IUST
Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

DRL-based Joint Beamforming and Power Allocation in Beyond Diagonal Reconfigurable Intelligence Surface 6G Systems

Mousa Abdollahvand*(C.A.) and Sima Sobhi-Givi*

Abstract: This paper introduces a new method for improving wireless communication systems by employing beyond diagonal reconfigurable intelligent surfaces (BD-RIS) and unmanned aerial vehicle (UAV) alongside deep reinforcement learning (DRL) techniques. BD-RIS represents a departure from traditional RIS designs, providing advanced capabilities for manipulating electromagnetic waves to optimize the performance of communication. We propose a DRL-based framework for optimizing the UAV and configuration of BD-RIS elements, including hybrid beamforming, phase shift adjustments, and transmit power coefficients for non-orthogonal multiple access (NOMA) transmission by considering max-min fairness. Through extensive simulations and performance evaluations, we demonstrate that BD-RIS outperforms conventional RIS architectures. Additionally, we analyze the convergence speed and performance trade-offs of different DRL algorithms, emphasizing the importance of selecting the appropriate algorithm and hyper-parameters for specific applications. Our findings underscore the transformative potential of BD-RIS and DRL in enhancing wireless communication systems, laying the groundwork for next-generation network optimization and deployment.

Keywords: Unmanned aerial vehicle (UAV), beyond diagonal-reconfigurable intelligent surface (BD-RIS), Non-orthogonal multiple access (NOMA), hybrid beamforming, reinforcement learning (RL).

In Press

1 Introduction

1.1 Motivation

RECONFIGURABLE intelligent surfaces (RIS) represent a promising technology, and the evolution towards beyond-diagonal RIS (BD-RIS) signifies a ground breaking departure from traditional RIS architectures. BD-RIS introduces innovative opportunities for manipulating electromagnetic waves, endowed with advanced signal processing capabilities. These intelligent surfaces harbor the potential to revolutionize communication systems by augmenting link quality and optimizing data rates, thereby enhancing wireless communication. Conversely, unmanned aerial vehicles (UAVs) have emerged as dynamic and adaptable solutions to tackle the challenges of conventional ground-based networks. Serving as aerial communication platforms, UAVs possess the ability to swiftly navigate diverse terrains, delivering on-demand connectivity and extending coverage to remote or disaster-stricken area [1].

Non-orthogonal multiple access (NOMA) emerges as a pivotal facilitator for boosting data rates in wireless networks. By enabling multiple users to utilize the same time-frequency resource, NOMA initiates a paradigm shift in multiple access schemes, opening avenues for heightened system throughput and enhanced user experience. In NOMA, intra-cell interference can be

Iranian Journal of Electrical & Electronic Engineering, YYYY.
Paper first received DD MONTH YYYY and accepted DD MONTH YYYY.
* The author is with the Department of XXX, YYY University, Address.
E-mail: xxx@xxx.xxx.xxx.
** The authors are with the Department of XXX, YYY University, Address.
E-mails: yyy@yyy.yyy.yyy, zzz@zzz.zzz.zzz.
Corresponding Author: Second B. Author.

mitigated using multi-user detection (MUD) and successive interference cancelation (SIC) techniques [2]. Millimeter-wave (mmWave) communications leverage higher frequency bands to unlock unparalleled data rates, fostering the evolution of ultra-fast and low-latency networks. This technology plays a vital role in addressing the exponentially increasing demand for data-intensive applications and services [2]. In mmWave communications, directional antennas and beamforming significantly enhance transmission quality. Various types of beamforming, including analog, digital, and hybrid beamforming, contribute to this enhancement. Hybrid Beamforming represents a sophisticated signal processing technique that amalgamates the benefits of digital and analog beamforming. This approach augments the flexibility and efficiency of beamforming operations, optimizing signal coverage and quality in intricate communication scenarios. The fusion of NOMA and mmWave transmissions augments the network's capacity [3].

Reinforcement Learning (RL), a subset of artificial intelligence, introduces autonomous decision-making capabilities to wireless communication networks. By enabling systems to learn and adapt in real-time, RL empowers networks to optimize resource allocation, adjust to dynamic environments, and elevate overall system performance [4].

1.2 Related Works

As previously discussed, advancements in RIS architectures have surpassed conventional diagonal phase shift matrices, with recent endeavors aiming to enhance their adaptability in shaping the wireless channel. In [5], the authors tackled the challenge of optimizing the signal-to-noise ratio (SNR) in both single and multiple antenna links with the aid of a group-connected BD-RIS. They addressed the Max-SNR problem by deriving a closed-form solution, relying on the Takagi factorization of a specific complex and symmetric matrix. In [6], scientists investigated optimizing both the transmit precoder and the BD-RIS matrix jointly to enhance the total data transmission rate in a system utilizing RIS technology. They employed Lagrangian dual transform and Quadratic transform techniques to address the optimization problem. Efficient BD-RIS architectures were proposed in [7], utilizing group- and fully-connected reconfigurable impedance networks. The authors provided a closed-form solution for the globally optimal scattering matrix, enabling the attainment of theoretical performance upper bounds across various channel configurations. Furthermore, in [8], a dynamically group-connected BD-RIS was investigated, employing a dynamic grouping strategy to categorize RIS antennas into multiple subsets based on channel state information (CSI). This led to the creation of a permuted block-diagonal scattering matrix, and an efficient algorithm was introduced to enhance the total data transmission rate for multiuser multiple-input single-output (MU-MISO) systems. In [9], an RIS architecture was presented, allowing a signal received at one element to be redirected by another element by adjusting phase shift. The authors tackled the challenge of maximizing attainable rates within this new RIS framework by concurrently optimizing non-diagonal phase shift matrix and active beamforming. This optimization was conducted for both single-user MISO systems and multi-user multiple-input multiple-output (MU-MIMO) systems using alternating optimization techniques and SDR methods, respectively. The study outlined in [10] explores a unified methodology for optimizing BD-RIS configurations under various non-convex constraints. The authors introduced an approach capable of addressing energy efficiency maximization problem in BD-RIS system considering multiple QoS constraints. Additionally, their methodology efficiently tackles the challenge of maximizing sum-rate in BD-RIS assisted systems. Moreover, the study highlights the notable advantages of BD-RIS compared to conventional diagonal RIS (D-RIS) configurations.

In [11], researchers investigated an innovative wireless powered communication network enabled by RIS and UAVs. The primary goal was to improve the minimum throughput for all ground users by concurrently optimizing several factors, including the user transmit power, horizontal placement of UAVs, passive beamforming vectors at the RIS, and transmission time allocation. To tackle this intricate optimization problem, the researchers introduced an algorithm that decomposes the issue into sequentially solved four subproblems. Their findings highlighted the superiority of the RIS-assisted UAV-enabled network over traditional networks in terms of enhancing minimum throughput. In another study, discussed in [12], the focus was on a UAV-enabled wireless system where hybrid active-passive RIS aided the communication between UAV and users. The objective revolved around achieving fairness in the system by maximizing the minimum rate of users. To achieve this, the researchers conducted jointly optimizing the transmit beamformer, UAV’s trajectory, and RIS coefficients. To address these complex optimization challenges, efficient algorithms based on block coordinate ascent and successive convex approximation (SCA) were devised by the researchers, enabling effective problem-solving in an iterative manner.

In [13], researchers tackled the joint optimization challenge of power allocation and hybrid beamforming to enhance the minimum user signal-to-leakage-and-

IEEE

noise ratio (SLNR), with the aim of balancing computational complexity and ensuring fair treatment of users. Meanwhile, in [14], the emphasis was placed on the coordinated optimization of power allocation and beamforming within a multi-cell multiuser MIMO-NOMA network. The objective was to maximize the total data transmission rate for users while ensuring that their target rates are preserved. The authors proposed an iterative sub-optimal algorithm based on SCA to synchronize base stations (BSs) and address the optimization task. The study discussed in [15] investigated the max-min fairness problem in a downlink (DL) MIMO mmWave-NOMA system, incorporating user clustering, power allocation, and hybrid beamforming. Here, the SLNR metric served as a focal point, guiding the development of user clustering strategies, hybrid beamforming matrices, and power allocation schemes. Similarly, [16] explored power allocation, hybrid precoding, and user clustering for massive MIMO in the mmWave-NOMA setup. The paper introduced a clustering technique to designate initial cluster heads and iteratively incorporate users into clusters while managing intrabeam interference. Furthermore, it transformed the non-convex data rate maximization problem into a convex inter-cluster problem to address hybrid precoding and power allocation.

In [17], hybrid beamforming techniques for UAV-assisted communications employing massive MIMO systems were investigated. The authors derived an approximate closed-form expression for data rate through the application of hybrid beamforming methodologies and formulated a power allocation strategy tailored for line-of-sight (LoS) channels. Furthermore, the study explored optimal UAV location designs. In [18], researchers studied a mmWave DL communication system assisted by RIS, with hybrid beamforming implemented at the BS. They tackled a power minimization challenge by concurrently optimizing the response matrix at the RIS and the hybrid beamforming at the BS, while ensuring all users signal-to-interference-plus-noise ratio (SINR) constraints. Manifold optimization techniques were utilized to manage the non-convex constraints. Furthermore, the research explored the interplay between the max-min fairness problem and power minimization, underscoring the pivotal contribution of the RIS in decreasing power consumption across the system.

1.3 Contributions

The main objective of this study is to examine the overall performance of the system with a particular focus on the contributions of BD-RIS and UAVs. This integration aims to jointly optimize phase shift matrix at the BD-RIS, hybrid beamforming at the UAV, and NOMA power gains in a mmWave-NOMA system, with the overarching goal of max-min rate fairness as a α-fair utility function. To tackle this optimization problem, we leverage deep reinforcement learning (DRL) techniques. We demonstrate that BD-RIS and fairness optimization can substantially enhance system performance and elucidate the specific contributions of BD-RIS to these improvements. Therefore, the main contributions of this paper can be categorized as follows:

- We propose a novel system architecture that combines BD-RIS with UAVs in mmWave-NOMA systems. This innovative integration aims to leverage the unique capabilities of BD-RIS and UAVs to enhance the efficiency and performance of wireless communication networks.
- In practical scenarios, NOMA transmission may suffer from imperfect SIC. We incorporate this aspect into our system model to accurately reflect real-world conditions, ensuring the robustness and reliability of our proposed solution.
- We propose an optimization problem that jointly designs the BD-RIS phase shift matrix, UAV hybrid beamforming, and power allocation. Our objective is to maximize the max-min rate fairness, which is represented by an α-fair utility function. This comprehensive optimization framework accounts for the diverse characteristics and requirements of modern wireless communication systems.
- To tackle the complex optimization problem, we employ DRL algorithms. These algorithms offer a powerful and adaptive approach to optimizing system parameters in dynamic and uncertain environments, enabling efficient and effective decision-making.
- Through extensive simulations, we showcase that our proposed BD-RIS architecture combined with DRL algorithms outperforms conventional D-RIS architectures with DRL in terms of max-min rate fairness.
- We conduct a thorough analysis to investigate the impact of varying the number of UAV transmit antennas on the performance of the DRL algorithm. This analysis provides valuable insights into the scalability, adaptability, and robustness of our proposed solution, offering guidance for practical deployment and optimization of mmWave-NOMA systems.

The remainder of this paper is structured as follows: Section II delineates the proposed mmWave-NOMA transmission system model. Section III elucidates the proposed design of RIS phase shift matrix, hybrid beamforming, and power allocation to users. Numerical

IEEE Proof

results are presented Section IV, and lastly, Section V offers concluding remarks.

2 System Description

As illustrated in Fig. 1, we examine a communication system at mmWave frequencies using NOMA with the assistance of a RIS and an UAV. The unmanned aerial vehicle (UAV) is outfitted with M antennas, catering to K single-antenna user equipment's (UEs), and engages in collaboration with an RIS consisting N elements. The RIS is forming an $N = N_x \times N_z$ uniform rectangular array (URA) and includes a controller capable of intelligently adjusting the phase shift of each element [15]. Employing a hybrid beamforming structure with N_{RF} radio frequency (RF) chains (where $N_{RF} < M$), the UAV transmits N_s independent data streams to the UE terminals. To maximize multiplexing gain, we assume the data streams undergo initial precoding through digital beamforming denoted as $\mathbf{D} \in \mathbb{C}^{N_{RF} \times N_s}$ in the baseband. Following the relevant RF processing, the digitally processed signal traverses M phase shifters with a constant modulus for analog beamforming, utilizing the beamforming matrix $\mathbf{A} \in \mathbb{C}^{M \times N_{RF}}$. Consequently, the hybrid beamforming matrix is expressed as $\mathbf{W} = \mathbf{AD} = \left[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{N_{RF}}\right]$, with each column having a unit norm, i.e., $\| \mathbf{w}_n \| = 1, 1 \le n \le N_{RF}$. Additionally, the locations of the UEs, RIS, and UAV are denoted by $\mathbf{U}_i = [x_i, y_i, z_i]$, $\mathbf{R} = [x_r, y_r, z_r]$, and $\mathbf{Q} = [x_q, y_q, z_q]$, respectively.

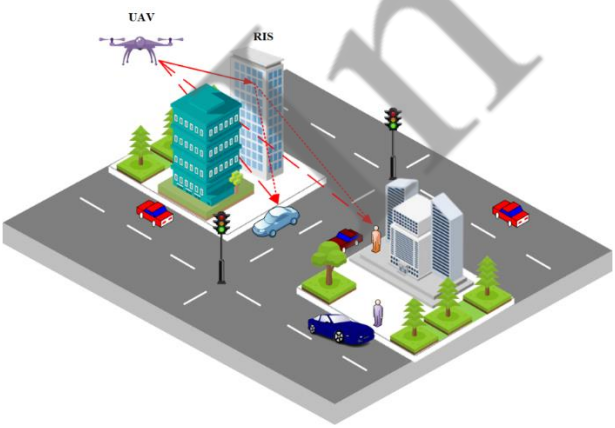

Fig. 1. System model for UAV-assisted BS using NOMA and hybrid beamforming

In traditional RIS configurations, each RIS element functions independently. Put differently, the incoming signal aimed at the i -th element is exclusively reflected by that same i -th element, undergoing a phase shift adjustment. However, in our approach, we assume the establishment of interconnections among RIS elements, enabling the signal arriving at the i -th element to be redirected by other elements. This interconnected arrangement provides greater flexibility in crafting the RIS phase shift matrix, offering enhanced configurability that can potentially result in improved performance. Our research presents an innovative RIS scheme where the signal received by the i -th element can be redirected to another element, indicated as the $i^{'}$ -th element, after undergoing a phase shift adjustment, as detailed below [20]:

$$\mathbf{\Theta} = \begin{bmatrix} a_{1,1}e^{j\theta_{1,1}} & a_{1,2}e^{j\theta_{1,2}} & \cdots & a_{1,N}e^{j\theta_{1,N}} \\ a_{2,1}e^{j\theta_{2,1}} & a_{2,2}e^{j\theta_{2,2}} & \cdots & a_{2,N}e^{j\theta_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}e^{j\theta_{N,1}} & a_{N,2}e^{j\theta_{N,2}} & \cdots & a_{N,N}e^{j\theta_{N,N}} \end{bmatrix} \tag{1}$$

Here, the phase shift matrix assumes a non-diagonal structure. It's crucial to highlight that the RIS phase shift matrix comprises just N non-zero elements. In the context of NOMA transmission, K UEs are categorized into N_{RF} clusters (for simplicity in Fig. 1, we show one cluster), where each cluster corresponds to an independent data stream. Within each cluster, UEs can leverage NOMA and successive interference cancellation (SIC) techniques to address intra-cluster interference, while inter-cluster interference is alleviated through beamforming. As N_{RF} RF chains can accommodate up to N_{RF} data streams, ensuring at least one UE in each cluster is necessary to prevent idle RF resources. The proposed system operates in a DL mode, with the UAV transmitting to the UEs through BD-RIS.

2.1 Channel Model

We adopt a Rician fading channel model for all communication links in our analysis. Consequently, the channel between the UAV-RIS, UAV-UE (u_k), and RIS-UE (u_k) are respectively denoted by $\mathbf{G} \in \mathbb{C}^{N \times M}$, $\mathbf{h}_U^k \in \mathbb{C}^{1 \times M}$ and $\mathbf{h}_R^k \in \mathbb{C}^{N \times 1}$, which can be mathematically expressed as [19]:

$$\mathbf{G} = \sqrt{\frac{\rho_0}{\| \mathbf{Q} - \mathbf{R} \|^2}}\overline{\mathbf{G}}, \tag{2}$$

$$\overline{\mathbf{G}} = \mathbf{g}_y \otimes \mathbf{g}_x, \tag{3}$$

where $\otimes$ is the tensor multiplication operand and

In Press

$$\mathbf{g}_x = \left[1, e^{-j\frac{2\pi}{\lambda}d\cos\phi\sin\psi}, \ldots, e^{-j\frac{2\pi}{\lambda}(N_x-1)d\cos\phi\sin\psi}\right]^T,$$

$$\mathbf{g}_y = \left[1, e^{-j\frac{2\pi}{\lambda}d\sin\phi\sin\psi}, \ldots, e^{-j\frac{2\pi}{\lambda}(N_z-1)d\sin\phi\sin\psi}\right]^T,$$

$$\cos\phi\sin\psi = \frac{x_q - x_r}{\| \mathbf{Q}-\mathbf{R} \|},$$

$$\sin\phi\sin\psi = \frac{z_q - z_r}{\| \mathbf{Q}-\mathbf{R} \|}.$$

$$\mathbf{h}_U^k = \sqrt{\frac{\rho_0}{\| \mathbf{Q}-\mathbf{U}_k \|^{\alpha_U}}}\left(\sqrt{\frac{\beta_U}{1+\beta_U}}\bar{\mathbf{h}}_U^k + \sqrt{\frac{1}{1+\beta_U}}\hat{\mathbf{h}}_U^k\right), \quad (4)$$

$$\mathbf{h}_R^k = \sqrt{\frac{\rho_0}{\| \mathbf{R}-\mathbf{U}_k \|^{\alpha_R}}}\left(\sqrt{\frac{\beta_R}{1+\beta_R}}\bar{\mathbf{h}}_R^k + \sqrt{\frac{1}{1+\beta_R}}\hat{\mathbf{h}}_R^k\right), \quad (5)$$

where ρ_0 denotes the path loss at the reference distance of one meter, ϕ and ψ represent the azimuth and elevation angles of the LoS component, respectively, d is the antenna separation, and λ is the carrier wavelength. α_U and α_R represent the path loss exponents for the UAV-UE and RIS-UE links, respectively. Additionally, β_U and α_R indicate the Ricean factors, while $\bar{\mathbf{G}}$, $\bar{\mathbf{h}}_U^k$, and $\bar{\mathbf{h}}_R^k$ signify the deterministic LoS components and defined similar to (3) for $\bar{\mathbf{h}}_U^k$, and $\bar{\mathbf{h}}_R^k$. The random Rayleigh distributed non-LoS (NLoS) components are represented by $\hat{\mathbf{h}}_U^k$, $\hat{\mathbf{h}}_R^k$.

Based on the previously discussed channel models, the effective channel power gain between the UAV and the k-th UE with the assistance of the RIS, denoted as $\mathbf{H}_k \in \mathbb{C}^{1\times M}$, is expressed as $\mathbf{H}_k = \left|(\mathbf{h}_U^k + (\mathbf{h}_R^k)^H \Theta\mathbf{G})\mathbf{W}\right|^2$, where $(.)^H$ is the Hermitian.

2.2 Signal to Interference and Noise Ratio (SINR)

Here, we assume that each UAV employs NOMA to provide communication service for UEs. Therefore, two UEs are grouped to perform NOMA transmission. Without loss of generality, we assume that $\mathbf{H}_i \geq \mathbf{H}_j$. In the NOMA scheme, SIC is applied at the receiver, allowing the UE with a higher channel gain to eliminate or reduce the interference from UEs with lower channel gains. Hence, the SINR of j-th UE at the i-th UE ($\gamma_j^{j\to i}$), can be expressed as:

$$\gamma_j^{j\to i} = \frac{\beta_j P\left|(\mathbf{h}_U^i + (\mathbf{h}_R^i)^H \Theta\mathbf{G})\mathbf{W}\right|^2}{\beta_i P\left|(\mathbf{h}_U^i + (\mathbf{h}_R^i)^H \Theta\mathbf{G})\mathbf{W}\right|^2 + \sigma^2}, \quad (6)$$

where β_i is the transmit power coefficient for the i-th UE and $\sum_{i=1}^{K}\beta_i = 1$. The SINR of i-th UE can be expressed as:

$$\gamma_i = \frac{\beta_i P\left|(\mathbf{h}_U^i + (\mathbf{h}_R^i)^H \Theta\mathbf{G})\mathbf{W}\right|^2}{\xi\beta_j P\left|(\mathbf{h}_U^i + (\mathbf{h}_R^i)^H \Theta\mathbf{G})\mathbf{W}\right|^2 + \sigma^2}, \quad (7)$$

where ξ represents the imperfect SIC coefficient, e.g., $\xi = 1$ for no SIC and $\xi = 0$ for perfect SIC. Also, the SINR of j-th UE can be written as [2, 21]:

$$\gamma_j = \frac{\beta_j P\left|(\mathbf{h}_U^j + (\mathbf{h}_R^j)^H \Theta\mathbf{G})\mathbf{W}\right|^2}{\beta_i P\left|(\mathbf{h}_U^j + (\mathbf{h}_R^j)^H \Theta\mathbf{G})\mathbf{W}\right|^2 + \sigma^2}, \quad (8)$$

2.3 Date Rate and Fairness Function

Data rate refers to the information rate that can be transmitted over a given bandwidth. Hence, the data rate of k-th UE (R^k) is obtained as:

$$R^k = \log_2\left(1+\gamma_k\right),\ k \in \{i, j\}. \quad (9)$$

Therefore, the sum-rate of users (R) can be obtained as follows:

$$R = \sum_{k=1}^{K} R^k = \sum_{k=1}^{K}\log_2\left(1+\gamma_k\right). \quad (10)$$

In wireless communication, users with better channel conditions typically achieve higher data-rate compared to those with poorer channels, leading to unfairness among users. To address this issue, we propose defining utility function that incorporates various levels of fairness. In this paper, we consider max-min fairness (η_α) as a α-fair utility function as follows [22]:

$$\eta_\alpha(R) = \frac{R^{1-\alpha}}{1-\alpha}, \qquad \forall\alpha \neq 1, \alpha \geq 0. \quad (11)$$

where α refers to the different levels of rate-fairness.

2.4 Problem Formulation

In the system under consideration, the UAV aims to maximize the utility function which is defined as a max-min fairness in (11). To this end, we need to optimally design the hybrid beamforming matrix ($\mathbf{W}$) at the UAV, passive beamforming matrix ($\mathbf{\Theta}$) at the RIS, and power allocation coefficients (β_k). Therefore, the optimization problem is formulated as:

In Press

$$\max_{\mathbf{w},\boldsymbol{\Theta},\beta_k} \eta_\alpha(R), \tag{12}$$

subject to:

$$\mathbf{S}_1 : \gamma_k > \gamma_{th}, \ \ k \in \{i, j\},$$

$$\mathbf{S}_2 : \gamma_j^{i \to i} > \gamma_j,$$

$$\mathbf{S}_3 : \sum_{k=1}^{K} \beta_k = 1,$$

$$\mathbf{S}_4 : a_{i,i'} \in \{0,1\}, \ \ \vee \ i, i' = 1, \ldots, N$$

$$\mathbf{S}_5 : \theta_{i,i'} \in [0, 2\pi), \ \ \vee \ i, i' = 1, \ldots, N$$

$$\mathbf{S}_6 : \left\| \mathbf{w}_n \right\|_F^2 = N_s, \ \ \vee \ n,$$

$$\mathbf{S}_7 : \left| A \right| = \frac{1}{\sqrt{N}},$$

According to the constraint $\mathbf{S}_1$ each UE SINR must be higher than the predefined threshold value (γ_{th}) to satisfy the minimum QoS requirements. The constraint $\mathbf{S}_2$ ensures successful SIC performance. Sum of transmission power gains of NOMA clusters must be one according to constraint $\mathbf{S}_3$. The constraint $\mathbf{S}_4$ and $\mathbf{S}_5$ denotes the phase shift constraints of each RIS sub-surface. Finally, constraints $\mathbf{S}_6$ and $\mathbf{S}_7$ are the restriction of the power for hybrid beamforming matrix and restriction of constant modulus in the analog domain, respectively.

2.5 Proposed Solution

In this context, in this scenario, we present a method to ascertain the hybrid beamforming matrix at the UAV, the phase shift matrix at the BD-RIS, and the power gain coefficients for NOMA, as depicted in equation (10). Given the non-convex characteristics of the objective functions involved, traditional approaches designed for convex problems prove inadequate. Hence, we advocate for the application of DRL algorithms, incorporating the following parameters:

Agent: The UAV operates as the agent in our system setup. It takes actions based on its observations and aims to optimize its performance over time.

Action: The action ($\mathbf{a}_t$) refers to the set of feasible choices available to the UAV regarding hybrid beamforming configuration, RIS phase shift matrix adjustments, and power gain coefficients for NOMA transmission.

State: The state at time t ($\mathbf{s}_t$) represents the current status of the system, specifically the SINR of the most recent communication link.

Reward: The reward function (r_t) is a pivotal element of the DRL framework, serving as feedback for the agent's actions. In our proposed framework, the reward function is formulated as a weighted sum of the objective function and various constraints. It guides the agent's learning process by providing a quantitative measure of performance, incentivizing actions that lead to improved system behavior while penalizing deviations from desired outcomes.

Soft Actor-Critic (SAC) Method

The soft actor-critic (SAC) algorithm represents a online, off-policy, model-free actor-critic RL technique. It seeks to compute an optimal policy that not only maximizes the long-term expected reward but also increases the entropy of the policy. Entropy here serves as a gauge of policy uncertainty given a certain state, with higher entropy values encouraging more exploration. By simultaneously maximizing both the expected cumulative long-term reward and policy entropy, SAC strikes a balance between exploiting known strategies and exploring new possibilities within the environment.

SAC differs from its predecessors in its approach to policy optimization, blending stochastic policy optimization with elements of deep deterministic policy gradient (DDPG)-style methods. One key aspect of SAC is its incorporation of entropy regularization. The policy is trained to navigate a trade-off between expected return and entropy, where entropy represents the level of randomness in the policy. This linkage to the exploration-exploitation trade-off implies that higher entropy fosters greater exploration, thus potentially hastening learning processes. Moreover, it helps prevent the policy from prematurely converging to suboptimal solutions. The pseudo-code for the SAC method can be found in Algorithm (1).

Softmax Deep Double Deterministic Policy Gradients (SD3) Algorithm

SD3 represents a continuous control DRL algorithm that updates both the optimal $\mathcal{Q}$-value and policy functions iteratively using the actor-critic method [19]. It utilizes neural networks, specifically the critic and actor networks, to approximate the policy function $\pi(s)$ and $\mathcal{Q}$-value function $\mathcal{Q}(s, a)$, respectively. The actor network is responsible for selecting actions, while the critic network estimates the $\mathcal{Q}$-value to guide the actor toward maximizing them. To address continuous action spaces and minimize discretization errors, SD3 integrates techniques from deep $\mathcal{Q}$-networks (DQN) and double $\mathcal{Q}$-learning. The methodology incorporates dual-actor and dual-critic approaches, along with target network techniques.

In Press

Algorithm 1: Pseudo-code of SAC method

Initial $\mathcal{Q}$-function parameters ϕ_1, ϕ_2, policy parameters θ, empty replay buffer $\mathcal{D}$
Set main parameters into target parameters $\phi_{targ,1} \leftarrow \phi_1$, $\phi_{targ,2} \leftarrow \phi_2$
repeat
  Select action $a \sim \pi_\theta(.|s)$ by observing state s
  Execute a in the environment
  Observe reward r, next state s', and done signal d
  Store a tuple (s, a, r, s', d) in replay buffer $\mathcal{D}$
  If s' is terminal, reset environment state.
  if it's time to update **then**
    for j in range (update) **do**
      Sample a batch of transitions, randomly, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
      Compute targets:
      $$y(r, s', d) = r + \gamma(1-d)\Big(\min_{i=1,2} \mathcal{Q}_{\phi_i}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}'|s')\Big),\ \tilde{a} \sim \pi_\theta(.|s')$$
      Update $\mathcal{Q}$-functions:
      $$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \Big(\mathcal{Q}_{\phi_i}(s,a) - y(r,s',d)\Big)^2$$
      Update policy:
      $$\nabla_\theta \frac{1}{|B|} \sum_{s \in B} \Big(\min_{i=1,2} \mathcal{Q}_{\phi_i}(s, \tilde{a}_\theta(s)) - \alpha \log \pi_\theta(\tilde{a}_\theta(s)|s)\Big)$$
      , where $\tilde{a}_\theta(s)$ is a sample from $\pi_\theta(.|s)$.
      Update target networks:
      $\phi_{targ,i} \leftarrow \rho\phi_{targ,i} + (1-\rho)\phi_i$ for $i = 1, 2$
    end for
  end if
until Convergence

The SD3 learning framework employs eight neural networks, with only the critic and actor networks undergoing training. The target networks are updated through soft copying. SD3 utilizes experience replay for training, where interaction samples are stored in an experience buffer and randomly sampled mini-batches are used for network training. This enhances sample utilization and reduces sample correlation. In SD3, the critic network employs clipped double $\mathcal{Q}$-learning integrated with the Boltzmann softmax operator to approximate $\mathcal{Q}$-values and create temporal difference (TD) errors [23]. The pseudo-code for the SD3 method is provided in Algorithm (2).

Algorithm 2: Pseudo-code of SD3

Input: Initial critic networks $\mathcal{Q}_1$, $\mathcal{Q}_2$, and actor networks π_1, π_2 with random parameters θ_1, θ_2, ϕ_1, ϕ_2
Input: Initial target networks $\overline{\theta}_1 \leftarrow \theta_1$, $\overline{\theta}_2 \leftarrow \theta_2$, $\overline{\phi}_1 \leftarrow \phi_1$, $\overline{\phi}_2 \leftarrow \phi_2$
Input: Initialize replay buffer $\mathcal{D}$
for $t = 1, \ldots, T$ **do**
  Based on π_1 and π_2, select action a with exploration noise $\epsilon \sim \mathcal{N}(0, \sigma)$
  Observing reward r and new state s' by executing action a, and done d
  Store transition tuple (s, a, r, s', d) in $\mathcal{D}//d$
  for $i = 1, 2$ **do**
    Sample a mini-batch of N transitions $\{(s, a, r, s', d)\}$ from $\mathcal{D}$
    Sample K noises $\epsilon \sim \mathcal{N}(0, \overline{\sigma})$
    $\hat{a}' \leftarrow \pi_i(s'; \overline{\phi}_i) + \text{clip}(\epsilon, -c, c)$
    $\hat{\mathcal{Q}}(s', \hat{a}') \leftarrow \min_{j=1,2}(\mathcal{Q}_j(s', \hat{a}'; \overline{\theta}_j))$
    Compute:
    $$\text{softmax}_\beta\Big(\hat{\mathcal{Q}}(s', .)\Big) \leftarrow \mathbb{E}_{\hat{a}' \sim p}\left[\frac{e^{(\beta\hat{\mathcal{Q}}(s',\hat{a}'))}\hat{\mathcal{Q}}(s',\hat{a}')}{p(\hat{a}')}\right] / \mathbb{E}_{\hat{a}' \sim p}\left[\frac{e^{(\beta\hat{\mathcal{Q}}(s',\hat{a}'))}}{p(\hat{a}')}\right]$$
    $y_i \leftarrow r + \gamma(1-d)\,\text{softmax}_\beta\Big(\hat{\mathcal{Q}}(s', .)\Big)$
    Update the critic θ_i:
    $$\frac{1}{N}\sum_s \Big(\mathcal{Q}_i(s, a; \theta_i) - y_i\Big)^2$$
    Update actor ϕ_i:
    $$\frac{1}{N}\sum_s \Big[\nabla_{\phi_i}(\pi(s; \phi_i))\nabla_a \mathcal{Q}_i(s, a; \theta_i)|_{a=\pi(s;\phi_i)}\Big]$$
    Update target networks: $\overline{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\overline{\theta}_i$, $\overline{\phi}_i \leftarrow \tau\phi_i + (1-\tau)\overline{\phi}_i$
  end for
end for

2.6 Complexity Analysis

The computational complexity of DRL algorithms can vary significantly based on several factors, including the intricacy of the environment, the size of the state and

In Press

action spaces, and the underlying algorithm. In our proposed method, which utilizes the DDPG, TD3, and SAC algorithms, the complexity order differs among them. DDPG generally has a lower computational complexity compared to TD3 and SAC. This is because TD3 and SAC incorporate additional components that increase their computational cost. For example, TD3 employs two critic networks and takes the minimum value between them to reduce overestimation bias, adding complexity to the learning process. Similarly, SAC, with its entropy-regularized objective, introduces further computational overhead, particularly in environments with large state-action spaces or more complex dynamics. Despite the higher complexity of TD3 and SAC, there are techniques that can mitigate their computational demands. For instance, experience replay allows for more efficient learning by reusing past experiences, and prioritized experience replay further optimizes this process by focusing on more valuable experiences. Additionally, parallelization strategies can distribute the computation across multiple processors or GPUs, helping to alleviate some of the computational load.

It is important to note that the complexity order mentioned here serves as a rough estimate of computational cost. The actual execution time can be influenced by a variety of factors, such as the hardware being used (e.g., CPUs vs. GPUs), implementation optimizations, and other specifics related to how the algorithm is applied. Therefore, while DDPG may typically have a lower complexity, real-world performance can vary depending on these additional considerations.

3 Performance Evaluation

In this section, we assess the effectiveness of the algorithms we have put forward for beamforming design at the UAV, BD-RIS phase shift matrix, and power allocation. The value of the parameters used in the simulation are listed in Table I.

Table 1 Simulation Parameters

Parameter	Value
Carrier frequency	28 GHz
Bandwidth	100 MHz
Max transmission power	30, 35, 40 dBm
Number of UAV transmit antennas	8, 16, 32, 64
Number of RIS elements	16
Number of RF chains	3
Number of data streams	3
Number of clusters	3
Number of UEs	6
Noise power	-174 dBm/Hz
Minimum rate	2 bps/Hz

3.1 Sum Rate

Fig. 2 illustrates the sum rate of the proposed network across various maximum transmit power settings. It's evident that as the transmit power increases, the sum rate of clusters also rises. This escalation is attributed to the heightened SINR experienced by each user. Additionally, we conducted a comparative analysis of different DRL methods. Our findings indicate that the SAC algorithm consistently outperforms the TD3 and DDPG algorithms in terms of network performance.

Fig. 3 depicts the relationship between the sum rate and the number of UAV antennas across various DRL methods. We observe a notable trend: as the number of UAV antennas increases, the sum rate also rises. This phenomenon occurs due to the enhanced spatial diversity and multiplexing gain enabled by the increased number of antennas. With more antennas, the UAV can employ sophisticated beamforming techniques to optimize signal transmission and reception. As a result, the overall sum rate of the system improves. Furthermore, the comparison among different DRL methods reveals insights into their efficacy in optimizing the system performance.

3.2 Rate-Fairness Utility Function

Fig. 4 illustrates the cumulative distribution function (CDF) of the fairness function across various values of α. When α decreases, resources are distributed more evenly among users, resulting in an increase in the fairness function.

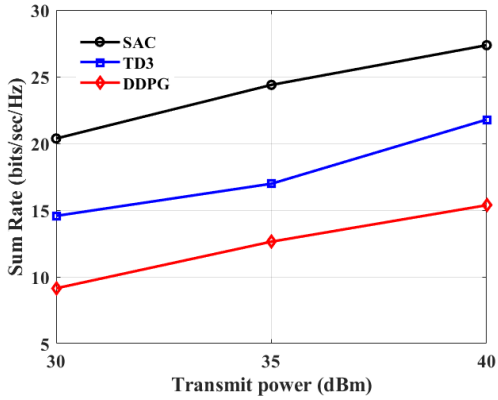

Fig. 2. Sum rate of the proposed network versus different levels of maximum transmit power

In Press

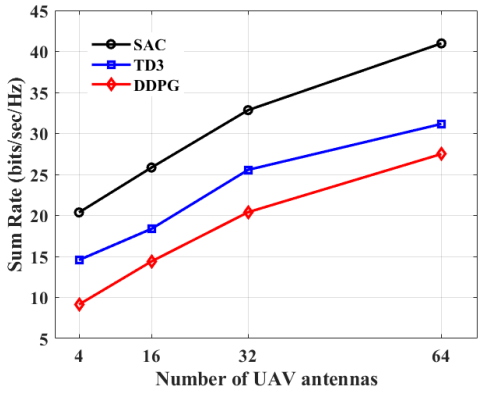

Fig. 3. Sum rate of the proposed network versus different number of UAV antennas.

This indicates that as α decreases, the system prioritizes equitable resource allocation, aiming to ensure that each user receives a fair share of resources. However, it's important to note that there is a trade-off involved: while reducing α enhances fairness among users, it may also lead to a decrease in the overall system performance or efficiency. Thus, selecting an appropriate value for α involves balancing the need for fairness with the desire to optimize system performance.

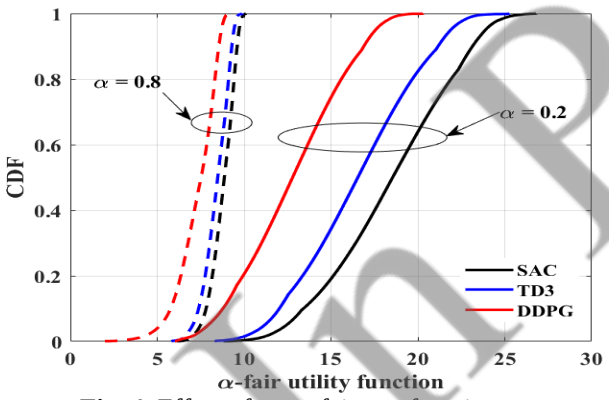

Fig. 4. Effect of α on fairness function

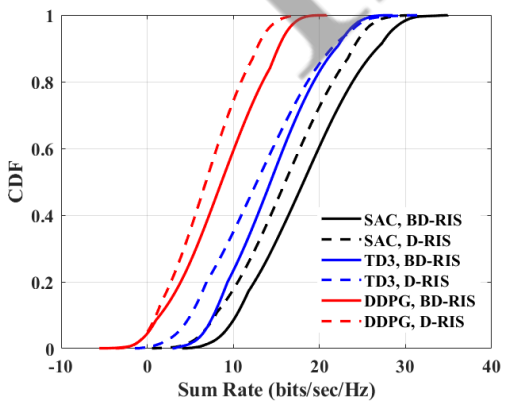

Fig. 5. Sum rate comparison of D-RIS with BD-RIS

3.3 Performance Comparison

Comparison of BD-RIS and D-RIS performance

Here, we evaluate how the performance of the proposed BD-RIS system compares to that of the traditional D-RIS architecture. Our findings clearly indicate that the BD-RIS system outperforms its D-RIS counterpart, which is shown in Fig. 5. The superiority of the BD-RIS system can be attributed to several factors. Firstly, BD-RIS offers enhanced adaptability and flexibility in manipulating electromagnetic waves compared to D-RIS. By enabling interconnections among RIS elements and facilitating dynamic phase shift adjustments, BD-RIS optimally shapes the wireless channel, resulting in improved signal quality and higher data rates. Moreover, BD-RIS's ability to exploit spatial multiplexing and diversity gains allows for more efficient utilization of available resources, leading to enhanced system performance. Additionally, BD-RIS's capacity to mitigate interference and optimize spectrum utilization further contributes to its superior performance compared to D-RIS. Overall, the comparison underscores the significant advantages of BD-RIS over traditional D-RIS architectures, highlighting its potential to revolutionize wireless communication systems by offering higher throughput, improved reliability, and enhanced spectral efficiency.

In Fig. 6, we present a comparison of the bit error rate (BER) for the proposed network across varying numbers of RIS elements. The results clearly demonstrate that increasing the number of RIS elements leads to a reduction in BER. This is because having more RIS elements enhances the ability of the system to intelligently manipulate the signal propagation environment, improving signal quality and reducing transmission errors. Additionally, it is observed that BD-RIS outperforms traditional D-RIS in terms of BER, showing consistently lower error rates. This is likely due to the more advanced design and functionality of BD-RIS, which offers greater flexibility in controlling the reflection properties of the signal.

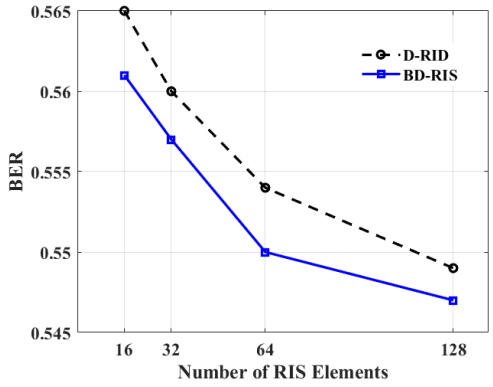

Fig. 6. BER comparison of D-RIS with BD-RIS

In Press

Comparison of Different Multiple Access methods

In this section, we compare the performance of the proposed NOMA transmission with two other multiple access schemes: space division multiple access (SDMA) and orthogonal multiple access (OMA). Fig. 7 illustrates the sum-rate achieved by various DRL algorithms across these different multiple access methods.

The results clearly show that NOMA outperforms both SDMA and OMA in terms of sum-rate. This is because NOMA allows multiple users to share the same frequency resources by superimposing their signals with different power levels, enabling more efficient spectrum utilization. In contrast, OMA allocates distinct time or frequency resources to different users, leading to more rigid resource usage. SDMA, on the other hand, separates users by their spatial location, which can be effective in certain conditions but may not fully exploit the available spectrum in more complex environments. The superior performance of NOMA can be attributed to its ability to accommodate a higher number of users on the same frequency band while maintaining a higher overall data rate, especially when combined with advanced DRL algorithms. The learning-based optimization provided by the DRL algorithms further enhances the sum-rate by dynamically adapting the power allocation and other transmission parameters in NOMA, making it a more efficient solution compared to SDMA and OMA.

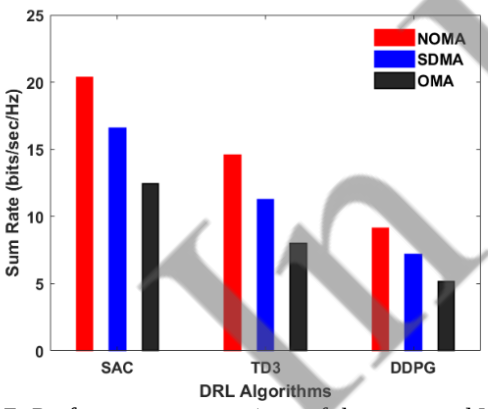

Fig. 7: Performance comparison of the proposed NOMA scheme, SDMA, and OMA

3.4 Convergence Analysis

In Fig. 8, the convergence speed of various DRL algorithms is depicted. Notably, DDPG demonstrates faster convergence compared to TD3 and SAC methods. However, as highlighted in the preceding subsection, SAC achieves a higher sum rate despite its slower convergence speed. This observation underscores a fundamental trade-off between convergence speed and performance in DRL. In essence, optimizing for superior performance may necessitate sacrificing some convergence speed, and vice versa. This trade-off is inherent in DRL methodologies. Therefore, selecting the appropriate algorithm and hyper-parameters for a given problem is crucial, depending on the specific requirements and objectives at hand. Striking the right balance between convergence speed and performance ensures the effectiveness and efficiency of the DRL approach in addressing real-world challenges.

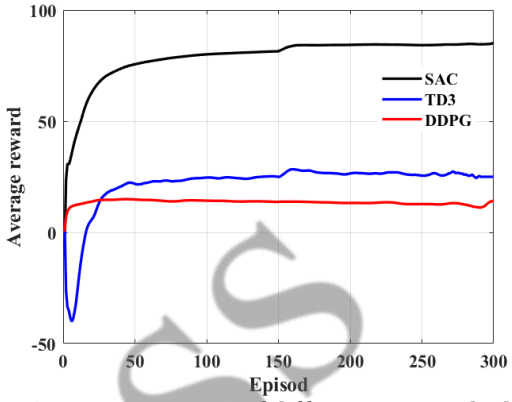

Fig. 8: Convergence of different DRL methods

4 Conclusion

This paper has presented a comprehensive investigation into the utilization of BD-RIS and UAV in wireless communication systems, coupled with DRL techniques. We defined an optimization problem for performance improvement as a α-fair utility function. Our proposed DRL-based framework offers a powerful means to optimize the dynamic phase shift adjustments, enabling efficient hybrid beamforming, and effective power allocation for NOMA transmission. Through extensive simulations and performance evaluations, we have showcased the superiority of BD-RIS over conventional D-RIS architectures, emphasizing its transformative potential in revolutionizing wireless communication systems. Furthermore, our analysis of various DRL algorithms has shed light on the trade-offs between convergence speed and performance, highlighting the importance of selecting the appropriate algorithm and hyper-parameters based on specific application requirements.

Acknowledgment

This work has been financially supported by the research deputy of University of Mohaghegh Ardabili. The grant number was 25888.

References

[1] T. Fang and Y. Mao, A low-complexity beamforming design for beyond-diagonal ris aided multi-user networks, *IEEE Communications Letters*, vol. 28, no. 1, pp. 203–207, 2024.

In Press

[2] S. Sobhi-Givi, M. G. Shayesteh, and H. Kalbkhani, Energy-efficient power allocation and user selection for mmwave-noma transmission in m2m communications underlaying cellular heterogeneous networks, *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9866–9881, 2020.

[3] Y. Xiu, J. Zhao, W. Sun, M. D. Renzo, G. Gui, Z. Zhang, and N. Wei, Reconfigurable intelligent surfaces aided mmwave noma: Joint power allocation, phase shifts, and hybrid beamforming optimization, *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 8393–8409, 2021.

[4] S. Sobhi-Givi, M. Nouri, M. G. Shayesteh, H. Kalbkhani, and Z. Ding, Reinforcement learning based joint resource allocation and user fairness optimization in mmwave-noma hetnets, in *2023 31st International Conference on Electrical Engineering (ICEE)*, 2023, pp. 781–786.

[5] I. Santamaria, M. Soleymani, E. Jorswieck, and J. Gutierrez, Snr maximization in beyond diagonal ris-assisted single and multiple antenna links, *IEEE Signal Processing Letters*, 2023.

[6] H. Li, S. Shen, and B. Clerckx, Beyond diagonal reconfigurable intelligent surfaces: From transmitting and reflecting modes to single-, group-, and fully-connected architectures, *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2311–2324, 2022.

[7] M. Nerini, S. Shen, and B. Clerckx, Closed-form global optimization of beyond diagonal reconfigurable intelligent surfaces, *IEEE Transactions on Wireless Communications*, 2023.

[8] H. Li, S. Shen, and B. Clerckx, A dynamic grouping strategy for beyond diagonal reconfigurable intelligent surfaces with hybrid transmitting and reflecting mode, *IEEE Transactions on Vehicular Technology*, 2023.

[9] Q. Li, M. El-Hajjar, I. Hemadeh, A. Shojaeifard, A. A. M. Mourad, B. Clerckx, and L. Hanzo, Reconfigurable intelligent surfaces relying on non-diagonal phase shift matrices, *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6367–6383, 2022.

[10] Y. Zhou, Y. Liu, H. Li, Q. Wu, S. Shen, and B. Clerckx, Optimizing power consumption, energy efficiency and sum-rate using beyond diagonal ris — a unified approach, *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[11] J. Zhang, J. Tang, W. Feng, X. Y. Zhang, D. K. C. So, K.-K. Wong, and J. A. Chambers, Throughput maximization for ris-assisted uav-enabled wpcn, *IEEE Access*, vol. 12, pp. 13 418–13 430, 2024.

[12] N. T. Nguyen, V.-D. Nguyen, H. Van Nguyen, Q. Wu, A. Tolli, ¨S. Chatzinotas, and M. Juntti, Fairness enhancement of uav systems with hybrid active-passive ris, *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[13] L. Pang, W. Wu, Y. Zhang, Y. Yuan, Y. Chen, A. Wang, and J. Li, Joint power allocation and hybrid beamforming for downlink mmwave-noma systems, *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 173–10 184, October 2021.

[14] X. Sun, N. Yang, S. Yan, Z. Ding, D. W. K. Ng, C. Shen, and Z. Zhong, Joint beamforming and power allocation in downlink noma multiuser mimo networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5367–5381, Augest 2018.

[15] J. Zhu, Q. Li, H. Chen, and H. V. Poor, Statistical csi based hybrid mmwave mimo-noma with max-min fairness, in *IEEE International Conference on Communications (ICC)*. June 2021.

[16] Z. Zhu, H. Deng, F. Xu, W. Zhang, G. Liu, and Y. Zhang, Hybrid precoding-based millimeter wave massive mimo-noma systems, *Symmetry*, vol. 14, no. 2, p. 412, February 2022.

[17] J. Du, W. Xu, Y. Deng, A. Nallanathan, and L. Vandendorpe, Energy saving uav-assisted multiuser communications with massive mimo hybrid beamforming, *IEEE Communications Letters*, vol. 24, no. 5, pp. 1100–1104,2020.

[18] R. Li, B. Guo, M. Tao, Y.-F. Liu, and W. Yu, Joint design of hybrid beamforming and reflection coefficients in ris-aided mmwave mimo systems, *IEEE Transactions on Communications*, vol. 70, no. 4, pp. 2404–2416, 2022.

[19] S. Li, B. Duo, M. D. Renzo, M. Tao, and X. Yuan, Robust secure uav communications with the aid of reconfigurable intelligent surfaces, *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6402–6417, 2021.

[20] S. Sobhi-Givi, M. Nouri, H. Behroozi, and Z. Ding, Joint bs and beyond diagonal ris beamforming design with drl methods for mmwave 6g mobile communications, *in 2024 IEEE Wireless Communications and Networking (WCNC 2024)*, 2024.

[21] S. Sobhi-Givi, M. G. Shayesteh, H. Kalbkhani, and N. Rajatheva, Resource allocation and user association for load balancing in nomabased cellular heterogeneous networks, in *2020 Iran Workshop on Communication and Information Theory (IWCIT)*, 2020, pp. 1–6.

[22] S. Sobhi-Givi, M. Nouri, M. G. Shayesteh, H. Kalbkhani, & Z. Ding, (2024). Joint Power

IEEE Press

Allocation and User Fairness Optimization for Reinforcement Learning Over mmWave-NOMA Heterogeneous Networks. *IEEE Transactions on Vehicular Technology*.

[23] K. Wang, R. Yang, Y. Zhou, W. Huang, and S. Zhang, Design and improvement of sd3-based energy management strategy for a hybrid electric urban bus, *Energies*, vol. 15, no. 16, p. 5878, 2022.

Mousa Abdollahvand received his M.Sc. degree in Communication Engineering from Shahed University, Tehran, and Ph.D. degree in Communication Engineering waves and fields from Tarbiat Modares University (TMU), Tehran, Iran. Also, he is a staff member of Mohaghegh Ardabili University, Iran. He was a research visitor at Universidad Politécnica de Madrid, Spain. His main interests are UWB antennas, antenna arrays, reflect array antennas, frequency selective surface, reconfigurable structures, equivalent circuit models, Artificial Intelligence (AI), and numerical methods in electromagnetics.

Sima Sobhi-Givi received her B.Sc., M.Sc., and Ph.D. degrees, all in electrical engineering, from Urmia University, Iran. She also completed a Post-doctoral Research Fellowship at Urmia University, supported by the Mobile Telecommunication Company of Iran (MCI), the largest mobile operator in the Middle East. Additionally, she served as the Project Manager for 5G Communication Systems at MCI. She is currently a faculty member at Mohaghegh Ardabili University. Her research interests include 5G and beyond 5G (B5G) wireless cellular networks, as well as machine learning applications in communication systems.

In Press