contributions of this study are summarized as follows:

- This study proposes a nonconvex optimization approach that maximizes the minimum achievable rate of users by jointly optimizing the precoding matrix, common rate allocation, and UAV trajectories.
- To effectively address the nonconvex problem, it was reformulated as a Markov decision process (MDP) problem and solved using PPO-based DRL. This approach enables the UAV to dynamically adapt to the environment without relying on prior CSI, allowing for efficient resource allocation and interference management in complex and dynamic wireless scenarios.
- Simulations confirmed that the proposed method outperformed existing DRL methods such as DDPG, SAC, and random methods. Particularly, the proposed approach improved the minimum user rate by up to 23.31%, 48.59%, 50.87%, 63.36%, 79.13%, and 145.67% over the Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), Trust Region Policy Optimization (TRPO), REINFORCE algorithm, Greedy algorithm, and Random baseline strategies, respectively, achieving significant gains in both fairness and adaptability in UAV-based RSMA scenarios.

The rest of this paper is organized as follows: Section 2 describes the system model and problem formulation. In Section 3, the proposed PPO-based optimization approach is introduced. Section 4 presents the simulation results. Finally, Section 5 presents the conclusions drawn from this study.

2. System Model and Problem Formulation

This study considered a scenario wherein a UAV equipped with U antennas flies above N stationary ground users,

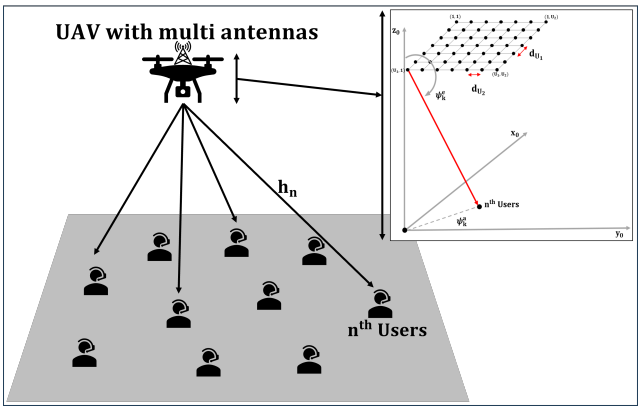

Figure 1: Unmanned aerial vehicle (UAV)-assisted rate-splitting multiple access (RSMA) downlink transmission system in which aerial base station serves N ground users while in motion.

as shown in Fig. 1. The UAV is assumed to have prior knowledge of user positions during each time slot. Communication occurs through an air-to-ground channel, where $n = 0$ represents the UAV, and $n \in [1, N]$ denotes ground users. Time is divided into discrete slots t, where m indicates the movement duration. The UAV's direction is represented by angle $\varphi_n(t) \in [0, 2\pi]$, and its speed is given by $V_n(t) \in [0, V_{\max}]$. After moving for m s, the UAV hovers to serve users for $|t - m|$ s.

2.1. Channel Model

Unlike terrestrial communication, UAV communication is significantly influenced by the altitude and elevation angles in the air-to-ground channel. Therefore, the Rician fading model has been applied to systems that are suitable for aerial base stations and provide a strong LoS [25]. The channel between the UAV and user n was modeled using a block-fading Rician channel, as follows [26]:

$$\mathbf{h}_n = L_n \left(\sqrt{\frac{n^*}{n^*+1}} \mathbf{h}_n^{LOS} + \sqrt{\frac{1}{n^*+1}} \mathbf{h}_n^{NLOS} \right), \quad (1)$$

where L_n denotes the large-scale path loss, n^* is the Rician factor, and $\mathbf{h}_n^{\mathrm{NLOS}}$ follows the distribution $\mathcal{CN}(0, 1)$. A small urban environment was considered to allow the use of the Hata model, which was used as a large-scale fading model and is defined as follows: [27]:

$$\begin{aligned} L_{un} =& 69.55 + 26.16 \log_{10} f \\ &+ \left[44.9 - 6.55 \log_{10} h_b\right] \log_{10} d \\ &- 13.82 \log_{10} h_b - C_H, \end{aligned} \quad (2)$$

$$C_H = (1.1 \log_{10} f - 0.7) h_M + 0.8 - 1.56 \log_{10} f, \quad (3)$$

where C_H, h_b, h_M, f, and d denote the antenna height correction factor, base station antenna height, mobile station antenna height, transmission frequency, and distance between the base station and user, respectively. $\mathbf{h}_n^{\mathrm{LOS}}$ is defined as follows [24]:

$$\mathbf{h}_n^{\mathrm{LOS}} = \mathbf{a}_n(\psi_n^e, \psi_n^a) = \mathrm{vec}\left(\mathbf{a}_{U_1}(\rho) \mathbf{a}_{U_2}^T(\zeta) \right), \quad (4)$$

where $\mathbf{a}_n(\cdot) \in \mathbb{C}^{U \times 1}$ represents the steering array, ψ_n^e is the elevation angle, and ψ_n^a is the azimuth angle. Additionally, vec$(\cdot)$ maps the $U_1 \times U_2$ matrix; $\mathbf{a}_{U_1}(\rho)$ and $\mathbf{a}_{U_2}(\zeta)$ are calculated as follows:

$$\mathbf{a}_{U_1}(\rho) = \left[1, e^{j\rho}, \ldots, e^{(U_1-1)j\rho}\right]^T, \quad (5)$$

$$\mathbf{a}_{U_2}(\zeta) = \left[1, e^{j\zeta}, \ldots, e^{(U_2-1)j\zeta}\right]^T, \quad (6)$$

where

$$\rho = \frac{2\pi}{\lambda} d_{U_1} \cos\psi_n^a \sin\psi_n^e = \frac{2\pi}{\lambda} d_{U_1} \frac{x_0 - x_n}{d_{0,n}}, \quad (7)$$

$$\zeta = \frac{2\pi}{\lambda} d_{U_2} \sin\psi_n^a \sin\psi_n^e = \frac{2\pi}{\lambda} d_{U_2} \frac{y_0 - y_n}{d_{0,n}}, \quad (8)$$

where λ denotes the wavelength, and d_{U_1} and d_{U_2} represent the distances between two consecutive vertical and horizontal antennas, respectively. The parameters ρ and ζ represent the vector from the elevation and azimuth angles, respectively. The distance between the UAV, which serves as an aerial base station, and the n-th user is denoted by $d_{0,n}$. The subsequent locations of the UAV are denoted by x_n and y_n.

2.2. One-layer RSMA Model

In this study, a one-layer RSMA was used, and the detailed design of RSMA is as follows. Messages intended for users are divided into common and private components. The common message shared among all n users is embedded into the common stream s_0, whereas the private message for IoT device n is represented as $\{s_n\}$. The stream vector $\mathbf{s}$ and precoding matrix $\mathbf{P}$ are defined as follows:

$$\mathbf{s} = \left[s^c, s_1^p, \ldots, s_N^p\right]^T \in \mathbb{C}^{(N+1)\times 1}, \quad (9)$$

$$\mathbf{P} = \left[\mathbf{p}^c, \mathbf{p}_1^p, \ldots, \mathbf{p}_N^p\right] \in \mathbb{C}^{U \times (N+1)}, \quad (10)$$

The transmitted signal is expressed as follows:

$$\mathbf{x} = \mathbf{p}^c s^c + \sum_{n \in \mathcal{N}} \mathbf{p}_n^p s_n^p. \quad (11)$$

Then, the signal received by the nth user is defined as follows:

$$y_k = \mathbf{h}_n^H \mathbf{p}^c s^c + \sum_{j \in \mathcal{N}} \mathbf{h}_j^H \mathbf{p}_j^p s_j^p + n_n, \quad (12)$$

where additive white Gaussian noise n_n follows the distribution $\mathcal{CN}(0, \sigma^2)$. The UAV's transmit power is subject to the constraint shown in the figure above, considering real-world conditions $\sum_{n \in \mathcal{N}} \|\mathbf{p}_n\|^2 \leq P_t$, where P_t represents the maximum allowable transmission power. The receiver initially treats all private streams as interference to decode

the common stream. Subsequently, user n decodes its private stream while treating the remaining private streams as noise. Finally, each user combines the decoded common and private messages. The signal-to-interference-plus-noise ratios (SINRs) are as follows:

$$\gamma_n^c = \frac{\left|\mathbf{h}_n^H \mathbf{p}^c\right|^2}{\sum_{j\in\mathcal{N}} \left|\mathbf{h}_j^H \mathbf{p}_j^p\right|^2 + \sigma^2}, \tag{13}$$

$$\gamma_n^p = \frac{\left|\mathbf{h}_n^H \mathbf{p}_n^p\right|^2}{\sum_{j\in\mathcal{N}, j\neq n} \left|\mathbf{h}_n^H \mathbf{p}_j^p\right|^2 + \sigma^2}, \tag{14}$$

where γ_n^c and γ_n^p denote the SINRs for user n associated with the common and private messages, respectively. The achievable rates are as follows:

$$R_n^c = \log_2(1+\gamma_n^c), \tag{15}$$

$$R_n^p = \log_2(1+\gamma_n^p), \tag{16}$$

where R_n^c and R_n^p denote the achievable rates for user n associated with the common and private messages, respectively.

A common message should be decoded by all users; therefore, it should be sent at the following rate:

$$R_c = \min(R_1^c, \dots, R_N^c), \tag{17}$$

which is referred to as the common rate R_c. Here, R_c can be further divided among users through a common rate-allocation vector $\mathbf{c} = [C_1, C_2, \dots, C_N]$, where C_n represents the portion of the common message rate assigned to user n that satisfies the following expression:

$$\sum_{n\in\mathcal{N}} C_n \leq R_c, \tag{18}$$

Finally, the total achievable rate for user n becomes the sum of the allocated common and private rates.

$$R_n = C_n + R_n^p. \tag{19}$$

2.3. Problem Formulation

This study aimed to maximize the minimum achievable rate to ensure user fairness. The optimization jointly considers the precoding matrix $\mathbf{P}$, common rate allocation vector $\mathbf{c}$, and UAV trajectory parameters $\varphi_0(t)$ and V_0. Based on the system model, the optimization problem can be formulated as follows:

$$\max_{\mathbf{P}(t),\mathbf{c}(t),\varphi_0(t),V_0(t)} \min_{n\in\mathcal{N}} R_n, \tag{20}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{N}} \|\mathbf{p}_n\|^2 \leq P_t, \tag{21}$$

$$\sum_{n\in\mathcal{N}} C_n \leq \min(R_1^c, \dots, R_N^c), \tag{22}$$

$$C_n \geq 0, \quad \forall n \in \mathcal{N}, \tag{23}$$

$$0 \leq \varphi_0 \leq 2\pi, \tag{24}$$

$$0 \leq V_0 \leq V_{\max}. \tag{25}$$

In this formulation, the transmission power constraint is given in (21), while constraint (22) ensures that all users can successfully decode the common stream. The constraint (23) ensures that the rate has a positive value. The constraints (24) and (25) define the allowable ranges of the directional angle and UAV speed, respectively.

The defined optimization problem is not easy to solve using a general method. First, joint optimization over the precoding matrix, common rate allocations, and UAV trajectory parameters are highly coupled. Second, because of the nonlinear dependence of SINRs on the precoding matrix and UAV locations, the problem is nonconvex and lacks a closed-form solution. Conventional optimization methods can be used to decompose the problem into a precoding matrix subproblem, common rate allocation subproblem, and UAV mobility subproblem. Subsequently, these problems are solved iteratively. This approach requires many iterations and inevitably entails suboptimality from the convexification of a nonconvex user achievable rate expression. Therefore, it does not satisfy the requirements for real-time optimal adaptation. However, DRL techniques can efficiently handle nonconvex, dynamic, and high-dimensional optimization settings [28]. These advantages of DRL techniques motivated this study to adopt the DRL approach.

3. Proposed Approach

3.1. MDP Formulation

To solve the problem using DRL, we reformulated (20) as an MDP. The detailed definitions of the MDP components are as follows:

1. *State Space ($\mathcal{S}$)*: The state s_t at time t is defined as follows: $s_t = (x_0(t), y_0(t), \{(x_k, y_k)\}_{k=1}^N, R_c(t-1))$. Here, $(x_0(t), y_0(t))$ denote the UAV's coordinates, (x_k, y_k) are the fixed ground user positions, and $R_c(t-1)$ is the common rate at the previous step. Thus, the state space is a $(2+2N+1)$-dimensional real vector.
2. *Action Space ($\mathcal{A}$)*: The action a_t at time t is defined as $a_t = (\mathbf{P}(t), \mathbf{c}(t), \varphi_0(t), V_0(t))$. Here, $\mathbf{P}(t)$ is the precoding matrix, $\mathbf{c}(t)$ is the common rate allocation, $\varphi_0(t)$ is the UAV's directional angle, and $V_0(t)$ is the UAV's speed. Each precoding element p_n in $\mathbf{P}(t)$ is represented by its real and imaginary parts as $p_n = \Re(p_n) + j\Im(p_n)$.
3. *Reward Function ($\mathbf{R}$)*: The objective is to maximize the minimum reward across users while satisfying the given constraints. The reward function $R : \mathcal{S} \times \mathcal{A}$ is defined based on the objective function, providing an instantaneous reward at each step.

$$r(t) = \min_{n\in\mathcal{N}} R_{n,t}, \tag{26}$$

where $\min_{n\in\mathcal{N}} R_{n,t}$ represents the minimum rate among all users at time slot t. Here, the agent selects actions to maximize the cumulative expected reward based on the defined objective.

Based on this formulation, we propose a DRL approach to determine the optimal solutions.

3.2. Proposed Algorithm

In the actor-critic framework of PPO, the actor network learns a policy distribution parameterized by $(\mu(t), \tau^2(t))$. This defines a policy $\pi_\theta(a(t)|s(t))$ that selects actions $a(t)$ based on state $s(t)$. Considering a realistic scenario wherein the UAV moves continuously, the policy for a continuous action space was designed as a Gaussian distribution. In the PPO algorithm, this policy function is parameterized by θ, which allows adjustments through learnable parameters to optimize the action selection for a given state. For the precoding action, the nth precoding vector $\mathbf{p}_k$ is redefined as follows:

$$\mathbf{p}_n = \left[\alpha_1, \alpha_2, \ldots, \alpha_{2L-1}, \alpha_{2L}\right]^T \in \mathbb{R}^{2L}, \tag{27}$$

where α ranges from 0 to 1. The true scaled value of the precoding action is calculated as follows:

$$\mathfrak{R}(\mathbf{p}_{n,l}) = \frac{P_t}{\sqrt{\Upsilon_n}}\alpha_{2l-1}, \quad \mathfrak{I}(\mathbf{p}_{n,l}) = \frac{P_t}{\sqrt{\Upsilon_n}}\alpha_{2l}, \tag{28}$$

$$\Upsilon_n = \sum_{l=1}^{2L} \alpha_l, \quad \forall n \in \mathcal{N}. \tag{29}$$

Regarding the common rate action $\mathbf{c}$, the softmax function was applied to ensure that the sum of its N elements is equal to 1:

$$\bar{\mathbf{c}} = \left[\beta_1, \ldots, \beta_N\right]^T \in \mathbb{R}^N. \tag{30}$$

The true scaled value of the common rate is computed as follows:

$$R_c = \min\left(R_1^c, \ldots, R_N^c\right)\left(\sum_{n=1}^{N} \beta_n\right). \tag{31}$$

The UAV's speed and direction angle functions are calculated as follows:

$$V_0 = \bar{V}_0 V_{\max}, \quad \varphi_0 = \bar{\varphi}_0 2\pi. \tag{32}$$

Therefore, the original action determined by the policy at each step is redefined as follows:

$$\bar{a}(t) = \left\{\alpha_1(t), \ldots, \alpha_{2L}(t), \beta_1(t), \ldots, \beta_N(t), \bar{V}_0(t), \bar{\varphi}_0(t)\right\}. \tag{33}$$

To update the actor network parameters, a batch of past experiences $B = (s(t), a(t), r(t), s(t+1))$ is sampled from the replay memory, where $s(t+1)$ represents the next state after performing action $a(t)$ in state $s(t)$.

In this implementation, the actor and critic networks were parameterized independently. The overall objective function integrates the following three components: the clipped surrogate objective stabilizes policy updates, value function loss for accurate state value estimation, and entropy bonuses to encourage exploration. The combined loss function is expressed as follows:

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\phi) + c_2 S[\pi_\theta(s(t))]], \tag{34}$$

First, PPO optimizes a clipped surrogate objective $L_B^{\text{CLIP}}(\theta)$, ensuring stable updates [29]:

$$L_B^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right], \tag{35}$$

where probability ratio $r_t(\theta) = \frac{\pi_\theta(a(t)|s(t))}{\pi_{\theta_{\text{old}}}(a(t)|s(t))}$ compares the new and old policies. The clipped objective in (35) constrains $r(\theta)$ to the range $[1-\epsilon, 1+\epsilon]$. This restriction enhances stability by preventing excessively large updates. To estimate the advantage function accurately, we this study adopted the generalized advantage estimation (GAE) method proposed in [30].

$$\hat{A}_t = \delta_t + (\gamma\iota)\delta_{t+1} + \cdots + (\gamma\iota)^{T-t-1}\delta_{T-1}, \tag{36}$$

where δ_t denotes the temporal difference (TD) error, γ is the discount factor, and ι is the weighting parameter for future rewards. T-step rewards are used as alternatives to recurrent network structures, enabling the agent to consider future rewards over multiple steps and learn strategies that maximize cumulative rewards.

Second, L_t^{VF} denotes the squared-error loss, which is defined as $(V_\phi(s(t)) - V_t^{\text{target}})^2$. Here, $V_\phi(s(t))$ and V_t^{target} represent the estimated state value and the actual reward of the advantage function $\hat{A}_t$, respectively.

Third, the entropy bonus S promotes exploration, whereas c_1 and c_2 are weighting coefficients balancing the policy and value updates.

The actor network updates its parameters θ using the clipped loss L^{CLIP}, as follows:

$$\theta_{\text{new}} = \arg\max_\theta \mathbb{E}_t\left[L_t^{\text{CLIP}}(\theta) - c_1 \cdot S[\pi_\theta(s(t))]\right]. \tag{37}$$

By maximizing this loss function, the actor network parameter θ is updated.

The critic network parameter ϕ is updated by minimizing L^{VF}:

$$\phi_{\text{new}} = \arg\min_\phi \mathbb{E}_t\left[\left(V_\phi(s(t)) - V_t^{\text{target}}\right)^2\right]. \tag{38}$$

The overall flow of the proposed PPO algorithm is presented in Algorithm 1.

Fig. 2 illustrates the PPO framework. The actor network takes the current observation as the input, including the UAV position, user locations, and previous common rate. Then, it outputs continuous actions for precoding, common-rate allocation, and UAV speed and direction. These actions are sampled from a Gaussian policy and scaled to satisfy the system constraints. The critic network receives the same input and estimates the state value to guide the computation. Both networks share a multilayer perceptron (MLP)

Algorithm 1 Proximal policy optimization (PPO) with actor-critic structure in rate-splitting multiple access (RSMA)-based unmanned aerial vehicle (UAV) network.

Initialize policy parameters θ, value network parameters ϕ
Initialize replay buffer $\mathcal{B}$, learning rate lr, discount factor γ, clipping factor ϵ
for each episode $e = 1$ to E **do**
  Reset environment and observe initial state s_0
  for each timestep $t = 1$ to T **do**
    Execute action $a(t)$ and observe reward r_t, next state s_{t+1}
    Store transition $(s(t), a(t), r(t), s(t+1))$ in $\mathcal{B}$
  end for
  Compute advantage estimates $\hat{A}_t$ using GAE (Eq. 36)
  Compute targets V_t^{target}
  for each update step **do**
    Compute clipped loss $L_t^{\text{CLIP}}(\theta)$ using Eq. (35)
    Compute value loss $L_t^{VF}(\phi)$ and entropy bonus $S[\pi_\theta(s_t)]$
    Compute total loss $L_t^{\text{CLIP+VF+S}}$ Eq. (34)
    Update θ and ϕ
  end for
end for
Return optimized policy π_{θ^*} and value function V_{ϕ^*}

structure with two hidden layers. Each layer is followed by layer normalization and ReLU activation to improve learning stability. The actor ends with sigmoid activation to constrain the output range and maintain a trainable log-standard deviation for action variance modeling. The critic ends with a single linear output for value estimation. The advantage estimates are computed using GAE. The actor is updated using a clipped surrogate objective and the critic minimizes the value function loss. An entropy bonus is also added to encourage exploration. This end-to-end framework enables stable and efficient learning in highly dynamic UAV-RSMA environments.

Remark. This study selected PPO over DDPG [31], SAC [32], TRPO [33] and REINFORCE algorithm [34] for the following reasons. First, PPO is an on-policy algorithm that allows multiple updates from a single batch of collected data. Consequently, sample usage is improved without requiring complex replay buffers, stabilizes policy updates using a clipped surrogate objective, prevents large changes, and improves robustness in dynamic and high-dimensional environments [29]. Second, PPO reduces implementation complexity and computational overhead by avoiding the need for replay buffers as required in DDPG, entropy tuning used in SAC, second-order optimization demanded by TRPO, and high-variance gradient estimation typical of REINFORCE algorithm, thereby delivering a more stable and scalable solution for our problem [35, 36, 37]. Therefore, given the complexity of RSMA-UAV control and continuous action spaces and the requirement for reliable and scalable learning, PPO is better suited to this study compared with DDPG, SAC, TRPO and REINFORCE algorithm.

4. Evaluations

This section presents an analysis on the complexity of the proposed algorithm and discusses the evaluation of the convergence and achievable maximum–minimum rate. For the simulation, the system parameters were determined based on previous studies [27, 29, 38] and are listed in Table 1.

Table 1
System parameters

System parameter	Value
Network Size [27]	500 m × 500 m × 50 m
Large-scale fading, L_{hk} [27]	$112.472 + 33.772 \cdot \log_{10}(d_k)$
UAV maximum speed	50 m/s
Rician factor, n^* [38]	10
Transmission frequency [27]	1500 MHz
Bandwidth [27]	1 MHz
Noise variance [38]	-64 dBm/Hz
Buffer capacity, C	$1e5$
Batch size, B	128
Clipping coefficient, ϵ [29]	0.2
Value function loss term, c_1 [29]	0.5
Entropy bonus term, c_2 [27]	$1e$-3
Learning rate, lr	$1e$-3
Discount factor for GAE, γ [29]	0.95

The proposed algorithm was compared with the following benchmark schemes.

- **PPO-NOMA**: This scheme adopts the PPO algorithm under a NOMA-based downlink transmission framework. The agent is trained to optimize the UAV trajectory, power allocation, and decoding order over 3000 episodes.
- **PPO-SDMA**: This scheme applies PPO to a conventional SDMA-based system. Beamforming vectors and UAV positions are jointly optimized during 3000 training episodes.
- **DDPG**: This scheme employs the DDPG algorithm to jointly optimize the UAV trajectory, RSMA precoding, and common rate allocation. The model was trained for 3000 episodes.
- **SAC**: This scheme uses the SAC algorithm under the same system configuration. The model was trained for 3000 episodes.
- **TRPO**: This scheme uses the TRPO algorithm under the same system configuration. The model was trained for 3000 episodes.
- **REINFORCE**: This scheme uses the REINFORCE algorithm under the same system configuration. The model was trained for 3000 episodes.

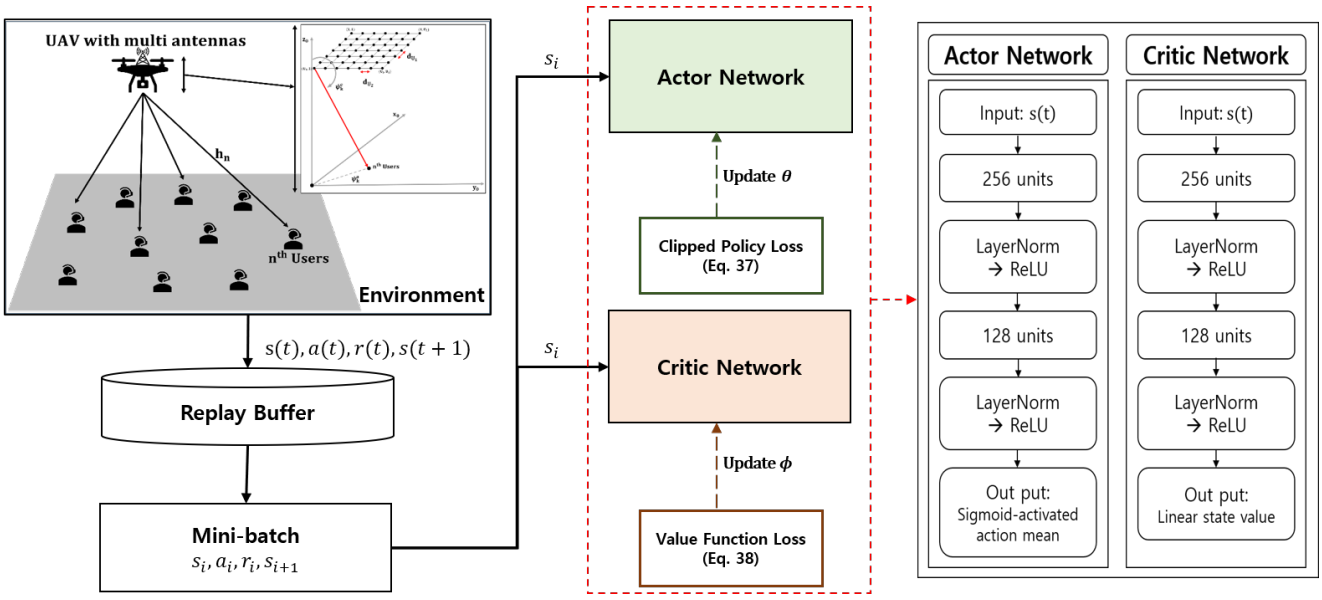

Figure 2: Proposed proximal policy optimization (PPO) framework based on actor-critic architecture.

- **Greedy**: This scheme uses the greedy algorithm under the same system configuration. The model was trained for 3000 episodes.
- **Random**: This non-learning baseline selects UAV positions and precoding vectors randomly at each time step. The minimum user rate was averaged over 50 independent trials.

Simulations were conducted on a server equipped with an Intel Core i7-10700F CPU, Nvidia RTX 2070 SUPER GPU, and 32 GB of memory using Python 3.9.13, PyTorch 2.3.1, and CUDA 11.8.

4.1. Complexity Analysis

During the training phase, the PPO agent updates its actor and critic networks using forward and backward propagations. Each network comprises two hidden layers. The time complexity of a single forward or backward pass is given by $O(SH_1 + H_1H_2 + H_2A)$, where H_1 and H_2 denote the number of neurons in the first and second hidden layers, respectively; $S = 2N + 3$ is the state dimension consisting of the UAV position, user positions, and previous common rate; $A = 2NL + N + 2$ is the action dimension consisting of precoding vectors, rate allocations, and UAV control variables. Here, N is the number of users and L is the number of UAV antennas. Considering that PPO performs E epochs of updates per episode with a mini-batch size B over T time steps and that both the actor and critic networks are trained, the total training complexity is $O(2TBE(SH_1 + H_1H_2 + H_2A))$. Factor 2 accounts for the independent training of both actor and critic networks. During the execution phase, only the actor network performs forward propagation, resulting in a per-step inference complexity of $O(SH_1 + H_1H_2 + H_2A)$. Therefore, the proposed PPO-based RSMA framework exhibits polynomial time complexity in both the training and execution phases while jointly optimizing precoding, rate allocation, and UAV control in a high-dimensional environment.

4.2. Convergence

Fig. 3 shows the average reward convergence during the training episodes for different learning rates ($\gamma \in \{0.0001, 0.0005, 0.001, 0.005\}$). Among the tested values, the learning rate of 0.001 achieved the most stable and highest reward convergence, indicating an appropriate balance between learning speed and stability. By contrast, high learning rates, such as 0.005, led to unstable learning behavior, with the average reward stagnating at approximately 0.02. Consequently, the learning rate of 0.001 was adopted in the proposed framework.

Fig. 4 shows the average reward and its convergence behavior over training episodes for different numbers of users ($N \in \{4, 6, 8, 10\}$). In this simulation, the number of antenna and transmit power are set to 25 and 15 dBm, respectively. The scenario with 4 users achieved the fastest convergence and the highest reward, consistently improving over episodes and reaching an average reward of approximately 0.09. For more than 4 users, there is a clear degradation in convergence speed and final reward values. For 6 users, the average reward is stabilized around 0.035, indicating moderate convergence performance. As the number of users increased to 8, the reward fluctuated more and saturated near 0.02. The scenario with 10 users showed the poorest performance, with the reward stagnating below 0.01.

Fig. 5 shows the average reward and its convergence behavior over training episodes for different UAV speeds ($V \in \{20, 30, 40, 50\} m/s$). In this simulation, the number of antenna and transmit power are set to 25 and 15 dBm,

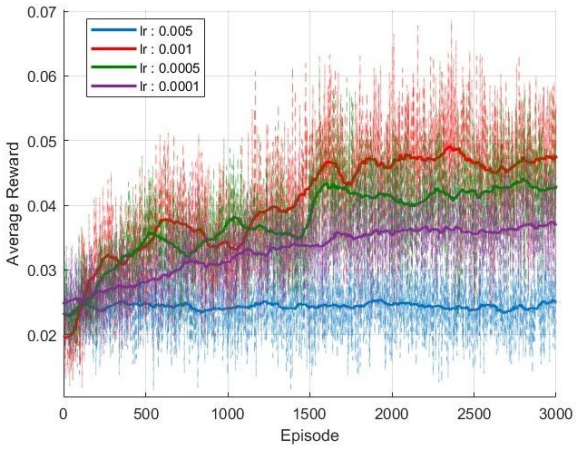

Figure 3: Reward convergence for different learning rates.

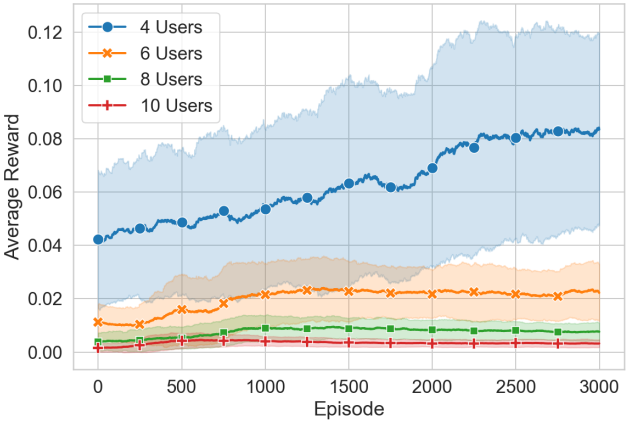

Figure 4: Reward convergence for different number of users.

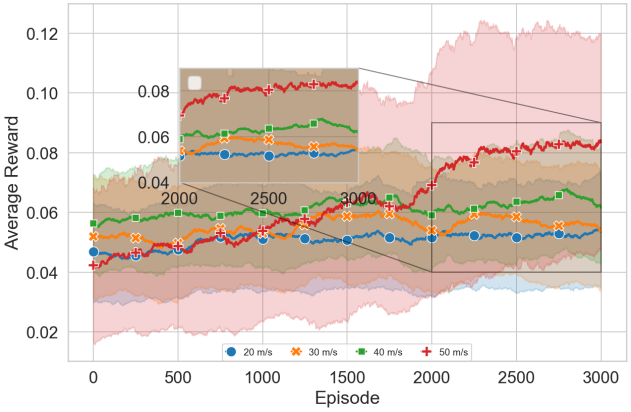

Figure 5: Reward convergence for different UAV speeds.

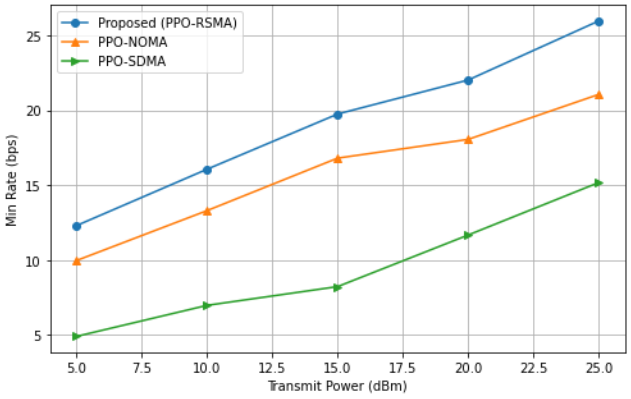

Figure 6: Minimum rate versus transmit power for different multiple access technologies.

respectively. Among the different speeds, the scenario with 50 m/s achieved the fastest convergence and the highest reward, consistently improving over episodes and reaching an average reward of approximately 0.1. As the UAV speed decreased, there was a noticeable degradation in both convergence rate and final reward performance. For 40 m/s, the average reward saturated around 0.06, indicating moderate performance. At 30 m/s and 20 m/s, the learning curves became flatter, and the rewards stabilized near 0.055 and 0.05, respectively, with increased fluctuations and slower convergence. These results suggest that higher UAV mobility facilitates faster and more effective learning in the considered environment, likely due to enhanced spatial diversity and improved LoS connectivity.

4.3. Comparison with Multiple Access Techniques

To demonstrate the superiority of the RSMA technique in the proposed framework, the PPO-based RSMA was compared with the PPO-based NOMA and SDMA schemes. Fig. 6 presents the minimum achievable user rate versus the transmit power levels ranging from 5 to 25 dBm. Across all power levels, PPO-RSMA consistently outperformed PPO-NOMA and PPO-SDMA. For example, at 10 dBm, PPO-RSMA achieved a minimum rate of approximately 15.5 bps, compared with 13 bps for PPO-NOMA and 9 bps for PPO-SDMA. At 25 dBm, the gap became more pronounced, with PPO-RSMA reaching approximately 25 bps, while PPO-NOMA and PPO-SDMA achieved approximately 21 bps and 17 bps, respectively. This performance gain is attributed to RSMA's inherent ability to split messages into common and private parts, enabling partial interference decoding. Such flexibility improves interference management and ensures a more balanced rate allocation among users. By contrast, NOMA relies on SIC, which is sensitive to channel estimation errors and may lead to degraded performance. Moreover, SDMA depends on spatial separation, which becomes less effective when users have correlated channels or limited antenna resources. These results confirm that RSMA, when combined with PPO-based control, provides more robust and fair performance across various power levels.

4.4. Comparison with Different DRL Algorithms

To validate the effectiveness of the PPO algorithm in high-mobility UAV environments, the proposed PPO-based method was compared with other DRL algorithms including DDPG, SAC, TRPO, REINFORCE, greedy algorithm and random policy baseline. Fig. 7 shows the minimum achievable user rate under different learning algorithms as the transmit power increased from 5 to 25 dBm. As can be seen, the proposed PPO method consistently outperformed all baseline methods across all power levels. At 25 dBm, it achieved a minimum rate of approximately 26 bps, while DDPG, SAC, TRPO, REINFORCE, and greedy methods reached around 21, 17, 17, 16, and 15 bps, respectively. The random policy remained below 11 bps even at higher power levels. These results validate the effectiveness of PPO's policy optimization in complex control tasks. The clipped surrogate objective enables stable and reliable policy updates, while structured exploration improves learning efficiency in continuous action spaces. These characteristics make PPO particularly well-suited for proposed dynamic UAV-assisted RSMA systems.

Remark. On the other hand, existing studies [39, 40] demonstrated that user demand imbalance can significantly affect system fairness and performance. The proposed PPO-based framework can be extended to incorporate task load heterogeneity without modifying the core optimization structure. By assigning user-specific weights (corresponding to each user's traffic load) to the reward function, the learning process can prioritize weighted fairness-aware scheduling, as proposed in [41], [42].

5. Conclusion

This study proposed a DRL-based framework for optimizing UAV-assisted RSMA downlink systems equipped with URA antennas. To enhance communication fairness, a max-min fairness problem that jointly optimizes the UAV trajectory, RSMA precoding matrix, and common rate allocation was formulated. To solve the joint optimization problem, a PPO-based algorithm was proposed. This algorithm supports highly dynamic UAV scenarios without relying on prior CSI and URA-based precoding, allowing for efficient resource allocation and interference management in complex and dynamic wireless scenarios. Simulation results revealed that the proposed PPO-based RSMA framework consistently outperformed baselines such as DDPG, SAC, TRPO, REINFORCE algorithm, greedy algorithm and random selection in terms of user fairness. In other words, the proposed method significantly improved the minimum user achievable rate compared with the other benchmarks.

In future work, the proposed framework will be extended to multi-UAV cooperative RSMA environments, and user mobility models will be considered. Also, hybrid techniques combining mathematical optimization-based methods and machine learning methods will be developed to achieve more accurate and faster control. Furthermore, asymmetric traffic scenarios will be applied to reflect heterogeneous user load distributions, enabling the proposed framework to optimize under more realistic network conditions.

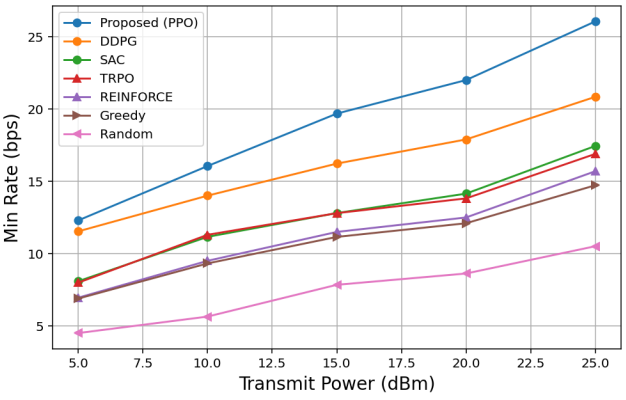

Figure 7: Minimum rate versus transmit power for different algorithms.

CRediT authorship contribution statement

Donghyeon Hur: Conceptualization of this study, Writing-original draft, Writing-editing, Methodology, Investigation. **Donghyun Lee:** Writing-original draft, Writing-review & editing, Validation, Methodology. **Manh Cuong Ho:** Writing-original draft, Validation, Methodology, Investigation, Software. **Wonjong Noh:** Writing-review & editing, Validation, Methodology. **Sungrae Cho:** Writing-review & editing, Validation, Methodology.

References

[1] Y. Zeng, Q. Wu, R. Zhang, Accessing from the sky: A tutorial on UAV communications for 5G and beyond, Proceedings of the IEEE 107 (2019) 2327–2375. doi:10.1109/JPROC.2019.2952892.

[2] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, M. Debbah, A tutorial on UAVs for wireless networks: Applications, challenges, and open problems, IEEE Communications Surveys & Tutorials 21 (2019) 2334–2360. doi:10.1109/COMST.2019.2902862.

[3] Z. Yao, W. Cheng, W. Zhang, H. Zhang, Resource allocation for 5g-UAV-based emergency wireless communications, IEEE Journal on Selected Areas in Communications 39 (2021) 3395–3410. doi:10.1109/JSAC.2021.3088684.

[4] C. Liu, M. Ding, C. Ma, Q. Li, Z. Lin, Y.-C. Liang, Performance analysis for practical unmanned aerial vehicle networks with LoS/NLoS transmissions, in: 2018 IEEE International Conference on Communications Workshops (ICC Workshops), 2018, pp. 1–6. doi:10.1109/ICCW.2018.8403635.

[5] A. A. Nasir, H. D. Tuan, T. Q. Duong, H. V. Poor, UAV-enabled communication using NOMA, IEEE Transactions on Communications 67 (2019) 5126–5138. doi:10.1109/TCOMM.2019.2906622.

[6] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, K. Higuchi, Non-orthogonal multiple access (NOMA) for cellular future radio access, in: 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), 2013, pp. 1–5. doi:10.1109/VTCSpring.2013.6692652.

[7] Y. Mao, B. Clerckx, V. O. Li, Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA, EURASIP Journal on Wireless Communications and Networking 2018 (2018). URL: http://dx.doi.org/10.1186/s13638-018-1104-7. doi:10.1186/s13638-018-1104-7.

[8] B. Clerckx, Y. Mao, R. Schober, H. V. Poor, Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis, 2019. URL: https://arxiv.org/abs/1906.04474. arXiv:1906.04474.

[9] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, G. Caire, Is NOMA efficient in multi-antenna networks? a critical look at next generation multiple access techniques, IEEE Open Journal of the Communications Society 2 (2021) 1310–1343. doi:10.1109/OJCOMS.2021.3084799.

[10] A. Rahmati, Y. Yapici, N. Rupasinghe, I. Guvenc, H. Dai, A. Bhuyan, Energy efficiency of RSMA and NOMA in cellular-connected mmwave UAV networks, in: 2019 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2019, pp. 1–6.

[11] X. Liu, J. Feng, F. Li, V. C. Leung, Downlink energy efficiency maximization for RSMA-UAV assisted communications, IEEE Wireless Communications Letters 13 (2023) 98–102.

[12] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, H. Yanikomeroglu, On the downlink performance of RSMA-based UAV communications, IEEE Transactions on Vehicular Technology 69 (2020) 16258–16263.

[13] B. Yao, R. Li, Y. Chen, L. Wang, Coordinated RSMA for integrated sensing and communication in emergency UAV systems, 2024. URL: https://arxiv.org/abs/2406.19205. arXiv:2406.19205.

[14] T.-H. Nguyen, L. V. Nguyen, L. M. Dang, V. T. Hoang, L. Park, Td3-based optimization framework for RSMA-enhanced UAV-aided downlink communications in remote areas, Remote Sensing 15 (2023) 5284.

[15] O. Dizdar, A. Sattarzadeh, Y. X. Yap, S. Wang, RSMA for overloaded MIMO networks: Low-complexity design for max–min fairness, IEEE Transactions on Wireless Communications 23 (2023) 6156–6173.

[16] B. Lee, W. Shin, Max-min fairness precoder design for rate-splitting multiple access: Impact of imperfect channel knowledge, IEEE Transactions on Vehicular Technology 72 (2022) 1355–1359.

[17] Y. Xu, Y. Mao, O. Dizdar, B. Clerckx, Max-min fairness of rate-splitting multiple access with finite blocklength communications, IEEE Transactions on Vehicular Technology 72 (2022) 6816–6821.

[18] M. Xiao, H. Cui, D. Huang, Z. Zhao, X. Cao, D. O. Wu, Traffic-aware energy-efficient resource allocation for RSMA based UAV communications, IEEE Transactions on Network Science and Engineering 11 (2023) 2537–2548.

[19] J. Feng, X. Liu, Z. Liu, T. S. Durrani, Optimal trajectory and resource allocation for RSMA-UAV assisted IoT communications, IEEE Transactions on Vehicular Technology 73 (2024) 8693–8704.

[20] L. Huang, S. Bi, Y.-J. A. Zhang, Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks, IEEE Transactions on Mobile Computing 19 (2020) 2581–2593. doi:10.1109/TMC.2019.2928811.

[21] Y. Sun, M. Peng, S. Mao, Deep reinforcement learning-based mode selection and resource management for green fog radio access networks, IEEE Internet of Things Journal 6 (2019) 1960–1971. doi:10.1109/JIOT.2018.2871020.

[22] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun, H. Gao, Cooperative UAV trajectory design for disaster area emergency communications: A multiagent PPO method, IEEE Internet of Things Journal 11 (2023) 8848–8859.

[23] G. Zhan, X. Zhang, Z. Li, L. Xu, D. Zhou, Z. Yang, Multiple-UAV reinforcement learning algorithm based on improved ppo in ray framework, Drones 6 (2022) 166.

[24] S. K. Yong, J. S. Thompson, Three-dimensional spatial fading correlation models for compact MIMO receivers, IEEE Transactions on Wireless Communications 4 (2005) 2856–2869.

[25] C. You, R. Zhang, 3D trajectory optimization in rician fading for UAV-enabled data harvesting, IEEE Transactions on Wireless Communications 18 (2019) 3192–3207.

[26] D. Tse, P. Viswanath, Fundamentals of wireless communication, Cambridge university press, 2005.

[27] M. Hata, Empirical formula for propagation loss in land mobile radio services, IEEE transactions on Vehicular Technology 29 (2013) 317–325.

[28] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, IEEE Signal Processing Magazine 34 (2017) 26–38.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[30] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International conference on machine learning, PmLR, 2016, pp. 1928–1937.

[31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, 2019. URL: https://arxiv.org/abs/1509.02971. arXiv:1509.02971.

[32] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 1861–1870. URL: https://proceedings.mlr.press/v80/haarnoja18b.html.

[33] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in: International conference on machine learning, PMLR, 2015, pp. 1889–1897.

[34] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning 8 (1992) 229–256.

[35] S. Liu, An evaluation of DDPG, TD3, SAC, and PPO: deep reinforcement learning algorithms for controlling continuous system, in: 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023), Atlantis Press, 2024, pp. 15–24.

[36] D. Kwon, J. Jeon, S. Park, J. Kim, S. Cho, Multiagent DDPG-based deep learning for smart ocean federated learning IoT networks, IEEE Internet of Things Journal 7 (2020) 9895–9903.

[37] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, A. Madry, Implementation matters in deep rl: A case study on ppo and trpo, in: International Conference on Learning Representations, 2020.

[38] D. W. Matolak, R. Sun, Air–ground channel characterization for unmanned aircraft systems—part III: The suburban and near-urban environments, IEEE Transactions on Vehicular Technology 66 (2017) 6607–6618. doi:10.1109/TVT.2017.2659651.

[39] C. Zhang, P. Patras, H. Haddadi, Deep learning in mobile and wireless networking: A survey, IEEE Communications Surveys Tutorials 21 (2019) 2224–2287. doi:10.1109/COMST.2019.2904897.

[40] M. Chen, U. Challita, W. Saad, C. Yin, M. Debbah, Artificial neural networks-based machine learning for wireless networks: A tutorial, IEEE Communications Surveys Tutorials 21 (2019) 3039–3071. doi:10.1109/COMST.2019.2926625.

[41] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: A selective overview of theories and algorithms, 2021. URL: https://arxiv.org/abs/1911.10635. arXiv:1911.10635.

[42] Y. Huang, J. Su, X. Lu, S. Huang, H. Zhu, H. Zeng, Deep reinforcement learning-based resource allocation for uav-gap downlink cooperative noma in iiot systems, Entropy 27 (2025). URL: https://www.mdpi.com/1099-4300/27/8/811. doi:10.3390/e27080811.

Donghyeon Hur received his B.S. degree from the Department of Advanced Materials Engineering, Kyong-Gi University, Suwon, South Korea, in 2023. He received an M.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2025. His current research interests include deep reinforcement learning, multiple access, wireless networks, and unmanned aerial vehicles.

Donghyun Lee received his B.S. degree in computer science and engineering from Dongguk University, South Korea, in 2020. He received an M.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2022. He is currently pursuing a Ph.D. in computer science and engineering at Chung-Ang University, Seoul, South Korea. His research interests include wireless networks, energy-efficient networks, multimedia communication, and multiple access.

Manh Cuong Ho received two B.S. degrees in information technology from the People's Police University of Technology and Logistics, Vietnam, in 2019, and Hanoi University of Science and Technology, Vietnam, in 2023. He received his M. S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2024. He is currently pursuing a Ph.D. in computer science and engineering at Chung-Ang University, Seoul, South Korea. His research interests include wireless networks, energy-efficient networks, federated learning, deep reinforcement learning, and multiple access.

Wonjong Noh received his B.S., M.S., and Ph.D. degrees from the Department of Electronics Engineering, Korea University, Seoul, Korea, in 1998, 2000, and 2005, respectively. From 2005 to 2007, he conducted postdoctoral research at Purdue University, IN, USA, and the University of California at Irvine, CA, USA. From 2008 to 2014, he was a Principal Research Engineer with Samsung Advanced Institute of Technology, Samsung Electronics, Korea. From 2014 to 2018, he was an Assistant Professor with the Department of Electronics and Communication Engineering, Gyeonggi University of Science and Technology, Korea. He is currently an Associate Professor at the School of Software, Hallym University, Korea. His current research interests include fundamental capacity analysis in 5G/6G wireless communication and networks, federated mobile edge computing, and machine learning-based system control. He received a government postdoctoral Fellowship from the Ministry of Information and Communications of Korea in 2005. He also received the Samsung Best Paper Gold Award in 2010, the Samsung Patent Bronze Award in 2011, and the Samsung Technology Award in 2013.

Sungrae Cho is a Professor with the School of Computer Sciences and Engineering, Chung-Ang University (CAU), Seoul, Korea. Before joining CAU, he was an Assistant Professor with the Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA, from 2003 to 2006, and a senior member of the technical staff at the Samsung Advanced Institute of Technology, Kiheung, South Korea, in 2003. From 1994 to 1996, he was a research staff member at the Electronics and Telecommunications Research Institute in Daejeon, South Korea. From 2012 to 2013, he was a Visiting Professor with the National Institute of Standards and Technology, Gaithersburg, MD, USA. He received his B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, South Korea, in 1992 and 1994, respectively, and his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2002. His current research interests include wireless networking, ubiquitous computing, and information and ICT convergence. He has served as a subject editor with IET Electronics Letters since 2018 and held the position of Editor for Ad Hoc Networks Journal (Elsevier) from 2012 to 2017. He has served as the organizing committee chair for numerous international conferences, including IEEE SECON, ICOIN, ICTC, ICUFN, TridentCom, and IEEE MASS, and as a program committee member for many conferences including IEEE ICC, MobiApps, SENSORNETS, and WINSYS.