RIS-Assisted ABS for Mobile Multi-User MISO Wireless Communications: A Deep Reinforcement Learning Approach

Walaa AlQwider, Aly Sabri Abdalla, and Vuk Marojevic
Department of Electrical and Computer Engineering, Mississippi State University, MS 39762, USA
Email: wq27@msstate.edu, asa298@msstate.edu, vuk.marojevic@msstate.edu

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622221

***Abstract*—In response to the evolving landscape of wireless communication networks and the escalating demand for unprecedented wireless connectivity performance in the forthcoming 6G era, this paper proposes a new 6G architecture to enhance the wireless network's sum rate performance. Therefore, we introduce an aerial base station (ABS) network with reconfigurable intelligent surfaces (RISs) while leveraging the multi-users multiple-input single-output (MU-MISO) antenna technology. The motivation behind our proposal stems from the imperative to address critical challenges in contemporary wireless networks and harness emerging technologies for substantial performance gains. We employ deep reinforcement learning (DRL) to jointly optimize the ABS trajectories, the active beamforming weights, and the RIS phase shifts. Simulation results show that this joint optimization effectively improves the system's sum rate while meeting minimum quality of service (Qos) requirements for diverse mobile users.**

***Index Words*—6G wireless, deep reinforcement learning, eavesdropping, RIS, sum rate, QoS, UAV, ABS, DDPg.**

I. Introduction

The rapid growth of wireless communications services has increased the need for advanced wireless technologies. Legacy communication systems often suffer from disruptions in connectivity and inadequate quality of services (QoS) resulting from wireless channel impairments. The reconfigurable intelligent surface (RIS) was introduced [1] to steer the radio frequency (RF) propagation and, thus, control the channel. A RIS is an engineered, planar meta-surface constructed from numerous passive antenna elements, each of which can be electronically controlled to create controllable radio environments [2]. The primary knobs of a RIS are its configurable phase shifters, which enable precise control over the RF propagation. By strategically manipulating the phase of incident electromagnetic waves, the RIS can manipulate the signal path, which can lead to improved coverage and communication quality [3] compared to conventional passive reflectors or antennas.

The rapid progress of unmanned aerial vehicle (UAV) technologies has motivated research on integrating UAVs into wireless communication networks [4]. UAVs, when employed as aerial base stations (ABS), offer a promising solution to increase coverage or capacity on the fly. ABSs can often establish line-of-sight (LoS) links with ground users, thus enhancing communication reliability [5]. The integration of ABSs with RISs to enhance the wireless system performance has attracted considerable attention in the recent years [6]. Theoretical research has shown significant improvements by leveraging multi-user multiple-input multiple-output (MU-MIMO) technology for RIS deployments.

There has been growing interest in using RISs and UAVs for mobile networks. However, there is a research gap on the integration of RIS with ABSs in conjunction with MISO-MU systems, particularly for scenarios involving user mobility. The authors of [7] focus on optimizing the passive beamforming of the RIS and designing the trajectory of the ABS in a wireless environment for a single ground user at a fixed location. Reference [6] introduces an alternating optimization algorithm to address the complex sum rate maximization problem for RIS-assisted UAV networks. This involves optimizing the UAV trajectory, phase shifter design, and resource allocation for an orthogonal frequency division multiple access (OFDM) system. It is worth noting that the RIS serves only one user at a time and that the user locations are assumed to be fixed. Another related work [8] tackles the problem of system sum rate maximization in an ABS-assisted network with an RIS. The sum rate is improved by jointly optimizing the RIS's phase shifts and ABS's altitude, employing a conjugate gradient particle swarm optimization (CG-PSO) scheme.

This paper stands at the intersection of several key advancements in wireless communication technologies—ABS, RIS, and MU-MIMO systems. The emphasis is on sum rate maximization while adhering to individual user QoS requirements in terms of minimum data rates by jointly optimizing the active beamforming MU-MISO system, the RIS phase shifts, and the ABS trajectory in a mobile multi-users scenario. Given the complexity of the problem, we propose applying deep reinforcement learning (DRL). The considered users encompass both vehicular and pedestrian users, exhibiting distinct mobility characteristics. By considering the mobility dynamics of these users, we aim to tailor our approach to scenarios where traditional communication systems often fall short, thereby paving the way for a more adaptable and robust communication infrastructure.

The rest of the paper is organized as follows: Section II introduces the system model, followed by the formulation of the optimization problem. Section IV introduces the user clustering and DRL schemes based on the deep deterministic

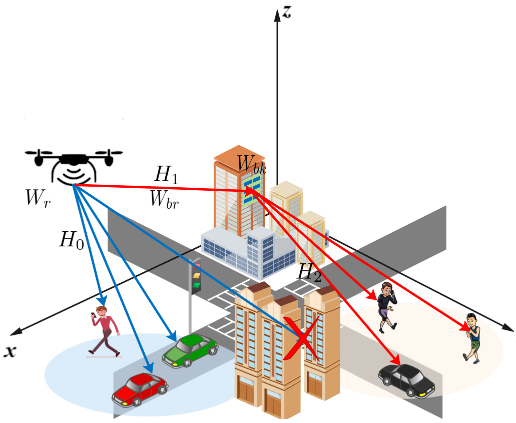


Fig. 1: RIS-assisted MU-MISO communications from an ABS.

policy gradient (DDPG) for jointly optimizing the active beamforming, ABS trajectories, and RIS phase shifts. The numerical analysis of Section V shows the effectiveness of the proposed approach. Section VI draws the conclusions.

II. System Model

A. System Model

We investigate a MU-MISO communication system, where the UAV is deployed as an aerial base station (ABS) equipped with M antennas and responsible for delivering downlink communications to multiple mobile ground terminals (GTs). An RIS with L reflecting elements is deployed on one of the surrounding buildings and is leveraged to steer the transmissions originating at the ABS to the GTs that suffer from high blockage or severe interference in their direct channels with the ABS. Figure 1 depicts this scenario.

The total flight time T of the ABS is split into N time slots of duration δ_t=$\frac{T}{N}$. The ABS hovers at $\boldsymbol{q}_A[n] = [x[n], y[n], z]^T \forall n \in N$ at a fixed height z_t. For the sake of simplicity and without loss of generality, we do not consider optimizing the ABS height in this paper. The height z_t is chosen to enable LoS communication links to ground users unobstructed by obstacles in ABS's proximity [9]. There are K single antenna GTs served by the ABS, where $K \leq M$.

Two types of GTs are assumed: vehicular and pedestrian GTs with different mobility models. The location of each GT in time slot n is $\boldsymbol{q}_k[n] = [x_k[n], y_k[n], 0]^T \forall k \in K, \forall n \in N$. The RIS, which is at a fixed location $\boldsymbol{q_R} = [x_r, y_r, z_r]^T$, receives incoming signals and utilizes its configurable reflective elements to redirect these signals toward the K GTs. For the considered MU-MISO communication system, each GT experiences signal reception through one of two primary communication routes: The first is the direct transmission from the ABS to the GT and the second is the indirect transmission through the RIS. The ABS employs its array of M antennas to transmit K distinct data streams to the RIS simultaneously, one for each each GT.

The communication channel between the M antennas of ABS A and the L reflecting elements of RIS R in time slot $n \in N$ is modeled as a multiple input, multiple output (MIMO) channel and denoted as $\boldsymbol{H_{AR}[n]} \in \mathbb{C}^{L \times M}$. The MISO channels between ABS A and GT k and between RIS R and GT k in time slot $n \in N$ are defined as $\boldsymbol{h_{Ak}[n]} \in \mathbb{C}^{M \times 1}$ and $\boldsymbol{h_{Rk}[n]} \in \mathbb{C}^{L \times 1}$, respectively, $\forall k$. We assume that the ABS has perfect knowledge of the channel state information (CSI) and conveys this information to the RIS controller through a dedicated control channel. The L reflecting elements of the RIS are interconnected to form a uniform linear array (ULA) as in [10]. The phase shift array in time slot $n \in N$ is denoted as $\boldsymbol{\phi}[n] \in \mathbb{C}^{L \times L}, \boldsymbol{\phi}[n] = diag\{e^{j\theta_1[n]}, e^{j\theta_2[n]}, \cdots, e^{j\theta_L[n]}\}$, where $\theta_l[n] \in [0, 2\pi), l \in [1, 2, ..., L]$ is the phase of the l^{th} element.

B. Channel Model

We model the air-to-ground communications channel between the ABS and the RISs and GTs using small-scale Rician fading, which includes line of sight (LoS) and non-LoS (NLoS) components [11]. Equation

$$\boldsymbol{H}_{AR}[n] = \frac{\sqrt{\lambda_0}}{D^{\alpha}_{AR}[n]} \left(\sqrt{\frac{\beta}{1+\beta}}\, \boldsymbol{H^L_{AR}}[n] + \sqrt{\frac{1}{\beta+1}}\, \boldsymbol{H^N_{AR}}[n] \right) \tag{1}$$

represents the MIMO communications channel between the M antennas of the ABS and the L reflecting elements of the RIS in time slot n. Parameter λ_0 is the path loss at the reference distance of 1 m, $D_{AR}[n]$ is the 3D distance between the ABS and the RIS, α is the path loss exponent, β corresponds to the Rician factor, and $\boldsymbol{H^L_{AR}[n]}$ and $\boldsymbol{H^N_{AR}[n]}$ represent the LoS and NLoS channel components, respectively.

Without loss of generality, the entries in $\boldsymbol{H^N_{AR}}$ are considered to be independent and identically distributed (i.i.d.). These entries are modeled as zero-mean and unit-variance circularly symmetric complex Gaussian (CSCG) variables: $\mathcal{CN}(0, 1)$. The LoS channel gain results from the angle of departure (AoD) channel at the ABS and the angle of arrival (AoA) channel at the RIS:

$$\boldsymbol{H^L_{AR}}[n] = \boldsymbol{H^{(A)}_{AR}}[n]\ \boldsymbol{H^{(D)}_{AR}}[n]. \tag{2}$$

The AoD channel contribution is

$$\boldsymbol{H^{(D)}_{AR}}[n] = \left[1, e^{-j\frac{2\pi}{\lambda}\Upsilon\Gamma^{AR}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(M-1)\Upsilon\Gamma^{AR}[n]}\right], \tag{3}$$

where λ represents the carrier wavelength, Υ the antenna separation, and $\Gamma^{AR}[n]$ the AoD component. The AoD component can be expressed as $\Gamma^{AR}[n] = \sin\vartheta[n]\cos\psi[n]$, with $\vartheta[n]$ representing the elevation AoD and $\psi[n]$ representing the azimuth AoD originating from the ULA antennas at the ABS.

The AoA can be calculated as

$$\boldsymbol{H^{(A)}_{AR}}[n] = \left[1, e^{-j\frac{2\pi}{\lambda}\Upsilon\Lambda^{AR}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(L-1)\Upsilon\Lambda^{AR}[n]}\right], \tag{4}$$

where $\Lambda^{AR}[n] = cos\ \Theta[n]\ sin\ \varphi[n]$ is the AoA component of the transmitted signal from the ABS to the RIS, $\Theta[n]$ corresponds to the azimuth AoA and $\varphi[n]$ represents the elevation AoA.

The MISO channels between the ABS and GT k and the RIS and GT k in time slot n are modeled as

$$\boldsymbol{h}_{Ak}[n] = \frac{\sqrt{\lambda_0}}{D_{Ak}^{\alpha}[n]}\Bigg(\sqrt{\frac{\beta}{1+\beta}}\ \boldsymbol{h}_{\boldsymbol{Ak}}^{\boldsymbol{L}}[n] + \sqrt{\frac{1}{\beta+1}}\ \boldsymbol{h}_{\boldsymbol{Ak}}^{\boldsymbol{N}}[n]\Bigg),$$
$$\boldsymbol{h}_{Rk}[n] = \frac{\sqrt{\lambda_0}}{D_{Rk}^{\alpha}[n]}\Bigg(\sqrt{\frac{\beta}{1+\beta}}\ \boldsymbol{h}_{\boldsymbol{Rk}}^{\boldsymbol{L}}[n] + \sqrt{\frac{1}{\beta+1}}\ \boldsymbol{h}_{\boldsymbol{Rk}}^{\boldsymbol{N}}[n]\Bigg). \tag{5}$$

The same CSCG distribution defined earlier is followed by $\boldsymbol{h}_{\boldsymbol{Ak}}^{\boldsymbol{N}}$ and $\boldsymbol{h}_{\boldsymbol{Rk}}^{\boldsymbol{N}}$. Parameters $D_{Ak}[n]$ and $D_{Ak}[n]$ represent the 3D distance between the ABS and the k^{th} GT and between the RIS and the k^{th} GT in time slot n, respectively.

The LoS MISO channel components between the ABS and a GT and between the RIS and a GT in time slot n are modeled as

$$\begin{aligned} \boldsymbol{h}_{\boldsymbol{Ak}}^{\boldsymbol{L}}[n] &= \Big[1, e^{-j\frac{2\pi}{\lambda}\Upsilon\chi^{Ak}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(M-1)\Upsilon\chi^{Ak}[n]}\Big],\\ \boldsymbol{h}_{\boldsymbol{Rk}}^{\boldsymbol{L}}[n] &= \Big[1, e^{-j\frac{2\pi}{\lambda}\Upsilon\chi^{Rk}[n]}, \cdots, e^{-j\frac{2\pi}{\lambda}(L-1)\Upsilon\chi^{Rk}[n]}\Big], \end{aligned} \tag{6}$$

where $\chi^{Ak} = cos\ \Phi^{Ak}[n]\ sin\ \Omega^{Ak}[n]$ and $\chi^{Rk}[n] = cos\ \Phi^{Rk}[n]\ sin\ \Omega^{Rk}[n]$ represent the AoD components of the transmissions originating from the ABS and RIS, respectively. These are determined by $\Phi[n]$ as the azimuth AoD and $\Omega[n]$ as the elevation AoD.

C. Data Rate

The M-antenna ABS serves the single antenna k-th user either directly or through the L-element RIS utilizing the same frequency, employing space-division multiple access (SDMA) and time-division multiple access (TDMA). In continuation we provide the user data rate calculations for both links, the direct and indirect links.

1) Data Rate of the Direct Link: The ABS simultaneously generates M concurrent beams to K spatially separated users using SDMA. Each user is assigned a dedicated beam vector for transmit beamforming. However, the presence of power leakage between beams within small proximity at the receivers introduces multi-user interference.

In time slot n,

$$\boldsymbol{x}[n] = \sum_{k=1}^{K} \boldsymbol{w}_k[n]s_k[n] \tag{7}$$

represents the downlink transmit signals, where $\boldsymbol{x}[n] \in \mathbb{C}^{M\times 1}$, $\boldsymbol{w}_k[n] \in \mathbb{C}^{M\times 1}$ is the ABS beamforming vector, and $s_k[n]$ is the transmitted information symbol for the k-th user in time slot n. The beamforming or precoding matrix of the ABS has K beamforming vectors $\boldsymbol{W}_k[n] = [\boldsymbol{w}_1[n], \dots, \boldsymbol{w}_K[n]] \in \mathbb{C}^{M\times K}$. The allocated transmit power for the k-th user can be computed as the squared norm of the beamforming vector: $\|\boldsymbol{w}_k[n]\|^2$. Therefore, the received signal at the k-th GT through the direct link can be expressed as

$$\begin{aligned} y_{0,k}[n] &= \boldsymbol{h}_{Ak}[n]\boldsymbol{x}[n] + n_k,\\ &= \boldsymbol{h}_{Ak}[n](\sum_{k=1}^{K}\boldsymbol{w}_k[n]s_k[n]) + n_k,\\ &= \boldsymbol{h}_{Ak}[n]w_k[n]s_k[n] + \boldsymbol{h}_{Ak}[n](\sum_{i=1,i\neq k}^{K}\boldsymbol{w}_i[n]s_i[n]) + n_k, \end{aligned} \tag{8}$$

where n_k represents the additive white Gaussian noise (AWGN). It is assumed that the noise at each user follows a complex normal distribution of zero-mean and unit variance: $n_k \sim \mathcal{CN}(0, 1)$. The signal-to-interference-plus-noise-ratio (SINR) at the k-th GT can be calculated as

$$\gamma_{0,k}[n] = \frac{|\ (\boldsymbol{h}_{Ak}[n]\boldsymbol{w}_k[n])\ |^2}{\sum\limits_{i\neq k}^{K} |\ (\boldsymbol{h}_{Ak}[n]\boldsymbol{w}_i[n])\ |^2 + \sigma_k^2}, \tag{9}$$

where the first term in the denominator corresponds to the multi-user interference of the MISO communications system and σ_k^2 is the noise variance. The resulting normalized data rate of the k-th GT served via the direct link in time slot n is then obtained as

$$R_{0,k}[n] = \log_2\left(1 + \gamma_{0,k}[n]\right), \tag{10}$$

2) Data Rate of the Indirect Link: The received signal at the k-th GT on the indirect link through the RIS can be expressed as follows:

$$\begin{aligned} y_{1,k}[n] &= \boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\boldsymbol{H}_{AR}[n]\boldsymbol{x}[n] + n_k\\ &= \boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\ \boldsymbol{H}_{AR}[n]\big(\sum_{k=1}^{K}\boldsymbol{w}_k[n]s_k[n]\big) + n_k,\\ &= \boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\ \boldsymbol{H}_{AR}[n]w_k[n]s_k[n] + \boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\\ &\qquad \boldsymbol{H}_{AR}[n]\big(\sum_{i=1,i\neq k}^{K}\boldsymbol{w}_i[n]s_i[n]\big) + n_k. \end{aligned} \tag{11}$$

The SINR at the k-th GT served through the RIS is

$$\gamma_{1,k}[n] = \frac{|\ (\boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\ \boldsymbol{H}_{AR}[n]\boldsymbol{w}_k[n])\ |^2}{\sum\limits_{i\neq k}^{K} |\ (\boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n]\ \boldsymbol{H}_{AR}[n]\boldsymbol{w}_i[n])\ |^2 + \sigma_k^2} \tag{12}$$

and the resulting normalized data rate

$$R_{1,k}[n] = \log_2\left(1 + \gamma_{1,k}[n]\right). \tag{13}$$

D. Total Data Rate

Considering the channel models and mobility models discussed in the previous section, the average achievable downlink data rate $R_k[n]$ in bits/s/Hz of the k-th GT up to time slot n can be calculated as

$$R_k[n] = \frac{1}{n}\cdot\sum_{i=1}^{n} u_k[i]R_{0,k}[i] + (1 - u_k[i])R_{1,k}[i]. \tag{14}$$

Expression $u_k[i] = 1$ if GT k is served by the direct link in time slot i and $u_k[i] = 0$, otherwise.

The average sum data rate over all GTs until time slot n is

$$R[n] = \sum_{i=k}^{K} R_k[n]. \tag{15}$$

III. Problem Formulation

Incorporating ABSs and RISs together in a dynamic mobility environment enables maintaining reliable multi-user communications despite heterogeneous user mobility patterns. This can be achieved by strategically positioning the ABS to accomplish direct LoS communications with ground users or establish controlled reflected propagation paths through the RIS. By employing an MU-MIMO ABS and an RIS, the optimization parameters are the beamforming matrix $\mathbf{W}$, the phase shifters of the RIS ϕ, and the trajectory of the ABS $\mathbf{q}_A$ in addition to the decision of which users should be served by the direct link and which should be served by the RIS, captured by $\mathbf{U}$. The optimization problem is formulated as

$$\begin{aligned}
&\mathcal{P}\left(\mathbf{U},\mathbf{W},\mathbf{q}_A,\boldsymbol{\phi}\right): \max_{\mathbf{U},\mathbf{W},\mathbf{q}_A,\boldsymbol{\phi}} \sum\nolimits_{n=1}^{N} R[n]\left(\mathbf{U},\mathbf{W},\mathbf{q},\boldsymbol{\phi}\right)\\
&\text{s.t.}\\
&\quad \text{C1}: \ u_k[n]\in\{0,1\},\ \forall k,n,\\
&\quad \text{C2}: \ R_k[n]\geq R_{\min,k},\forall k,\\
&\quad \text{C3}: \ \sum\nolimits_{k=1}^{K}\|\boldsymbol{w}_k[n]\|^2\leq P_{\max}, \qquad\qquad (16)\\
&\quad \text{C4}: \ \left\|e^{j\theta_l[n]}\right\|=1,\ \forall l,n,\\
&\quad \text{C5}: \ \|\mathbf{q}_A\left[n\right]-\mathbf{q}_A\left[n-1\right]\|\leq\delta_{\mathrm{t}}V_{\max},\forall n,\\
&\quad \text{C6}: \ \mathbf{q}\left[0\right]=\mathbf{q}_A(\text{Initial}),\\
&\quad \text{C7}: \ \mathbf{q}_A\left[N\right]=\mathbf{q}_A(\text{Final}).
\end{aligned}$$

Expression $R_{min,k}$ in constraint C2 denotes the minimum average data rate required for the k-th GT. This criterion is required to meet the user-specific QoS. Constraint C3 sets a boundary on the maximum allowable transmit power P_{max} for the ABS. Constraint C5 ensures that the ABS does not travel beyond the specified maximum speed limit V_{max}. Constraint C6 establishes the ABS's initial location $\boldsymbol{q}_A(Initial)$ and C7 the final location $\boldsymbol{q}_A(final)$.

IV. Proposed Solution

The optimization problem (16) is a non-convex mixed-integer optimization problem, which is known for its inherent complexity. This complexity primarily stems from the binary variable $u_k[n]$ and the non-convex nature of the achievable rate function embedded within both the objective function and constraint C2. Additionally, there exist intricate interdependencies among the optimization variables $\boldsymbol{\phi}$, $\boldsymbol{q}_A$, $\boldsymbol{W}$, and $\boldsymbol{U}$. Notably, the unit modulus constraint imposed on $\boldsymbol{\phi}$ has been demonstrated to be non-convex [10]. In addition, the optimization complexity associated with the phase shifts of the RIS directly scales with the number of elements, which is typically large. Therefore, it is imperative to devise optimization solutions that can efficiently handle a large number of reflective elements.

We propose solving (16) by leveraging data-driven approaches, which have demonstrated their efficacy in solving similar optimization problems [12]. Initially, we employ k-means clustering to assign GTs to either direct or indirect links. Subsequently, we employ DDPG to optimize the joint ABS trajectory, beamforming matrix, and RIS phase shifts.

A. User Clustering

The objective of user clustering is to partition the K users into two groups, one served directly by the ABS and the other served indirectly through the RIS. We employ K-means clustering, an unsupervised learning technique that maximizes the similarity within groups and the dissimilarity across groups. K-means clustering is known for its computational efficiency compared to alternative techniques such as graph theory, fuzzy c-means clustering, and hierarchical clustering [13].

We use the normalized channel coefficients between GTs and the ABS as the data points for the K-means clustering algorithm. These data points capture the fluctuations in channel gains arising from diverse propagation factors, including small-scale fading and shadow fading. Hence, we can define

$$h_{Ak}^{no}[n]=\frac{h_{Ak}[n]}{\parallel h_{Ak}[n]\parallel_2}, \qquad (17)$$

where $h_{Ak}^{no}[n]$ is the normalized channel gain and $h_{Ak}[n]$ is the channel gain between the ABS and k-th GT. Starting with random centroids for the two clusters the K-means algorithm starts to calculate the distances in terms of the normalized channel coefficient between each GT and the two cluster centers to assign each GT to its nearest center. Then, the centroids are updated to minimize the sum of the squared Euclidean distances between a clustered data point and its centroid. The Euclidean distance is the chosen metric in this paper to measure the similarities between data points, other metrics such as Manhattan distance, can be applied, instead.

B. DRL-Based Optimizer

We employ DDPG to tackle the joint optimization of beamforming, ABS trajectory, and RIS phase shifts. DDPG frames the problem as a Markov Decision Process (MDP), where the environment undergoes transitions from one state to another based on the actions taken and governed by transition probabilities.

1) MDP Settings: The MDP is structured around three components: the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the reward space $\mathcal{R}$. In time slot n, the agent observes the current state $s_n\in\mathcal{S}$ and, guided by its policy, selects an action $a_n\in\mathcal{A}$. Subsequently, the agent transitions into the new state s_{n+1} and receives reward $r_n\in\mathcal{R}$.

State: State s_n,

$$\begin{aligned}
&s_n=\{(G_1[n],R_1[n]),..,(G_k[n],R_k[n]),..,(G_K[n],R_K[n])\}\\
&G_k[n]=u_k[n]\cdot\mid(\boldsymbol{h}_{Ak}[n]\boldsymbol{w}_k[n-1])\mid^2+\\
&(1-u_k[n])\cdot\mid(\boldsymbol{h}_{Rk}[n]\boldsymbol{\phi}[n-1]\ \boldsymbol{H}_{AR}[n]\boldsymbol{w}_k[n-1])\mid^2,
\end{aligned} \qquad (18)$$

encompasses a collection of $2K$ elements that pertain to the CSI and the average data rate of each GT.

Action: Action a_n,

$$a_n=\left\{\boldsymbol{W}[n],\boldsymbol{\Phi}[n],\varphi[n],\omega[n]\right\}, \qquad (19)$$

has $K+L+2$ elements: K elements pertain to beamforming, L elements are associated with phase shifts, while the remaining two elements contribute to defining the trajectory of the ABS. Parameter φ represents the UAV's horizontal flight direction, while ω captures the distance of movement in this direction.

Reward: The reward function considers the average data rate and incorporates two penalties:

$$\begin{aligned}
&r_n=\left(\frac{\sum_{i=1}^{n}R[i]}{n}\right)-(P_1[n]+P_2[n]),\\
&P_1[n]=\max(0,\sum_{k=1}^{K}(R_{min,k}-R_k[n])),\\
&P_2[n]=\max(0,(\|\mathbf{q}_A\left[N\right]-\mathbf{q}_A\left[n\right]\|-(N-n)\cdot\delta_{\mathrm{t}}\cdot V_{\max})).
\end{aligned} \qquad (20)$$

The first penalty considers the discrepancy between the average data rate and the minimum data rate requirement for each GT, satisfying constraint C2. The second penalty considers the distance between the current location of the ABS and its final destination, while also assessing if there is enough remaining time to reach that destination, addressing constraint C7.

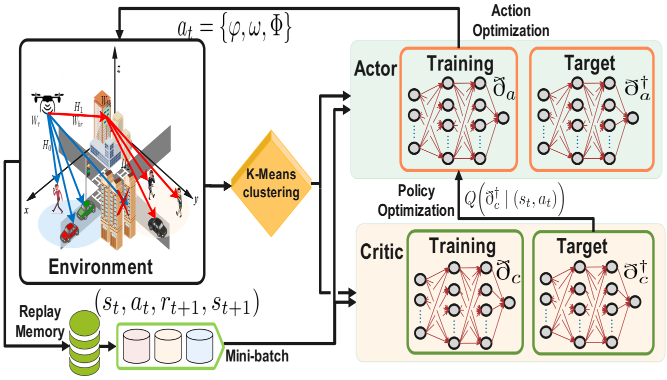


Fig. 2: Block diagram of the proposed DDPG architecture.

2) Deep Deterministic Policy Gradient: DDPG excels at handling complex, high-dimensional action spaces and continuous action domains [14]. It leverages two deep neural networks (DDN), the actor and critic network, to approximate both the policy and the value function [15]. This is illustrated in Fig. 2. In each time slot n, the DDPG agent gets the output of the clustering algorithm, the previous $\boldsymbol{W}$ and $\boldsymbol{\phi}$ values, and the CSI for each GT to construct state s_n, then feeds s_n to the actor network $\mho$ to determines action a_n, sends a_n to the ABS and the RIS controller, which execute the actions, calculates the reward, and generates a record of experience consisting of s_n, a_n, r_n, and the next state s_{n+1}, or $e_n = (s_n, a_n, r_n, s_{t+n})$. This experience is sent to a replay buffer of capacity $\aleph$ so that $\mathcal{M} = \{e_1, ..., e_n, ..., e_\aleph\}$ are used for training the actor and critic networks.

The actor network weight parameters are updated by taking a mini batch from the replay buffer and applying

$$\eth_a = \eth_a - \varrho_a \ \Delta_a Q\left(\eth_c^\dagger \mid (s_n, a_n)\right) \ \Delta_{\eth_a} \mho(\eth_a \mid s_n), \quad (21)$$

where $\eth_a$ denotes the actor network weights $\mho(\eth_a \mid s_n)$, $\eth_c^\dagger$ denotes the critic network weights, ϱ_a is the learning rate, $\Delta_a Q(\cdot)$ is the gradient of the target critic network output with reference to the taken action, and $\Delta_{\eth_a} \ \mho(\cdot)$ is the gradient of the training actor network with respect to $\eth_a$. The updates of the training critic network are obtained as

$$\eth_c = \eth_c - \varrho_c \ \Delta_{\eth_c} \ell(\eth_c). \quad (22)$$

Parameter $\ell(\eth_c)$ is the loss function of the training critic network and can be calculated as

$$\ell(\eth_c) = \mathbb{E}\Bigg[\Bigg(\left[r_t + \zeta \times \ Q\left(\eth_c^\dagger \mid (s_{t+1}, \tilde{a})\right)\right] - \left[Q\left(\eth_c \mid (s_t, a_t)\right)\right]\Bigg)^2\Bigg], \quad (23)$$

where $\tilde{a}$ is the agent's action that follows the deterministic policy drafted by the target actor network.

V. Numerical Analysis

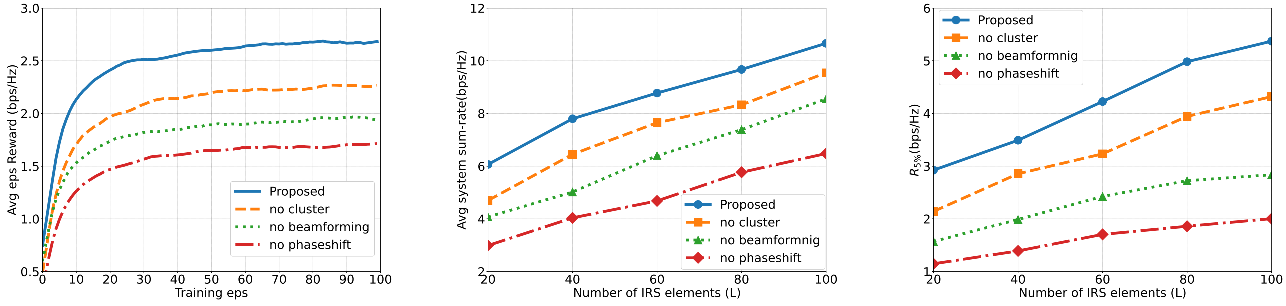
We evaluate the performance of the proposed scheme through simulations. The K GTs are distributed within an urban environment that features an intersection with vehicular users, a sidewalk with pedestrians, and a park alongside the road. Moreover, the environment is characterized by a multitude of high-rise buildings that may obstruct the signals from the ABS to certain GTs. The ABS maintains a fixed altitude of 100 m. The RIS is mounted on a building at a height of 70 m and faces the park.

The DDPG agent is constructed with the critic and actor networks employing the same DNN architecture. This architecture consists of six layers, including the input layer, four fully connected hidden layers, and the output layer. The input layer's dimension is set to $2K$, matching the state dimension. The hidden layers have $600, 400, 200, 100$ neurons. The output layer of the actor network has a dimension of $2L + 2M + 2$, where $2L$ represents the real and imaginary components of the complex phase shifts for the L-element RIS, $2M$ captures the complex beamforming weights of the M-antenna ABS, and the remaining two elements handle the ABS trajectory. All layers in both DNNs utilize the $(tanh)$ activation function and the Adam optimizer.

We assess the performance of the clustering step by introducing Baseline 1, which mirrors the proposed scheme but excludes the clustering. Baseline 2 implements the DDPG agent to solely optimize the RIS phase shifts and the ABS trajectory, while the active beamforming is constant and identical for all users. Baseline 3 assumes fixed RIS phase shifts, while employing the DDPG agent to optimize the active beamforming matrix and ABS trajectory. We analyze the convergence of the reward function for our proposed scheme and the baselines and evaluate the achievable average system sum data rate and the 5th percentile data rate, which represents the minimum data rate achieved by 95% of the users.

Figure 3a illustrates the average episode rewards over training episodes for our proposed solution and the baseline schemes with 32 antennas, 32 GTs and 40 RIS elements. The models were trained over 100 episodes, with each episode comprising 10,000 time steps N. The results illustrate that the proposed DDPG agent achieves higher rewards compared to any of the baseline schemes. However, it takes longer to converge when compared to the scheme without beamforming and the one without phase shift optimizations. This is attributed to the fact that in these two schemes, a smaller number of elements are being optimized, resulting in a quicker convergence time. The convergence time of the scheme without clustering is similar to the proposed solution with clustering, but it achieves a lower reward.

Figure 3b presents the average system sum rate achieved by the proposed solution and the baseline schemes as a function of the reflecting elements with 32 antennas and 32 GTs. The results illustrate that as the number of RIS elements increases, the average sum rate improves for all schemes. Furthermore, the scheme that does not optimize the RIS phase shifts performs worst, as also observed in the previous result. This emphasizes the importance of the RIS for improving the network sum rate. These results also show the importance of active beamforming over clustering.



(a)

(b)

(c)

Fig. 3: Average episode reward rate over learning episodes (a), average system sum rate rate over the number of RIS elements (b), and average 5-th percentile rate over the number of RIS elements (c) for the proposed scheme and other baselines.

Figure 3c displays the 5th percentile rate as a function of the number of RIS elements with 32 antennas and 32 GTs for the proposed DDPG-based scheme and the considered baselines. The scenario demands a QoS of $R_{min} = 2$ bps/Hz, which represents the minimum data rate target for each user. Both the proposed solution and the no clustering scheme ensure that 95% of the users achieve data rates exceeding R_{min} even with a relatively low number of RIS elements. Without proper beamforming, approximately 40 RIS elements are needed for meeting the 5th percentile rate of 2 bps/Hz, whereas without proper RIS phase shift optimization, 100 RIS elements are necessary to satisfy this threshold. This highlights the effectiveness of the proposed scheme and the tradeoff between phase shift optimization and the number of RIS resources.

VI. Conclusions

In this paper, we investigate the use of DRL to jointly optimize active beamforming, ABS trajectory, and RIS phase shifts in a MU-MISO communication system. Employing user clustering to serve them by the ABS directly or through the RIS is a simple yet effective scheme to improve sum rate performance, of the CSI is known. Most important, however, is the RIS phase shift optimization, followed by active beamforming. While taking longest to converge, the proposed DRL scheme outperforms the simpler solutions. Future work will further analyze the performance of the proposed solution in different scenarios, the complexity-performance tradeoff, and the scalability with ground and aerial base stations.

References

[1] E. Basar *et al.*, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.

[2] A. S. Abdalla *et al.*, "UAVs with reconfigurable intelligent surfaces: Applications, challenges, and opportunities," *arXiv 2012.04775*, 2020.

[3] A. S. Abdalla and V. Marojevic, "Aerial RIS for MU-MISO: Joint Base Station Beamforming and RIS Phase Shifter Optimization," in *2022 IEEE SECON Workshops*, 2022, pp. 19–24.

[4] H. Tataria *et al.*, "6g wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.

[5] Q. Wu *et al.*, "Joint trajectory and communication design for multi-uav enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, 2018.

[6] Z. Wei *et al.*, "Sum-rate maximization for irs-assisted uav ofdma communication systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2530–2550, 2021.

[7] S. Li *et al.*, "Reconfigurable intelligent surface assisted uav communication: Joint trajectory design and passive beamforming," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 716–720, 2020.

[8] M. Misbah *et al.*, "Phase and 3-d placement optimization for rate enhancement in ris-assisted uav networks," *IEEE Wireless Communications Letters*, vol. 12, no. 7, pp. 1135–1138, 2023.

[9] B. Galkin, J. Kibilda, and L. A. DaSilva, "Coverage analysis for low-altitude UAV networks in urban environments," in *2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[10] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, 2019.

[11] F. Jiang and A. L. Swindlehurst, "Optimization of UAV heading for the ground-to-air uplink," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 5, pp. 993–1005, 2012.

[12] A. S. Abdalla and V. Marojevic, "Securing mobile multiuser transmissions with UAVs in the presence of multiple eavesdroppers," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11 011–11 016, 2021.

[13] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[14] D. Silver, *et al.*, "Deterministic policy gradient algorithms," in *31st Int. Conf. Machine Learning*, Bejing, China, 22–24 Jun. 2014, pp. 387–395.

[15] A. S. Abdalla and V. Marojevic, "DDPG learning for aerial RIS-assisted MU-MISO communications," in *2022 IEEE 33nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2022, pp. 1–6.