Deep Reinforcement Learning for Trajectory and Phase Shift Optimization of Aerial RIS in CoMP-NOMA Networks

Muhammad Umer*, Muhammad Ahmed Mohsin*, Aamir Mahmood†, Kapal Dev‡, Haejoon Jung§, Mikael Gidlund†, and Syed Ali Hassan*
*School of Electrical Engineering and Computer Science (SEECS), NUST, Pakistan
†Department of Computer and Electrical Engineering, Mid Sweden University, Sweden
‡Department of Computer Science, Munster Technological University, Ireland
§Department of Electronics and Information Convergence Engineering, Kyung Hee University, Republic of Korea
Email: {mumer.bee20seecs, mmohsin.bee20seecs, ali.hassan}@seecs.edu.pk,
{aamir.mahmood, mikael.gidlund}@miun.se, kapal.dev@ieee.org, haejoonjung@khu.ac.kr

***Abstract*—This paper explores the potential of aerial reconfigurable intelligent surfaces (ARIS) to enhance coordinated multi-point non-orthogonal multiple access (CoMP-NOMA) networks. We consider a system model where a UAV-mounted RIS assists in serving multiple users through NOMA, while coordinating with multiple base stations. The optimization of UAV trajectory, RIS phase shifts, and NOMA power control constitutes a complex problem due to the hybrid nature of the parameters, involving both continuous and discrete values. To tackle this challenge, we propose a novel framework utilizing the multi-output proximal policy optimization (MO-PPO) algorithm. MO-PPO effectively handles the diverse nature of these optimization parameters and through extensive simulations, we demonstrate its effectiveness in achieving near-optimal performance and adapting to dynamic environments. Our findings highlight the benefits of integrating ARIS in CoMP-NOMA networks for improved spectral efficiency and coverage in future wireless networks.**

***Index Terms*—Deep reinforcement learning, unmanned aerial vehicle, RIS, NOMA, CoMP, trajectory design.**

I. Introduction

Driven by the ever-growing demand for ubiquitous connectivity and high data rates, future wireless networks necessitate the exploration of novel solutions that surpass the limitations of traditional approaches. Unmanned aerial vehicles (UAVs), with their inherent mobility and flexible deployment as aerial base stations, have emerged as a promising technology to address these challenges. This enables them to provide wireless service in diverse scenarios, ranging from temporary hotspots during events and disaster-stricken areas with compromised infrastructure to remote locations with limited coverage [1], [2]. However, UAV-assisted networks face challenges such as limited energy capacity and constrained coverage area, hindering the full realization of their potential benefits.

To overcome these limitations and unlock the full potential of UAV-assisted networks, researchers are actively investigating the integration of enabling technologies such as reconfigurable intelligent surfaces (RIS) [3]. RIS, composed of numerous passive reflecting elements that can be dynamically adjusted, offer the ability to control electromagnetic wave propagation. By intelligently manipulating the phase shifts of the incident signals, RIS can enhance the desired signal strength, suppress interference, and extend coverage area. In the context of UAV-assisted networks, mounting RIS on UAVs creates aerial RIS (ARIS) networks, offering greater flexibility in optimizing the wireless environment through dynamic adaptation of RIS location and orientation [4]. This dynamic adaptability enables ARIS to proactively respond to changing channel conditions and user distribution, fostering efficient and robust communication links.

Furthermore, non-orthogonal multiple access (NOMA) and coordinated multi-point (CoMP) transmission offer complementary benefits for improving spectral efficiency and user fairness in wireless networks. NOMA enables multiple users to share the same time-frequency resources, enhancing spectrum utilization and performance metrics such as outage probability and spectral efficiency [5]. CoMP, on the other hand, enables cooperation between multiple base stations to jointly serve users, mitigating inter-cell interference and enhancing user experience. The integration of CoMP and NOMA (CoMP-NOMA) further amplifies these benefits by allowing multiple BSs to collaboratively serve NOMA users with coordinated power allocation and SIC decoding [6], [7].

The convergence of RIS, CoMP, and NOMA within UAV-assisted networks holds immense potential for enhancing the performance and efficiency of future wireless communication systems. While recent research has demonstrated the benefits of combining these technologies with UAVs, existing works often assume static RIS deployments, limiting network adaptability [8], [9]. Although some studies have investigated ARIS-assisted CoMP-NOMA networks and optimized UAV trajectory and RIS phase shifts for sum rate maximization [10], the employed optimization approaches, such as double-layer alternating optimization, may face scalability and convergence challenges in large-scale networks.

To address the challenges of optimizing ARIS-assisted CoMP-NOMA networks, we propose a deep reinforcement

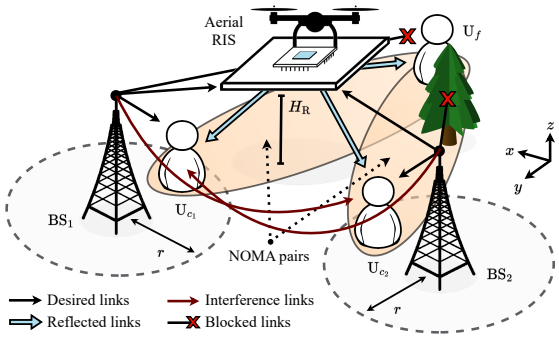

Fig. 1. Aerial RIS-assisted coordinated NOMA cluster.

learning (DRL) approach based on the multi-output proximal policy optimization (MO-PPO) algorithm to effectively address the hybrid continuous-discrete action space inherent in these networks. Our framework jointly optimizes UAV trajectory, RIS phase shifts, and NOMA power control to maximize network sum rate while adhering to user quality of service (QoS) constraints. Through extensive simulations, we evaluate the efficacy of our proposed approach, assess the convergence of MO-PPO, and highlight the benefits of CoMP-NOMA and RIS in UAV-assisted networks.

II. System Model & Problem Formulation

A. System Description

We consider a downlink transmission scenario in a multi-cell CoMP-NOMA network assisted by a UAV-mounted RIS, as illustrated in Figure 1. The network consists of I cells, each modeled as a circular disk of radius r with a single-antenna BS at its center, denoted as BS_i, where $i \in \mathcal{I} \triangleq \{1, 2, \ldots, I\}$. Each BS_i invokes two-user downlink NOMA to serve its respective cell-center and edge user, each also equipped with a single-antenna. The cell-center users are defined as users that lie within the disk of their associated cell and are denoted as U_{c_i}, $\forall i$ and $c_i \in \mathcal{C}^i \triangleq \{1, 2, \ldots, C_i\}$, where C_i is the number of cell-center users in cell i. Conversely, the edge users are defined as users that do not lie within any cell and are denoted as U_f, $\forall f$ and $f \in \mathcal{F} \triangleq \{1, 2, \ldots, F\}$, where F is the number of edge users in the network. Furthermore, let $\mathcal{U} \triangleq \bigcup_{i \in \mathcal{I}} \mathcal{C}^i \cup \mathcal{F}$ be the set of all the users in the network. Without loss of generality and for ease of exposition, we assume $I = 2$, and $C_i = F = 1$, $\forall i$.

For coordinated operation, the BSs are assumed to be interconnected via a high-speed backhaul network to a central processing unit (CPU). Moreover, to improve the signal quality for edge users, an ARIS, denoted as R, is deployed at a fixed altitude H_R over area A to create reflection links between the BSs and the users, and is equipped with K passive elements. For tractability, we discretize the entire system operation into time slots of equal length τ, where each time slot is indexed by $t \in \mathcal{T} \triangleq \{1, 2, \ldots, T\}$, such that T is the total flight time of the UAV. Furthermore, we assume the presence of O obstacles in the network, denoted as $\mathcal{O} \triangleq \{1, 2, \ldots, O\}$, where each obstacle O_o, $o \in \mathcal{O}$ has its own *forbidden zone* represented as a circular disk of radius d_min, centered at the obstacle's location, where the UAV is not allowed to fly due to safety and regulatory constraints.

Before proceeding with the channel and signal model, we define the positions of the various entities in the network. Specifically, $\forall i \in \mathcal{I}$, $u \in \mathcal{U}$, and $o \in \mathcal{O}$, the positions of BS_i, U_u, and O_o are represented by $\mathbf{p}_i = (x_i, y_i, H_\text{B})$, $\mathbf{p}_u = (x_u, y_u, 0)$, and $\mathbf{p}_o = (x_o, y_o, H_\text{O})$, respectively, where H_B and H_O are the heights of the BSs and obstacles, respectively. Moreover, the position of R at time slot t is denoted as $\mathbf{p}_\text{R}[t] = (x_\text{R}[t], y_\text{R}[t], H_\text{R})$. In this paper, we assume that the users are stationary, and the UAV is capable of adjusting its horizontal position in the xy-plane, while maintaining a fixed altitude.

B. Channel Model & RIS Configuration

Our analysis considers both large-scale path loss and small-scale fading effects on signal propagation. Similar to [8], we assume a rich scattering environment, leading to the modeling of direct links between BS_i and U_u as Rayleigh fading channels, denoted as $h_{i,u}$. Mathematically, the channel $h_{i,u}$ at time slot t is given by

$$h_{i,u}[t] = \sqrt{\frac{\rho_o}{PL(d_{i,u})}} v_{i,u}[t], \tag{1}$$

where ρ_o is the reference path loss at 1 m, $PL(d_{i,u}) = (d_{i,u})^{-\alpha_{i,u}}$ is the large-scale path loss, such that $\alpha_{i,u}$ is the path loss exponent, $d_{i,u} = \|\mathbf{p}_i - \mathbf{p}_u\|$ is the distance between BS_i and U_u and $\|\cdot\|$ denotes the Euclidean norm. Moreover, $v_{i,u}[t] \in \mathbb{C}^{1\times 1}$ is the small-scale Rayleigh fading coefficient with zero mean and unit variance, and is assumed to be independent and identically distributed (i.i.d) across different time slots and users. In this work, as a special case, we assume that the direct link between BS_i and U_f is blocked due to the presence of obstacles, thus $h_{i,f}[t] = 0$, $\forall i, f$.

Contrary to the direct links, the reflection links between BS_i and R are modeled as Rician fading channels, denoted as $\mathbf{h}_{i,\text{R}}[t]$, due to the presence of a dominant line-of-sight (LoS) component. At time slot t, the channel $\mathbf{h}_{i,\text{R}}[t]$ is given by

$$\mathbf{h}_{i,\text{R}}[t] = \sqrt{\frac{\rho_o}{PL(d_{i,\text{R}}[t])}} \left(\sqrt{\frac{\kappa}{1+\kappa}} \mathbf{g}_{i,\text{R}}^{\text{LoS}}[t] + \sqrt{\frac{1}{1+\kappa}} \mathbf{g}_{i,\text{R}}^{\text{NLoS}}[t] \right), \tag{2}$$

where κ is the Rician factor, and $d_{i,\text{R}}[t] = \left\|\mathbf{p}_i - \mathbf{p}_\text{R}[t]\right\|$ is the distance between BS_i and R. Moreover, the deterministic LoS represented, i.e., $\mathbf{g}_{i,\text{R}}^{\text{LoS}}[t] \in \mathbb{C}^{K\times 1}$, is given by

$$\mathbf{g}_{i,\text{R}}^{\text{LoS}} = \left[1, \ldots, e^{j(k-1)\pi \sin(\omega_i)}, \ldots, e^{j(K-1)\pi \sin(\omega_i)}\right]^T,$$

where $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$ indexes elements of R and ω_i is the angle of arrival (AoA) whereas $\mathbf{g}_{i,\text{R}_i}^{\text{NLoS}} \in \mathbb{C}^{K\times 1}$ is the NLoS component following Rayleigh fading as previously described. Similarly, the channel between R and U_u, denoted as $h_{\text{R},u}$, can also be modeled as a Rician fading channel.

For the ARIS configuration, we assume that the phase shift of the k-th element can be set independently of other elements

and that both the UAV trajectory and the phase response are controlled by the CPU. Furthermore, the phase shift (PS) matrix at time slot t is expressed as

$$\mathbf{\Theta}[t] = \text{diag}\left(a_1 e^{j\theta_1[t]}, a_2 e^{j\theta_2[t]}, \ldots, a_k e^{j\theta_K[t]}\right), \quad (3)$$

where $a_k \in (0, 1]$ is the amplitude coefficient and $\theta_k[t] \in [-\pi, \pi)$ is the phase shift of the k-th element. For simplicity, we assume an ideal RIS with perfect phase control and unit amplitude reflection coefficients for all elements, i.e., $a_k = 1, \forall k$. Additionally, perfect channel state information (CSI) is assumed to be available at the CPU.

C. Signal Model

In accordance with the NOMA principle, each BS_i serves two users, U_{c_i} and U_f, simultaneously, by superimposing their signals. The transmitted signal from BS_i at time slot t is given by $x_i[t] = \sqrt{(1-\lambda_i)P_i}x_{i,c_i}[t] + \sqrt{\lambda_i P_i}x_{i,f}[t]$, where $x_{i,c_i}[t]$ and $x_{i,f}[t]$ represent the respective signals for U_{c_i} and U_f, P_i is the transmit power of BS_i and λ_i is the power allocation factor assigned to U_f. To ensure successful decoding at the U_{c_i}, we constrain $\lambda_i \in (0.5, 1)$, as deduced in [11], [12].

The signal received by U_f can be expressed as

$$y_f[t] = H_{i,f}[t]x_i[t] + H_{i',f}[t]x_{i'}[t] + n_o[t] \quad (4)$$

where $i' \in \mathcal{I} \setminus \{i\}$, $n_o[t] \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN), and $H_{i,f}[t] = \mathbf{h}_{\text{R},f}^T[t]\mathbf{\Theta}[t]\mathbf{h}_{i,\text{R}}[t]$ represents the effective channels between BS_i and U_f through R, respectively. To minimize synchronization overhead, we employ non-coherent JT-CoMP, thus, the signal-to-interference-plus-noise ratio (SINR) is given by

$$\gamma_f[t] = \frac{\lambda_i \left|H_{i,f}[t]\right|^2 + \lambda_{i'}\left|H_{i',f}[t]\right|^2}{(1-\lambda_i)\left|H_{i,f}[t]\right|^2 + (1-\lambda_{i'})\left|H_{i',f}[t]\right|^2 + \frac{1}{\rho}}, \quad (5)$$

where $\rho = P_t/\sigma^2$ is the transmit SNR and $P_t = P_i, \forall i$ is the transmit power of each BS.

On the other hand, the signal received by U_{c_i} can be expressed as

$$y_{c_i}[t] = H_{i,c_i}[t]x_i[t] + h_{i',c_i}[t]x_{i'}[t] + n_o[t], \quad (6)$$

where $H_{i,c_i}[t] = h_{i,c_i}[t] + \mathbf{h}_{\text{R},c_i}^T[t]\mathbf{\Theta}[t]\mathbf{h}_{i,\text{R}}[t]$ represents the effective channels between BS_i and U_{c_i} through R, respectively. Also, the term $h_{i',c_i}[t]x_{i'}[t]$ represents the ICI caused by the transmission of $\text{BS}_{i'}$ at U_{c_i}. Based on the SIC principle, U_{c_i} first decodes $x_{i,f}[t]$ and then cancels it from $y_{c_i}[t]$ to decode $x_{i,c_i}[t]$. The SINR at U_{c_i} for decoding $x_{i,f}[t]$ is given by

$$\gamma_{c_i \to f}[t] = \frac{\lambda_i \left|H_{i,c_i}[t]\right|^2}{(1-\lambda_i)\left|H_{i,c_i}[t]\right|^2 + \left|h_{i',c_i}[t]\right|^2 + \frac{1}{\rho}}, \quad (7)$$

whereas the SINR at U_{c_i} for decoding $x_{i,c_i}[t]$ is

$$\gamma_{c_i}[t] = \frac{(1-\lambda_i)\left|H_{i,c_i}[t]\right|^2}{\left|h_{i',c_i}[t]\right|^2 + \frac{1}{\rho}}. \quad (8)$$

Finally, the achievable sum rate of the network at time slot t can be expressed as

$$R_{\text{sum}}[t] = \sum_{i\in\mathcal{I}} R_{c_i}[t] + \sum_{f\in\mathcal{F}} R_f[t]. \quad (9)$$

where $R_{c_i}[t] = \log_2(1+\gamma_{c_i}[t])$ and $R_f[t] = \log_2(1+\gamma_f[t])$ are the achievable rates of U_{c_i} and U_f, respectively.

D. Problem Formulation

The primary objective of this work is to maximize the cumulative sum rate achieved over a period of T time slots. To accomplish this, we jointly optimize three key control variables: the UAV trajectory denoted as $\mathbf{P} \triangleq \{\mathbf{p}_\text{R}[t], \forall t\}$, the RIS phase shifts represented by $\mathbf{\Theta} \triangleq \{\mathbf{\Theta}[t], \forall t\}$, and the power allocation factors denoted as $\mathbf{\Lambda} \triangleq \{\lambda_i, \forall i\}$. Mathematically, the optimization problem is formulated as

$$\begin{aligned}
\max_{\mathbf{P},\mathbf{\Theta},\mathbf{\Lambda}} \quad & \sum_{t\in\mathcal{T}} R_{\text{sum}}[t] && \text{(10a)}\\
\text{s.t.} \quad & x_\text{R}[t], y_\text{R}[t] \in A, \ \forall t \in \mathcal{T}, && \text{(10b)}\\
& \left\|\mathbf{p}_\text{R}[t] - \mathbf{p}_o\right\| \geq d_{\min}, \ \forall o \in \mathcal{O}, \ t \in \mathcal{T}, && \text{(10c)}\\
& \theta_k[t] \in [-\pi, \pi), \ \forall k \in \mathcal{K}, \ t \in \mathcal{T}, && \text{(10d)}\\
& R_{c_i}[t] \geq R_{c_i}^{\min}, \ \forall i \in \mathcal{I}, \ t \in \mathcal{T}, && \text{(10e)}\\
& R_f[t] \geq R_f^{\min}, \ \forall f \in \mathcal{F}, \ t \in \mathcal{T}, && \text{(10f)}\\
& \lambda_i \in (0.5, 1), \ \forall i \in \mathcal{I}, && \text{(10g)}
\end{aligned}$$

where constraint (10b) restricts the UAV trajectory to lie within A, and constraint (10c) enforces a minimum safety distance between the UAV and any obstacles present, thus guaranteeing the UAV's safety. Constraint (10d) limits the phase shifts applied by the RIS elements. To meet the quality of service (QoS) requirements, constraints (10e) and (10f) impose minimum rate thresholds, denoted by $\mathcal{R}_{c_i}^{\min}$ and $\mathcal{R}_f^{\min}$, for U_{c_i} and U_f, respectively. Lastly, constraint (10g) defines the permissible range for power allocation factors, ensuring successful SIC. The optimization problem in (10) is non-convex due to the coupled variables $\{\mathbf{P}, \mathbf{\Theta}, \mathbf{\Lambda}\}$. To address this, we propose a DRL-based solution in the next section.

III. Deep Reinforcement Learning-based Proposed Solution

A. MDP Formulation

To enable the applicability of DRL, we first recast the problem in (10) as a single-agent Markov Decision Process (MDP) operating in discrete time steps. The MDP is represented by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ denotes the set of possible environment states, $\mathcal{A}$ represents the action space, $\mathcal{P}$ defines the state transition probabilities, $\mathcal{R}$ is the reward function guiding the agent's learning, and γ is the discount factor that determines the importance of future rewards. At each time slot t, the agent observes the current state s_t, selects an action a_t based on its policy, transitions to a new state s_{t+1}, and receives a reward $\mathcal{R}(s_t, a_t)$. We define $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ as follows.

1) *State Space $\mathcal{S}$:* The environment state at time slot t consists of the UAV's current position $\mathbf{p}_\text{R}[t]$, its distances to

the center of each obstacle $\mathbf{d}_{\mathrm{R}}[t] = \{\left\|\mathbf{p}_{\mathrm{R}}[t] - \mathbf{p}_o\right\|, \forall o \in \mathcal{O}\}$, the power allocation factors $\mathbf{\Lambda}$, and the achievable rates $\mathbf{R}[t] = \{R_{c_i}[t], R_f[t], \forall i, f\}$. Formally, the state space can be expressed as

$$s_t = \{\mathbf{p}_{\mathrm{R}}[t], \mathbf{d}_{\mathrm{R}}[t], \mathbf{\Lambda}, \mathbf{R}[t]\} \in \mathbb{R}^{\dim_{\mathcal{S}}}. \tag{11}$$

where $\dim_{\mathcal{S}} = 2 + O + I + \sum_{i \in \mathcal{I}} C_i + F$ is the dimension of the state space.

2) *Action Space $\mathcal{A}$:* The actions available to the agent consist of controlling the movement of the UAV in the xy-plane, adjusting the phase shifts of the RIS elements, and managing the power allocation factors. Specifically, the action space at time slot t contains the manuevering actions $a_{\mathrm{R}}[t] \in \{(-1,0),(1,0),(0,-1),(0,1),(0,0)\}$, representing left, right, down, up, and hover, respectively, the phase shifts $a_\phi[t] = \{\phi_k[t], \forall k\}$, and the power allocation factors $a_\lambda = \{\lambda_i, \forall i\}$. Thus, the action space can be represented as

$$a_t = \{a_{\mathrm{R}}[t], a_\phi[t], a_\lambda\} \in \mathbb{R}^{\dim_{\mathcal{A}}}. \tag{12}$$

where $\dim_{\mathcal{A}} = 2 + K + I$ is the dimension of the action space.

3) *Reward Function $\mathcal{R}$:* The reward function plays a crucial role in shaping the learning behavior of the RL agent. Our design prioritizes maximizing the sum rate while simultaneously ensuring the UAV's safety and adherence to QoS requirements. The reward function is defined as

$$\mathcal{R}(s_t, a_t) = R_{\mathrm{sum}}[t]\left(1 - \frac{\sum_{u \in \mathcal{U}} \zeta_u[t]}{|\mathcal{U}|}\right) - \xi_{\mathrm{R}}[t] K_{\mathrm{viol}}, \tag{13}$$

where K_{viol} is the penalty factor for UAV's safety constraint violation, and $\zeta_u[t] = \mathbb{I}\{R_u[t] \leq R_u^{\min}\}$ is the indicator function for the QoS constraints, i.e., $\zeta_u[t] = 1$ if QoS constraints are violated, and 0 otherwise. Similarly, $\xi_{\mathrm{R}}[t] = \mathbb{I}\{x_{\mathrm{R}}[t], y_{\mathrm{R}}[t] \notin A \wedge \left\|\mathbf{p}_{\mathrm{R}}[t] - \mathbf{p}_o\right\| < d_{\min}, \forall o \in \mathcal{O}\}$ is the indicator function for UAV's safety constraints.

B. MO-PPO Algorithm

The considered action space is a hybrid continous-discrete space, which poses a challenge for traditional RL algorithms. While discretizing the continuous actions is a possible solution, it can lead to a large action space, significantly increasing computational complexity and potentially hindering the performance. To address this challenge, we propose employing a multi-output Proximal Policy Optimization (MO-PPO) algorithm. MO-PPO extends the standard PPO [13] framework by employing two parallel actor networks, each responsible for generating the discrete action a_{R} and the continuous actions a_ϕ and a_λ, respectively. The actor networks share the first few layers, allowing for the extraction of common features and encoding the state information. Furthermore, a single critic network is employed to estimate the value function $V(s_t)$, which is used to compute a variance-reduced advantage function estimate $\hat{A}_t$ for policy optimization. Following the implementation details used in [14], the policy is executed for $\hat{T}$ time steps, and $\hat{A}_t$ is computed as

$$\hat{A}_t = \sum_{k=0}^{\hat{T}-1} \gamma^k r_{t+k} + \gamma^{\hat{T}} V(s_{t+\hat{T}}) - V(s_t), \tag{14}$$

where $\hat{T}$ is much smaller than the length of the episode T.

To generate the stochastic policy $\pi_{\theta_d}(a_t|s_t)$ for the discrete actions, the corresponding actor network outputs $|a_{\mathrm{R}}|$ logits, which are then passed through a $\mathrm{softmax}$ function to obtain a probability distribution over the available discrete actions. Conversely, the continuous actor network generates the continuous actions a_ϕ and a_λ by sampling from Gaussian distributions parameterized by the mean and standard deviation outputs of the network, as dictated by the stochastic policy $\pi_{\theta_c}(a_t|s_t)$. Both $\pi_{\theta_d}(a_t|s_t)$ and $\pi_{\theta_c}(a_t|s_t)$ are optimized independently using their respective clipped surrogate objective functions. For the discrete actions, the objective function is given by

$$L_d^{\mathrm{CLIP}}(\theta_d) = \hat{\mathbb{E}}_t\left[\min(r_t^d(\theta_d)\hat{A}_t, \Im(r_t^d, \theta_d, \epsilon)\hat{A}_t\right], \tag{15}$$

where $\Im(r_t^d, \theta_d, \epsilon) = \mathrm{clip}(r_t^d(\theta_d), 1-\epsilon, 1+\epsilon)$, $r_t^d(\theta_d) = \pi_{\theta_d}(a_t|s_t)/\pi_{\theta_d}^{\mathrm{old}}(a_t|s_t)$ is the importance sampling ratio, and ϵ is the clipping parameter. The objective function for the continuous actions can be expressed in a similar manner but is left out for brevity.

Algorithm 1: MO-PPO Algorithm

```
Initialize the policy parameters θ_d and θ_c
for episode = 1, 2, ..., N do
    Receive initial state s_0
    for time step t = 0, 1, ..., T do
        Generate discrete action a_R using π_{θ_d}(a_t|s_t)
        Generate continuous actions a_φ and a_λ using
          π_{θ_c}(a_t|s_t)
        Execute actions a_t = {a_R, a_φ, a_λ}
        if UAV violates (10b) or (10c) then
            Set ξ_R[t] = 1, cancel the UAV's movement,
              and update the state s_{t+1}
        end
        Observe reward R as (13) and next state s_{t+1}
        Collect a set of partial trajectories D with T̂
          transitions
        Compute the variance-reduced advantage
          estimate Â_t as (14)
    end
    for epoch = 1, 2, ..., E do
        Sample a mini-batch of transitions B from D
        Compute the clipped surrogate objectives
          L_d^CLIP(θ_d) and L_c^CLIP(θ_c) as (15)
        Optimize overall objective and update the
          policy parameters θ_d and θ_c using Adam [15]
    end
    Synchronize the sampling policies as
                θ_d^old ← θ_d and θ_c^old ← θ_c
    Clear the collected trajectories D
end
```

It is worth noting that while both policies collaborate within the environment, their optimization objectives remain decoupled, i.e., $\pi_{\theta_d}(a_t|s_t)$ and $\pi_{\theta_c}(a_t|s_t)$ are treated as independent

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
Reference path loss ρ_o	-30 dBm	Rician factor κ	3 dB
Target data rate $R_f^{\min}$	0.2 bps/Hz	Learning rate	$2.75\text{e}-4$
Target data rate $R_{c_i}^{\min}$	0.5 bps/Hz	Clipping parameter ϵ	0.1
Penalty constant K_{viol}	7	Discount factor γ	0.98
Minimum distance $d_{\min}$	10 m	Number of episodes N	750
Time slots per episode T	250	Number of epochs E	20
Number of neurons	64	Batch size B	128

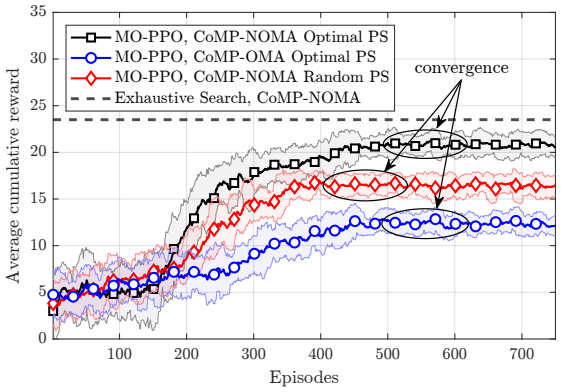

Fig. 2. Average cumulative reward vs. number of training episodes with $P_t = 20$ dBm and $K = 120$ elements.

distributions during policy optimization, rather than a joint distribution encompassing both action spaces. The MO-PPO algorithm is summarized in Algorithm 1.

IV. NUMERICAL RESULTS

A. Simulation Setup

To evaluate the efficacy of the proposed MO-PPO algorithm, we construct a simulated urban environment spanning an area of 150×150 m^2 with $I = 2$ BSs, $U = 3$ users, and $O = 2$ obstacles. The initial position of the UAV is set to $(0, 35, 50)$ m, while BS_1 and BS_2 are located at $(-35, -35, 25)$ m and $(35, 35, 25)$ m, respectively. All remaining entities are randomly placed within the environment.

Both BSs are assumed to transmit at an identical power level, i.e., $P_1 = P_2 = P_t$. Furthermore, the network operates at a carrier frequency of $f_c = 2.4$ GHz, utilizing a bandwidth of $BW = 10$ MHz and the noise power is set to $\sigma^2 = -174 + 10\log_{10}(BW)$ dBm. To model the signal propagation characteristics, we employ path loss exponents of $\alpha_{i,u} = 3$, $\alpha_{i,\text{R}} = \alpha_{\text{R},u} = 2.2$, and $\alpha_{i',u} = 3.5$, for direct, reflection, and interference links, respectively. Table I summarizes the remaining simulation parameters.

B. Results

Fig. 2 illustrates the average cumulative reward achieved by the MO-PPO algorithm with different network configurations. As shown, the algorithm consistently converges to a stable reward value after approximately 500 episodes, indicating the successful acquisition of an effective policy. Notably, MO-PPO

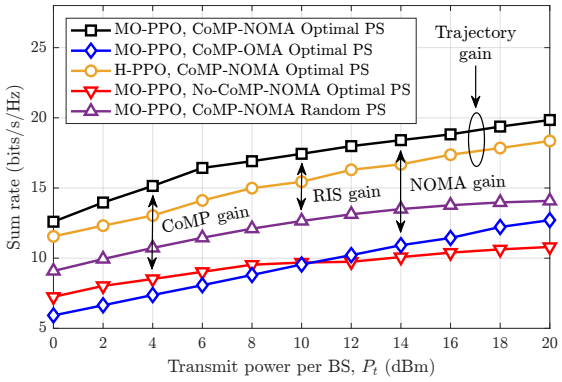

Fig. 3. Sum rate vs. transmit power for different algorithms and configurations with $K = 120$ elements.

with random phase shifts (PS) exhibits a faster convergence rate compared to its counterpart with optimal PS. This observation can be attributed to the increased complexity associated with optimizing the PS within the action space, leading to a slower convergence process. Moreover, the CoMP-NOMA configuration achieves a superior average cumulative reward compared to the CoMP-OMA configuration, underscoring the benefits of NOMA in enhancing overall network performance. A comparison with the optimal solution obtained through exhaustive search reveals that the proposed MO-PPO algorithm achieves near-optimal performance, effectively demonstrating its capability to solve the formulated problem.

Next, we investigate the sum rate achieved by the network as a function of the transmit power P_t as shown in Fig. 3. As expected, the sum rate exhibits an upward trend with increasing transmit power, emphasizing the crucial role of power control in optimizing network performance. The results clearly showcase the advantages of incorporating CoMP, RIS, and NOMA techniques to enhance spectral efficiency across all power levels. Furthermore, we compare the proposed MO-PPO algorithm against the hover PPO (H-PPO) algorithm, which maintains a fixed UAV position. This comparison highlights the significant improvement in network performance achieved by MO-PPO, directly attributable to its dynamic trajectory optimization capabilities. By adapting the UAV's position, MO-PPO effectively exploits favorable channel conditions, outperforming the static H-PPO approach. This underscores the crucial role of trajectory optimization in maximizing the potential of ARIS-assisted CoMP-NOMA networks.

Our investigation extends to analyzing the impact of the number of RIS elements on the network's achievable sum rate, providing further insights into the performance of the MO-PPO algorithm Fig. 4 illustrates the positive correlation between the achievable sum rate and the number of RIS elements, emphasizing the advantages of utilizing a larger RIS to enhance network performance. However, we observe a subtle, yet noteworthy trend: the difference in sum rate between the exhaustive search baseline and the MO-PPO algorithm, while remaining small, increases with the number of RIS elements. This observation highlights the importance

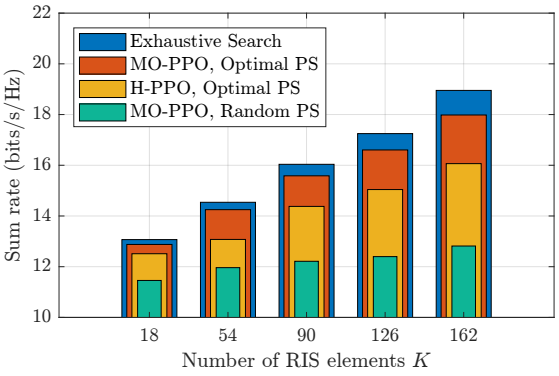

Fig. 4. Impact of the number of RIS elements on the achievable sum rate with $P_t = 10$ dBm.

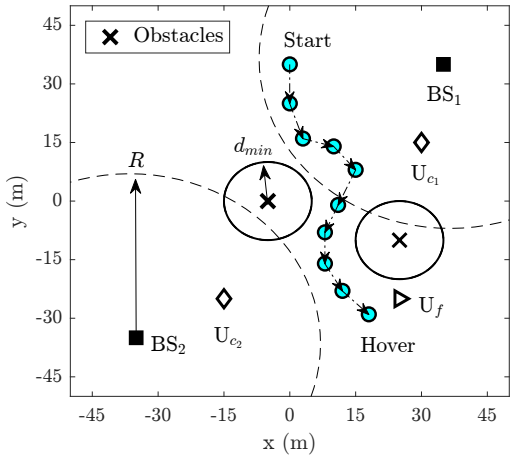

Fig. 5. Top view of the UAV trajectory obtained by the MO-PPO algorithm sampled every 25 time slots and averaged over 10 evaluation episodes

of carefully considering the trade-off between performance and complexity when determining the optimal number of operational RIS elements, a challenge we aim to address in future work.

Finally, we visualize the UAV trajectory generated by the MO-PPO algorithm in Fig. 5. It is observed that the UAV adopts a *cautious approach*, navigating around obstacles while minimizing its distance to U$_f$. Such a trend is commonly observed in DRL algorithms that operate on the principle of exploration-exploitation. The agent learns to strike a balance between exploring the environment and exploiting its current knowledge to maximize the cumulative reward. The generated trajectory further highlights the agent's ability to adapt to the dynamic environment and optimize network performance by effectively leveraging both RIS and NOMA techniques.

V. Conclusion

This paper explored the potential of ARIS in enhancing CoMP-NOMA networks. We proposed a novel framework utilizing the MO-PPO algorithm to jointly optimize UAV trajectory, RIS phase shifts, and NOMA power control, aiming to maximize network sum rate while satisfying user QoS constraints. Our results demonstrated the effectiveness of MO-PPO in handling the hybrid action space and achieving near-optimal performance, with significant gains observed in sum rate compared to benchmark schemes. This highlights the advantages of integrating ARIS, CoMP, and NOMA for future wireless networks, paving the way for more efficient and adaptable communication systems.

Although our analysis focused on a two-cell network, the proposed framework can be readily extended to accommodate a larger number of cells. Future research directions include exploring more sophisticated DRL algorithms with improved sample efficiency and convergence speed, as well as investigating the impact of imperfect CSI on the optimization process. Moreover, delving into the integration of ARIS with other emerging technologies, such as millimeter-wave and terahertz communication, presents exciting future research avenues.

References

[1] Z. Mohamed and S. Aissa, "Leveraging UAVs with intelligent reflecting surfaces for energy-efficient communications with cell-edge users," in *2020 IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, pp. 1–6, IEEE, 2020.

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

[3] N. Gao, S. Jin, X. Li, and M. Matthaiou, "Aerial RIS-assisted high altitude platform communications," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 10, pp. 2096–2100, 2021.

[4] T. N. Do, G. Kaddoum, T. L. Nguyen, D. B. Da Costa, and Z. J. Haas, "Aerial reconfigurable intelligent surface-aided wireless communication systems," in *2021 IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, pp. 525–530, IEEE, 2021.

[5] X. Yue, Z. Qin, Y. Liu, S. Kang, and Y. Chen, "A unified framework for non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5346–5359, 2018.

[6] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, 2018.

[7] M. Elhattab, M.-A. Arfaoui, and C. Assi, "CoMP transmission in downlink NOMA-based heterogeneous cloud radio access networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7779–7794, 2020.

[8] J. Zhao, L. Yu, K. Cai, Y. Zhu, and Z. Han, "RIS-aided ground-aerial NOMA communications: A distributionally robust DRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1287–1301, 2022.

[9] I. Budhiraja, V. Vishnoi, N. Kumar, D. Garg, and S. Tyagi, "Energy-efficient optimization scheme for RIS-assisted communication underlaying UAV with NOMA," in *ICC 2022 - IEEE Int. Conf. Commun.*, pp. 1–6, IEEE, 2022.

[10] S. Lv, X. Xu, S. Han, and P. Zhang, "UAV-RIS assisted coordinated multipoint finite blocklength transmission for MTC networks," *IEEE Internet Things J.*, 2023.

[11] M. Elhattab, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "RIS-assisted joint transmission in a two-cell downlink NOMA cellular system," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1270–1286, 2022.

[12] M. Obeed, H. Dahrouj, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "User pairing, link selection, and power allocation for cooperative NOMA hybrid VLC/RF systems," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 3, pp. 1785–1800, 2020.

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

[14] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Int. Conf. Mach. Learn.*, pp. 1928–1937, PMLR, 2016.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.