# Deep Reinforcement Learning for Trajectory and Beamforming Optimization of Aerial RIS in CoMP-NOMA Networks

Muhammad Umer*, Muhammad Ahmed Mohsin*, Syed Ali Hassan*, and Haejoon Jung†
*School of Electrical Engineering and Computer Science (SEECS), NUST, Pakistan
†Department of Electronics and Information Convergence Engineering, Kyung Hee University, Republic of Korea
Email: {mumer.bee20seecs, mmohsin.bee20seecs, ali.hassan}@seecs.edu.pk, 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 propose a novel framework utilizing the multi-output proximal policy optimization (MO-PPO) algorithm to jointly optimize UAV trajectory, RIS phase shifts, and NOMA power control, aiming to maximize the network sum rate while meeting user quality of service (QoS) requirements. MO-PPO effectively handles the hybrid nature of these optimization parameters, encompassing both continuous and discrete values. Through extensive simulations, we demonstrate the effectiveness of MO-PPO 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) have emerged as a promising technology to address these demands, offering inherent mobility and flexibility for deployment as aerial base stations. 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 consist of numerous passive reflecting elements that can be dynamically adjusted to control the propagation of electromagnetic waves. 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 allows multiple users to share the same time-frequency resources, improving spectrum utilization and performance metrics like 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 integration of RIS, CoMP, and NOMA technologies within UAV-assisted networks holds immense potential for enhancing the performance and efficiency of future wireless communication systems. Recent research has explored the benefits of combining UAVs with RIS, CoMP, and NOMA, demonstrating significant improvements in network capacity and energy efficiency [8], [9]. However, these works often assume static RIS deployments, which limits the degree of freedom and adaptability of the network. While some studies have investigated ARIS-assisted CoMP-NOMA networks, optimizing UAV trajectory and RIS phase shifts to maximize the network sum rate [10], the optimization approaches used, such as double-layer alternating optimization, may not be scalable for large-scale networks due to their complexity and potential convergence issues.

To address the challenges of optimizing ARIS-assisted CoMP-NOMA networks, we propose a deep reinforcement learning (DRL) approach, specifically leveraging the multi-output proximal policy optimization (MO-PPO) algorithm to handle 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 the network sum rate while adhering to user QoS constraints. Through extensive simulations, we evaluate the efficacy of our proposed approach, assess the convergence of

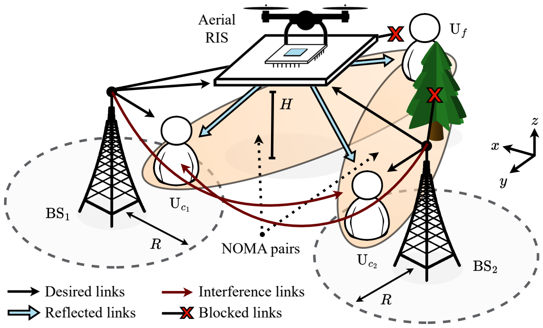

Fig. 1. Aerial RIS-assisted coodinated NOMA cluster.

MO-PPO, and highlight the benefits of CoMP-NOMA and RIS in UAV-assisted networks.

## II. System Model & Problem Formulation

### A. System Description

As shown in fig. 1, we consider a multi-cell CoMP-NOMA network assisted by a UAV-mounted RIS in a downlink transmission scenario. 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 $\mathrm{BS}_i$, where $i \in \mathcal{I} \triangleq \{1, 2, \ldots, I\}$. Each $\mathrm{BS}_i$ 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 $\mathrm{U}_{c_i}$, $\forall i$ 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 $\mathrm{U}_f$, $\forall f$ 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 $H$ 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 $\tau$, 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 $\mathrm{O}_o$, $o \in \mathcal{O}$ has its own *forbidden zone* represented as a circular disk of radius $d_{\text{min}}$, centered at the obstacle's location. The forbidden zone is an area 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 $\mathrm{BS}_i$, $\mathrm{U}_u$, and $\mathrm{O}_o$ are represented by $\mathbf{p}_i = (x_i, y_i, H_\mathrm{B})$, $\mathbf{p}_u = (x_u, y_u, 0)$, and $\mathbf{p}_o = (x_o, y_o, H_\mathrm{O})$, respectively, where $H_\mathrm{B}$ and $H_\mathrm{O}$ are the heights of the BSs and obstacles, respectively. Moreover, the position of $R$ at time slot $t$ is denoted as $\mathbf{p}_\mathrm{R}[t] = (x_\mathrm{R}[t], y_\mathrm{R}[t], H_\mathrm{R})$. 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 $H_\mathrm{R}$.

### B. Channel Model & RIS Configuration

In our analysis, both the large-scale path loss and small-scale fading effects are considered. Similar to [8], we assume the presence of numerous scatterers in the environment, and thus the direct links between $\mathrm{BS}_i$ and $\mathrm{U}_u$, denoted as $h_{i,u}$, are modeled as Rayleigh fading channels. 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 $\rho_o$ 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, and $d_{i,u} = \|\mathbf{p}_i - \mathbf{p}_u\|$ is the distance between $\mathrm{BS}_i$ and $\mathrm{U}_u$ 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 $\mathrm{BS}_i$ and $\mathrm{U}_f$ 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 $\mathrm{BS}_i$ and $R$ are modeled as Rician fading channels, denoted as $\mathbf{h}_{i,\mathrm{R}}[t]$, due to the presence of a dominant line-of-sight (LoS) component. At time slot $t$, the channel $\mathbf{h}_{i,\mathrm{R}}[t]$ is given by

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

where $\kappa$ is the Rician factor representing the ratio of the power of the LoS component to the power of the scattered components, and $d_{i,\mathrm{R}}[t] = \|\mathbf{p}_i - \mathbf{p}_\mathrm{R}[t]\|$ is the distance between $\mathrm{BS}_i$ and $R$. Moreover, the deterministic LoS represented, i.e., $\mathbf{g}_{i,\mathrm{R}}^{\mathrm{LoS}}[t] \in \mathbb{C}^{K \times 1}$, is given by

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

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

For the RIS 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

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

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. In this work, we assume an ideal RIS with perfect phase control and all reflection elements having a unit amplitude, i.e., $a_k = 1, \forall k$. Furthermore, we assume the availability of perfect channel state information (CSI) at the CPU. While this is a challenging assumption in practice, recent advancements in channel estimation techniques for RIS-assisted wireless networks have demonstrated the feasibility of achieving accurate CSI with reasonable overhead [11]–[13].

### C. Signal Model

In accordance with the NOMA principle, each $\text{BS}_i$ serves two users, $\text{U}_{c_i}$ and $\text{U}_f$, simultaneously, by superimposing their signals. Let $x_{i,c_i}[t]$ and $x_{i,f}[t]$ be the desired signals intended for $\text{U}_{c_i}$ and $\text{U}_f$ at time slot $t$, then the transmitted signal from $\text{BS}_i$ can be expressed as $x_i[t] = \sqrt{(1-\lambda_i)P_i}x_{i,c_i}[t] + \sqrt{\lambda_i P_i}x_{i,f}[t]$, where $P_i$ is the transmit power of $\text{BS}_i$ and $\lambda_i$ is the power allocation factor assigned to $\text{U}_f$, such that $\lambda_i \in (0.5, 1)$ to ensure successful decoding at $\text{U}_{c_i}$ [14], [15].

The received signal at $\text{U}_f$ can be expressed as

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

where $i' \in \mathcal{I} \setminus \{i\}$, $n_o[t] \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN), and $H_{i,f}[t] = \mathbf{h}^T_{\text{R},f}[t]\mathbf{\Theta}[t]\mathbf{h}_{i,\text{R}}[t]$ represents the effective channels between $\text{BS}_i$ and $\text{U}_f$ 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

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

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 received signal at $\text{U}_{c_i}$ can be expressed as

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

where $H_{i,c_i}[t] = h_{i,c_i}[t] + \mathbf{h}^T_{\text{R},c_i}[t]\mathbf{\Theta}[t]\mathbf{h}_{i,\text{R}}[t]$ represents the effective channels between $\text{BS}_i$ and $\text{U}_{c_i}$ 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 $\text{U}_{c_i}$. Based on the SIC principle, $\text{U}_{c_i}$ 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 $\text{U}_{c_i}$ for decoding $x_{i,f}[t]$ is given by

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

whereas the SINR at $\text{U}_{c_i}$ for decoding $x_{i,c_i}[t]$ is

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

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

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

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 $\text{U}_{c_i}$ and $\text{U}_f$, respectively.

### D. Problem Formulation

In this work, our primary objective is to maximize the sum rate achieved over $T$ time slots. To achieve this goal, 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 $\mathbf{\Lambda} \triangleq \{\lambda_i, \forall i\}$. The optimization problem can be mathematically formulated as

$$\max_{\mathbf{P}, \mathbf{\Theta}, \mathbf{\Lambda}} \quad \sum_{t \in \mathcal{T}} R_{\text{sum}}[t] \tag{10a}$$

$$\text{s.t.} \quad x_\text{R}[t], y_\text{R}[t] \in A, \ \forall t \in \mathcal{T}, \tag{10b}$$

$$\big\|\mathbf{p}_\text{R}[t] - \mathbf{p}_o\big\| \geq d_{\min}, \ \forall o \in \mathcal{O}, \ t \in \mathcal{T}, \tag{10c}$$

$$\theta_k[t] \in [-\pi, \pi), \ \forall k \in \mathcal{K}, \ t \in \mathcal{T}, \tag{10d}$$

$$R_{c_i}[t] \geq R^{\min}_{c_i}, \ \forall i \in \mathcal{I}, \ t \in \mathcal{T}, \tag{10e}$$

$$R_f[t] \geq R^{\min}_f, \ \forall f \in \mathcal{F}, \ t \in \mathcal{T}, \tag{10f}$$

$$\lambda_i \in (0.5, 1), \ \forall i \in \mathcal{I}, \tag{10g}$$

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 $\mathcal{R}^{\min}_{c_i}$ and $\mathcal{R}^{\min}_f$, for $\text{U}_{c_i}$ and $\text{U}_f$, 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

Before proceeding with the DRL-based solution, we model it as a single-agent Markov Decision Process (MDP) with discrete time steps. This 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 $\gamma$ 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 $\mathbf{p}_\text{R}[t]$, the distance from the UAV to the center of obstacles $\mathbf{d}_\text{R}[t] = \{\big\|\mathbf{p}_\text{R}[t] - \mathbf{p}_o\big\|, \forall o \in \mathcal{O}\}$, the power allocation factors $\mathbf{\Lambda}$, and the achievable rates $\mathbf{R}[t] = \{R_{c_i}[t], R_f[t], \forall i, f\}$. Thus, the state space can be expressed as

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

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

2) *Action Space $\mathcal{A}$:* The action space of the formulated MDP consists of the UAV's movement in the horizontal $xy$-plane, the phase shifts of the RIS elements, and the power allocation factors. Specifically, the action space at time slot $t$ contains the manuevering actions $\mathbf{a}_{\text{R}}[t] \in \{(-1,0),(1,0),(0,-1),(0,1),(0,0)\}$, representing left, right, down, up, and hover, respectively, the phase shifts $\mathbf{a}_{\Phi}[t] = \{\phi_k[t], \forall k\}$, and the power allocation factors $\mathbf{a}_{\Lambda} = \{\lambda_i, \forall i\}$. Thus, the action space can be expressed as

$$\mathbf{a}_t = \{\mathbf{a}_{\text{R}}[t], \mathbf{a}_{\Phi}[t], \mathbf{a}_{\Lambda}\} \in \mathbb{R}^{\text{dim}_{\mathcal{A}}}. \tag{12}$$

where $\text{dim}_{\mathcal{A}} = 2 + K + I$ 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 encourages maximizing the sum rate while ensuring UAV safety and meeting QoS requirements by penalizing constraint violations. The reward function is defined as

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

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

### B. MO-PPO Algorithm

In this work, the considered action space is a hybrid continous-discrete space, which poses a challenge for traditional RL algorithms. While discretization of continuous actions is a possibility, it can lead to a large action space, significantly increasing computational complexity and potentially hindering performance. To address this challenge, we propose employing a multi-output Proximal Policy Optimization (MO-PPO) algorithm. MO-PPO extends the standard PPO [17] framework by employing two parallel actor networks, each responsible for generating the discrete action $\mathbf{a}_{\text{R}}$ and the continuous actions $\mathbf{a}_{\Phi}$ and $\mathbf{a}_{\Lambda}$, respectively. The actor networks share the first few layers, allowing for 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 [18], 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 $|\mathbf{a}_{\text{R}}|$ logits, which are then passed through a $\text{softmax}$ function to obtain a probability distribution over the available discrete actions. Conversely, the continuous actor network generates the continuous actions $\mathbf{a}_{\Phi}$ and $\mathbf{a}_{\Lambda}$ 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

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

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

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

**Algorithm 1:** MO-PPO Algorithm

```
Initialize the policy parameters $\theta_d$ and $\theta_c$
for episode $= 1, 2, \dots, N$ do
    Receive initial state $s_0$
    for time step $t = 0, 1, \dots, T$ do
        Generate discrete action $a_{\text{R}}$ using $\pi_{\theta_d}(a_t|s_t)$
        Generate continuous actions $\mathbf{a}_{\Phi}$ and $\mathbf{a}_{\Lambda}$ using
          $\pi_{\theta_c}(a_t|s_t)$
        Execute actions $\mathbf{a}_t = \{a_{\text{R}}, \mathbf{a}_{\Phi}, \mathbf{a}_{\Lambda}\}$
        if UAV violates (10b) or (10c) then
            Set $\xi_{\text{R}}[t] = 1$, cancel the UAV's movement,
              and update the state $s_{t+1}$
        end
        Observe reward $\mathcal{R}$ as (13) and next state $s_{t+1}$
        Collect a set of partial trajectories $\mathcal{D}$ with $\hat{T}$
          transitions
        Compute the advantage estimate $\hat{A}_t$ as (14)
    end
    for epoch $= 1, 2, \dots, E$ do
        Sample a mini-batch of transitions $B$ from $\mathcal{D}$
        Compute the clipped surrogate objectives
          $L_d^{\text{CLIP}}(\theta_d)$ and $L_c^{\text{CLIP}}(\theta_c)$ as (15)
        Optimize overall objective and update the
          policy parameters $\theta_d$ and $\theta_c$ using Adam [16]
    end
    Synchronize the sampling policies as
        $\theta_d^{\text{old}} \leftarrow \theta_d$ and $\theta_c^{\text{old}} \leftarrow \theta_c$
    Clear the collected trajectories $\mathcal{D}$
end
```

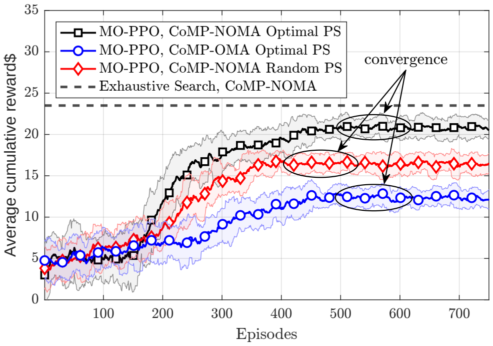

Fig. 2. Average cumulative reward vs. number of training episodes with $P_t = 20$ dBm and $K = 120$ elements.

## 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 $150 \times 150$ m$^2$ with $\mathcal{I} = 2$ BSs, $\mathcal{U} = 3$ users, and $\mathcal{O} = 2$ 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 $\alpha_{i,u} = 3$, $\alpha_{i,\mathrm{R}} = \alpha_{\mathrm{R},u} = 2.2$, 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 successful acquisition of an effective policy. Notably, MO-PPO with random 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.

TABLE I
Simulation Parameters

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Reference path loss $\rho_o$ | $-30$ dBm | Rician factor $\kappa$ | 3 dB |
| Target data rate $R_f^{\min}$ | 0.2 bps/Hz | Learning rate | $2.75\mathrm{e}-4$ |
| Target data rate $R_{c_i}^{\min}$ | 0.5 bps/Hz | Clipping parameter $\epsilon$ | 0.1 |
| Penalty constant $K_{\text{viol}}$ | 7 | Discount factor $\gamma$ | 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 |

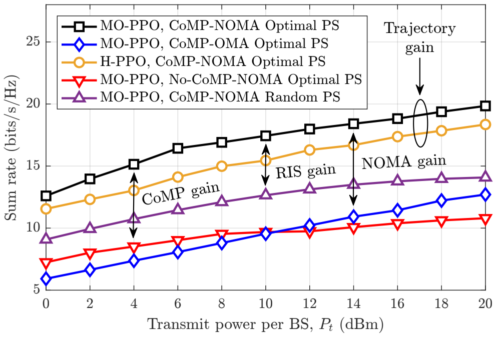

Fig. 3. Sum rate vs. transmit power for different algorithms and configurations with $K = 120$ elements.

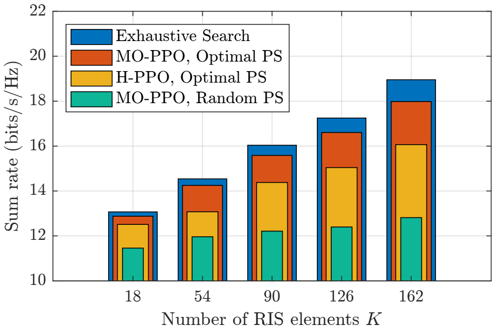

Fig. 4. Impact of the number of RIS elements on the achievable sum rate with $P_t = 10$ dBm.

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. Additionally, we benchmark the hover PPO (H-PPO) algorithm, which maintains a fixed UAV position at the center of the user clusters, against the proposed MO-PPO algorithm, and highlight the improvement in network performance achieved due to the trajectory optimization.

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 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

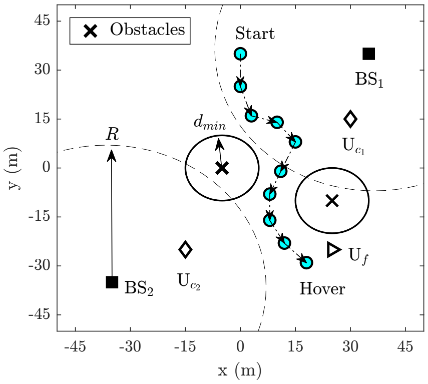

Fig. 5. Top view of the UAV trajectory obtained by the MO-PPO algorithm sampled every 25 time slots and averaged over 10 evaluation episodes

baseline and the MO-PPO algorithm, while remaining small, increases with the number of RIS elements. This observation highlights the importance 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, showcasing the advantages of integrating ARIS, CoMP, and NOMA for future wireless networks. 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] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 2, pp. 1035–1071, 2022.

[12] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, "Channel estimation for RIS-aided multiuser millimeter-wave systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, 2022.

[13] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, 2021.

[14] 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.

[15] 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.

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

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

[18] 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.