DRL-based User Fairness in Beyond Diagonal Reconfigurable Intelligent Surface-assisted Extremely Large Antenna Array Systems

Muhammad Abdullah Khan, *Graduate Student Member, IEEE,* Mahnoor Anjum, *Graduate Student Member, IEEE,* Deepak Mishra, *Senior Member, IEEE*, Haejoon Jung, *Senior Member, IEEE*, and Trung Q. Duong, *Fellow, IEEE*

***Abstract*—This paper investigates the user fairness for multi-user downlink communications systems with beyond diagonal-reconfigurable intelligent surface (BDRIS) and extremely large antenna array (ELAA), which takes advantage of interconnected elements and spherical wavefront characteristics in the near field. We formulate a user fairness-oriented optimization problem and develop a deep reinforcement learning (DRL)-based algorithm to effectively maximize fairness among users. The proposed framework provides a fast converging solution and simultaneously designs the transmit beamformers of the ELAA and the reconfiguration matrix of the BDRIS. Simulation results show that the proposed system provides better performance with perfect and imperfect channel state information (CSI) than typical DRL algorithms and improves the min-rate performance by 3.86% and 13.6%, compared to the benchmark ELAA systems with the conventional reconfigurable intelligent surface (RIS) and without RIS, respectively.**

***Index Terms*—Reconfigurable intelligent surface, deep reinforcement learning, extremely large antenna arrays.**

I. Introduction

Reconfigurable intelligent surfaces (RISs) are emerging as a promising, scalable, and energy-efficient solution for next-generation systems. RISs implement passive beamforming to improve signal quality, extend coverage, and mitigate interference. Concurrently, owing to the research and adoption of higher frequencies and increasing antenna array sizes in upcoming communication systems, near-field communications has emerged as a highly relevant area of interest in the current research landscape. The entities operating in the near-field experience spherical wavefronts as opposed to entities operating in the far-field, where the spherical nature of the waves becomes effectively negligible [1]. The elements of RISs operate independently of each other, limiting the achievable performance gains. As a result, beyond-diagonal RISs (BDRISs) are being explored for advanced passive beamforming with fully interconnected elements [2]. BDRISs can provide various performance enhancements, such as sum-rate maximization and outage minimization, typically assuming the far-field (FF) channels, while the conventional RIS systems have been exploited for the NF communications [1], [3].

To meet the stringent requirements of 6G systems, a large number of antennas can be employed at the base station (BS) to enhance performance, leading to the development of extremely large antenna arrays (ELAAs). ELAAs extend the influence of the near field (NF) and fundamentally change the propagation characteristics of electromagnetic waves by realizing spherical wavefronts, which complicates beamforming designs [4]. The beamforming gains enabled by the far-reaching NF in ELAAs cannot be realized with techniques based on traditional FF models. Furthermore, the fully connected nature of the BDRIS allows for better beamforming gains utilizing the same number of control bits, enabling more efficient control [5]. Even though the 3rd Generation Partnership Project (3GPP) standards emphasize a balanced distribution of network resources among users to ensure fair data rates for users [6], existing works mostly focus on the sum-rate optimization, this service fairness has not been thoroughly investigated, especially utilizing DRL in the case of BDRIS that can improve the beamforming gains compared to conventional RIS [7]. In addition, existing works predominantly focus on conventional convexifying techniques [9], which suffer from high complexity and limited scalability.

Motivated by these research gaps, we propose a deep reinforcement learning (DRL)-based algorithm to achieve user fairness in BDRIS-aided ELAA systems. The key contributions of this work are as follows:

- We propose the user fairness optimization of a novel system model, where both the BDRIS and the users are placed in the NF of the ELAA-based BS, while the users are in the FF of the BDRIS.
- We formulate a service-fairness problem and design the ELAA beamformers and the reconfiguration matrix of the BDRIS, while meeting the BDRIS phase control constraint and the power budget at the BS.
- We develop a DRL-based framework to solve the tightly coupled, non-convex problem. We provide simulation results to verify the fast convergence and efficacy of our algorithm by comparing its performance with typically employed DRL algorithms and standard benchmarks, including comparison with RIS-free and conventional RIS-assisted ELAA systems.

M. A. Khan and H. Jung are with the Department of Electronics Engineering, Kyung Hee University, Yongin-si, 17104, Korea (e-mail: {abdullah.khan, haejoonjung}@khu.ac.kr).

M. Anjum and D. Mishra are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia (e-mail: {mahnoor.anjum, d.mishra}@unsw.edu.au).

T. Q. Duong is with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K., and also with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, South Korea (e-mail: tduong@mun.ca).

This work was supported by Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00397480, System Development of Smart Repeater in Upper-mid Band).

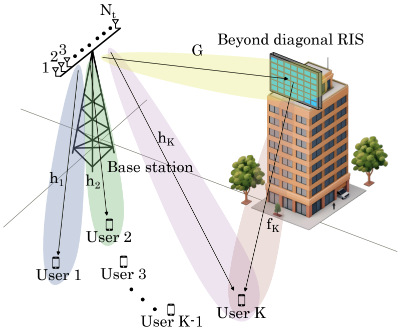

Fig. 1. System model of the BDRIS-enabled ELAA system.

II. System Model

In this work, we consider a BDRIS-enabled multi-user ELAA-based downlink communication system. As shown in Fig. 1, K users are assumed to be randomly distributed in the area covered by a BS equipped with a uniform linear array (ULA), and a BDRIS. The set of users is denoted by $\mathcal{K} = \{1, ..., K\}$. The BS consists of a large number of N_t isotropic antennas positioned at $(p_{tx}, p_{ty}, 0)$, while the fully connected BDRIS panel consists of $M_y \times M_z = M$ elements. The users and the BDRIS are placed in the NF of the ULA. At the same time, the users are assumed to be present in the far-field of the BDRIS. The users and the BDRIS have line-of-sight (LoS) links with the BS.

A. Wireless Channel Model

1) Near-field (NF) Channel: The NF of a communication entity is defined as the region covered by the Fraunhofer distance, calculated as $\frac{2D^2}{\lambda}$, where D is the largest dimension of the antenna array and λ is the transmission wavelength. The near field channel model is formulated in (1) based on [11], where $|\zeta_{tr}|$ represents the magnitude of the complex NF channel between the transmit and receive antennas. Also, (p_{rx}, p_{ry}, d) and A_r represent the location and area of the receive antenna element, respectively. The complete representation of the channel ζ_{tr} is given by

$$\zeta_{tr} = |\zeta_{tr}| e^{j 2\pi d_{tr}/\lambda}, \tag{2}$$

where d_{tr} is the distance between the transmit and the receive antennas.

2) Far-field (FF) Channel: Following [10], the FF LoS channel model between the BDRIS and the k^{th} user is described as the model considering the pathloss between the BDRIS and users with steering vectors designed in accordance with the far-field approximation, and is defined mathematically as

$$\mathbf{f}_k = \sqrt{M} \sqrt{\frac{\beta}{d_k^{\alpha}}} \mathbf{a}_t(\varphi_k, \vartheta_k), \tag{3}$$

where M is the number of elements, and d_k is the distance from the BDRIS to the k^{th} user. The channel gain at unit reference distance β is given by $(\frac{\lambda}{4\pi})^{\alpha}$, where α is the pathloss exponent. In addition, $\mathbf{a}_t(\varphi_k, \vartheta_k)$ is the transmit beam steering vector for the k^{th} user at azimuth angle φ_k and elevation angle ϑ_k. We define the steering vectors of the RIS as $\mathbf{a}_{M_y}(\varphi_k, \vartheta_k) = [e^{j 2\pi \frac{d(0)}{\lambda} \sin(\varphi_k) \sin(\vartheta_k)}, \ldots, e^{j 2\pi \frac{d(M_y - 1)}{\lambda} \sin(\varphi_k) \sin(\vartheta_k)}]^T$, and $\mathbf{a}_{M_z}(\vartheta_k) = [e^{j 2\pi \frac{d(0)}{\lambda} \cos(\vartheta_k)}, \ldots, e^{j 2\pi \frac{d(M_z - 1)}{\lambda} \cos(\vartheta_k)}]^T$, where d is the spacing between subsequent RIS elements. We model $\mathbf{a}_t(\varphi_k, \vartheta_k) = \mathbf{a}_{M_y}(\varphi_k, \vartheta_k) \otimes \mathbf{a}_{M_z}(\vartheta_k) \in \mathbb{C}^{M \times 1}$ with $\otimes$ denoting the Kronecker product.

B. Signal Model

The signal y_k received by the k^{th} user is given by

$$y_k = \mathbf{\Psi}_k^T \sum_{k \in \mathcal{K}} \mathbf{w}_k s_k + n_k, \tag{4}$$

where $\mathbf{w}_k \in \mathbb{C}^{N_t \times 1}$ is the beamforming vector for the k^{th} user at the ULA, while s_k is the signal for the k^{th} user transmitted by the BS such that $\mathbb{E}[|s_k|^2] = 1$, with $\mathbb{E}[.]$ denoting the expectation operator. Also, $\mathbf{\Psi}_k \in \mathbb{C}^{N_t \times 1}$ is the composite BS and BDRIS channel to the user k defined as $\mathbf{\Psi}_k = \mathbf{h}_k + \mathbf{G}\mathbf{\Theta}\mathbf{f}_k$, where $\mathbf{h}_k \in \mathbb{C}^{N_t \times 1}$ and $\mathbf{G} \in \mathbb{C}^{N_t \times M}$ are NF channels from the BS to k^{th} user and BDRIS respectively, defined in (2). On the other hand, $\mathbf{f}_k \in \mathbb{C}^{M \times 1}$ is the FF LoS channel between the BDRIS and user k, as defined in (3). Moreover, $\mathbf{\Theta} \in \mathbb{C}^{M \times M}$ is the phase-shift matrix of the BDRIS with ideal reflection coefficients such that $\mathbf{\Theta}\mathbf{\Theta}^H = \mathbf{I}_M$ [12], where $\mathbf{I}_M$ is the $M \times M$ identity matrix. It is important to note that the matrix $\mathbf{\Theta}$, as opposed to a typical singly-connected RIS, is not restricted to a diagonal matrix. Following [8], we model the entry in the p^{th} row and q^{th} column of $\mathbf{\Theta}$ as $r_{pq} e^{j\phi_{pq}}$, where $r_{pq} \in \{0, 1\}$ and ϕ is the phase reconfiguration provided by the BDRIS element. Lastly, $n_k \in \mathbb{C}$ is the additive complex Gaussian noise with zero mean and variance σ^2. Thus, the signal-to-interference-plus-noise ratio (SINR) of the k^{th} user is

$$\gamma_k(\{\mathbf{w}_k\}_{k \in \mathcal{K}}, \mathbf{\Theta}) = \frac{\|\mathbf{\Psi}_k^T \mathbf{w}_k\|^2}{\sigma^2 + \sum l \in \mathcal{K}, l \neq k \|\mathbf{\Psi}_k^T \mathbf{w}_l\|^2}. \tag{5}$$

Thus, the instantaneous achievable rate ε_k of the user k for $k \in \mathcal{K}$ is given by $\varepsilon_k = \log_2(1 + \gamma_k)$.Furthermore, the transmit power P_t is defined as $P_t(\{\mathbf{w}_k\}_{k \in \mathcal{K}}) = \sum_{k \in \mathcal{K}} \mathrm{Tr}(\mathbf{w}_k \mathbf{w}_k^{\mathrm{H}})$, where $\mathrm{Tr}(\cdot)$ is the trace of a square matrix. The maximum transmit power budget at the BS is $P_{\max}$, and $P_t \leq P_{\max}$ for all values of transmit beamformers.

C. Problem Formulation

In this section, we formulate the service fairness problem for the considered BDRIS-aided downlink ELAA system. We

$$|\zeta_{tr}| = \sqrt{\frac{d}{\sqrt{(p_{rx} - p_{tx})^2 + (p_{ry} - p_{ty})^2 + d^2}} \times \frac{(p_{rx} - p_{tx})^2 + d^2}{(p_{rx} - p_{tx})^2 + (p_{ry} - p_{ty})^2 + d^2} \times \frac{A_r}{4\pi((p_{rx} - p_{tx})^2 + (p_{ry} - p_{ty})^2 + d^2)}}. \tag{1}$$

jointly design the beamformers $\{\mathbf{w}_k\}_{k=1}^{K}$ and the reconfiguration matrix $\mathbf{\Theta}$ to maximize the minimum achievable rate, thereby ensuring user fairness. The service fairness optimization is formulated as problem (P1), while meeting the maximum transmit power constraint and the functional requirements of the BDRIS elements. The final optimization problem can be defined as

$$
\begin{aligned}
\mathrm{P1}:\ &\underset{\mathbf{\Theta},\{\mathbf{w}_k\}_{k\in\mathcal{K}}}{\text{maximize}} \quad \min(\varepsilon_k) \quad \forall k \in \mathcal{K}, && (6)\\
\text{subject to}\quad &\mathrm{C1}: \mathbf{\Theta}\mathbf{\Theta}^{H} = \mathbf{I}_M,\\
&\mathrm{C2}: P_t(\{\mathbf{w}_k\}_{k\in\mathcal{K}}) \leq P_{\max},
\end{aligned}
$$

where (C1) corresponds to the BDRIS reconfiguration matrix constraint [8], and (C2) represents the transmit power budget. It may be noted that in the case of a singly connected RIS, the reconfiguration matrix selected by an optimization algorithm can simply be normalized to fulfill the constraint (C1). However, in the case of fully connected BDRIS, normalizing the vectors of the matrix will not guarantee the fulfillment of this constraint. Hence, the procedures typically employed to handle RIS constraints are not applicable in the case of BDRIS. We also observe that this optimization problem is non-convex and the design variables are tightly coupled, including the symmetry of constraint (C1) [12]. For this reason, we utilize DRL to solve this using proximal policy optimization (PPO) with appropriate transformations to handle the constraints unique to BDRIS.

III. Service Fairness Optimization Using PPO

DRL algorithms represent a category of learning paradigms that rely on iterative interaction with, and adaptation to, the environment. Traditional reinforcement learning algorithms, such as Q-learning and deep Q-learning, require discrete variable spaces. To overcome this limitation, DRL algorithms have been developed to deal with continuous variables and action spaces. In this context, the PPO algorithm has emerged as a popular DRL algorithm for continuous action spaces and improves on previously developed algorithms such as Trust Region Policy Optimization (TRPO), achieving similar performance and a lower computational complexity. This off-policy DRL approach is markedly different from the typically used on-policy DRL algorithms, including SAC and DDPG. It relies on more stable updates owing to its clipped objective function and can mitigate its sample inefficiency through multiple environment interactions [13].

A. DRL Formulation

The learning process of a DRL algorithm is characterized by its interaction with the environment. This process is modeled as a Markov decision process (MDP) wherein an agent interacts with an environment by taking actions $a_t \in \mathcal{A}$ within the environment having state $s_t \in \mathcal{S}$ at time instant t. This action a_t, taken in accordance with the learned policy π, transitions the state s_t of the environment to the new state $s_{t+1} \in \mathcal{S}$, where $\mathcal{A}$ and $\mathcal{S}$ are the sets of all possible actions and states, respectively. This change of state is accompanied by a numerical reward r corresponding to the reward function based on the success of a_t in s_t.

Algorithm 1 Proposed PPO-based solution

```
Initialize actor and critic network weights $\vartheta_\varpi$ and $\vartheta_\delta$ respectively.
while No Convergence do
    Initialize random realizations of state variables $\{\{\mathbf{w}_k\}_{k\in\mathcal{K}}, \mathbf{\Theta}, \{\gamma_k\}_{k\in\mathcal{K}}\} \in \mathcal{S}$.
    Use policy $\pi_{\theta_\varpi}$ to take action $a_t \in \mathcal{A}$ at timestep $t$ which determines the values of the next state $s_{t+1} \in \mathcal{S}$ and outputs the reward $r_t$.
    Obtain the advantage value $\hat{A}_t$ from Eq. (7).
    Obtain the value of clipped objective function $L^{CLIP}(\theta_\varpi)$ defined in Eq. (8) using the advantage value $\hat{A}_t$.
    Update $\theta_\varpi$ and $\theta_\delta$ to obtain new weights for the policy and value neural networks, respectively, using gradient descent.
end while
```

The construction of an MDP for a DRL algorithm involves the realization of a triple tuple of state, action, and reward represented as $\{\mathcal{S}, \mathcal{A}, r\}$. The DRL algorithm then enables the agent to learn the mapping between the state and action space. The state, action, and reward for our problem formulation are defined as follows:

- *State*: Our system utilizes transmit beamforming vectors $\{\mathbf{w}_k\}_{k\in\mathcal{K}}$, the reconfiguration matrix $\mathbf{\Theta}$, and the instantaneous SINRs $\{\gamma_k\}_{k\in\mathcal{K}}$ as the state space, i.e., $\mathcal{S} \in \{\{\mathbf{w}_k\}_{k\in\mathcal{K}}, \mathbf{\Theta}, \{\gamma_k\}_{k\in\mathcal{K}}\}$.
- *Action*: The transmit beamforming vectors $\{\mathbf{w}_k\}_{k\in\mathcal{K}}$ and the reconfiguration matrix $\mathbf{\Theta}$ to define the action space, i.e., $\mathcal{A} \in \{\{\mathbf{w}_k\}_{k\in\mathcal{K}}, \mathbf{\Theta}\}$.
- *Reward*: The reward function r, considered for our user fairness-based sum-rate maximization, is formulated such that $r = \min\limits_{k\in\mathcal{K}}(\varepsilon_k)$.

B. Proximal Policy Optimization (PPO)

The objective function for a basic actor-critic-based DRL algorithm is given by $g = \mathbb{E}_t[\Delta_{\theta_\varpi} \log \pi_{\theta_\varpi}(a_t|s_t)\hat{A}_t]$, where $\mathbb{E}_t$ is the expectation, and Δ_{θ_ϖ} denotes the gradient of the weight matrix of the deep neural network. Furthermore, π_{θ_ϖ} denotes the policy of a network with weights configuration denoted by θ_ϖ considering the action a_t at state s_t and $\hat{A}_t$, if the advantage function defined as

$$
\hat{A}_t = \sum_{l=0}^{T-t} \nu^l r_{t+1} - V(\mathbf{\Theta}_t, \mathbf{w}_{\{k,t\}}, \gamma_{\{k,t\}}), \tag{7}
$$

where ν is the discount factor and V is the state value function of the critic network. A clipping function-based surrogate function is introduced in PPO and is defined as

$$
L^{CLIP}(\theta_\varpi) = \mathbb{E}_t[\min(c_t(\theta_\varpi)\hat{A}_t, \mathrm{clip}(c_t(\theta_\varpi), 1-\epsilon, 1+\epsilon)\hat{A}_t)], \tag{8}
$$

where ϵ is a hyperparameter for the clipping function of the probability ratio as

$$
c_t(\theta_\varpi) = \frac{\pi_{\theta_\varpi}\Big(\{\mathbf{w}_{k,t}\},\, \mathbf{\Theta}_t \,\Big|\, \{\mathbf{w}_{k,t}\},\, \mathbf{\Theta}_t,\, \{\gamma_{k,t}\}\Big)}{\pi_{\theta_{\varpi'}}\Big(\{\mathbf{w}_{k,t}\},\, \mathbf{\Theta}_t \,\Big|\, \{\mathbf{w}_{k,t}\},\, \mathbf{\Theta}_t,\, \{\gamma_{k,t}\}\Big)}, \tag{9}
$$

TABLE I
HYPERPARAMETERS.

Hyperparameter	Value
# of neurons in the hidden layers	$128, 64$
Learning rate η	0.0001
Discount factor γ	0.92
Batch size	32
Clipping parameter ϵ	0.32
Steps per update	256
Max. value for gradient clipping	0.51

where $\pi_{\theta_{\varpi'}}$ denotes the policy at instant $t-1$ preceding the policy $\pi_{\theta_{\varpi}}$ at instant t. PPO aims to improve the training stability, speed of convergence and focuses on the long-term rewards. To ensure that the solution derived from the optimization process meets the defined constraints, the following procedures are adopted:

- **Reconfiguration constraint (C1):** The obtained BDRIS reconfiguration matrix may or may not be unitary. In order to guarantee the fulfillment of the constraint, the matrix is projected to a space of unitary matrices using QR decomposition. The Q matrix, forming an orthonormal basis for the column space of the predicted reconfiguration matrix, is used as the reconfiguration matrix of the BDRIS. This effectively decouples the problem from the constraint (C1) and allows PPO to explore the space of all matrices in order to find the optimal reconfiguration matrix.
- **Power constraint (C2):** The transmit beamforming vector is normalized to keep the transmit power under the defined constraints.

C. Complexity Analysis

In this section, we analyze the complexity of the proposed scheme. As the PPO scheme exploits neural networks for learning, we utilize the computations involved in forward and backward passes to provide the big-$\mathcal{O}$ upper bound time complexity [14]. The input dimension D_I of the actor and critic networks is $D_I = KN_t + M^2 + K$, whereas the action dimension is $A = KN_t + M^2$. Let n_i denote the number of neurons in layer i. Since the proposed solution has 2 layers, the forward pass complexity of the actor and critic networks is $D_I n_1 + n_1 n_2 + n_2 A$. The reward computation requires QR-decomposition, which has a complexity of M^3 for our square-matrix $\mathbf{\Theta}$. The backward pass is expected to have twice the complexity of the forward pass. Hence, the total asymptotic complexity of the algorithm can be shown derived as $\mathcal{O}(M^3 + (KN_t + M^2 + K)n_1 + n_1 n_2 + (KN_t + M^2)n_2)$, which can be simplified to $\mathcal{O}((KN_t + M^2)(n_1 + n_2) + n_1(n_2 + K) + M^3)$. This asymptotic complexity is notably lower than the $\mathcal{O}(K^{\frac{3}{2}}M^4)$ complexity reported in [8], underscoring the computational efficiency of our proposed scheme.

IV. Results and Discussion

For the simulation, the BS and BDRIS are placed at $(x, y, z) = (0, 0, 0)$ and $(0, 60\text{m}, 0)$ respectively. The users are uniformly and randomly distributed in a 150m × 150m grid in the first quadrant jointly served by the BS and BDRIS [1]. The BS is equipped with $N_t = 50$ antennas [15], while the BDRIS is set to have $M = 36$ elements, each with element spacing of $\frac{\lambda}{2}$. The elevation for all system entities is constant. The largest dimension of the ULA is assumed to be $D = \frac{\lambda(N_t-1)}{2}$. The noise power σ^2, the frequency of operation and pathloss exponent α are set to -170 dBm, 2.4 GHz, and 2, respectively. The experiment is repeated over 10 random seeds with hyperparameters in Table I.

The performance of the proposed scheme is compared with Deep Deterministic Policy Gradient (DDPG) and Soft Actor-Critic (SAC) DRL algorithms under perfect and imperfect CSI as shown in Fig. 4. It can be seen that PPO significantly outperforms the typical DRL algorithms under perfect CSI and maintains its superior performance even under CSI uncertainty. The performance of the PPO algorithm is also evaluated under different architectures. These benchmarks include an RIS equivalent setup with a modified objective function according to the scheme proposed in [16], an ELAA setup optimized using DRL but without any RIS assistance, and an ELAA setup with random transmit beamforming vectors to emphasize the optimization gains from the DRL algorithm. The comparison is kept fair by fixing the number of elements of the RIS M and the number of transmit antennas N_t.

1) Convergence Analysis: The learning trend of the DRL algorithm is illustrated in Fig. 2, where the shaded region represents the 95% confidence interval over 10 training sequences. It can be seen that the training is highly stable over the training sequences and converges approximately at around 2×10^3 episodes each with 100 steps.

2) Performance Analysis of Proposed Scheme: As the sum-throughput is not a suitable metric for service fairness, we consider the outage probability to demonstrate the effectiveness of the proposed scheme by quantifying the outage likelihood. Fig. 5 shows the empirical CDF of the rates of the users obtained over 10^4 Monte Carlo experiments. It can be seen that the BDRIS-enabled PPO optimized system has the least probability of outage as compared to any of the considered schemes for a given rate threshold due to its superior reconfigurability performance. As expected, the RIS-enabled system performs better than the ELAA-only system due to the RIS beamforming gain. The non-optimized ELAA system tends to perform the worst due to its unoptimized transmit beamformers.

3) Impact of the Number of Antennas at the BS: The impacts of the number of antennas N_t and the number of BDRIS elements M are also analyzed using the CDF of user rates. In Fig. 6, the outage performance of the users is considerably improved with the increase in N_t.

4) Impact of the Number of Elements at the BDRIS: In Fig. 7, we observe that the user rates are enhanced by the increase in the number of BDRIS elements M. Thus, an increase in the number of antennas and reflecting elements positively affects the performance of the DRL optimized system due to an increase in the beamforming gains owing to the increase in the number of antennas and the number of BDRIS elements.

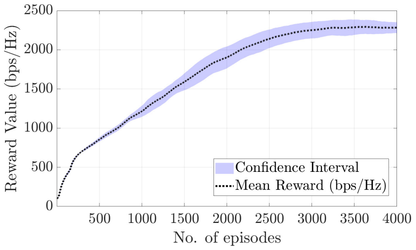

Fig. 2. The training curve of the proposed PPO-based learning scheme.

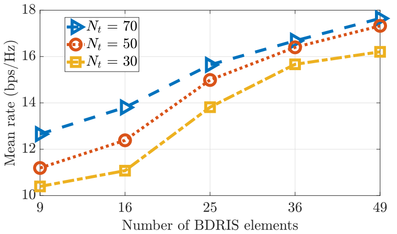

Fig. 3. The mean rates of the users with different N_t and M.

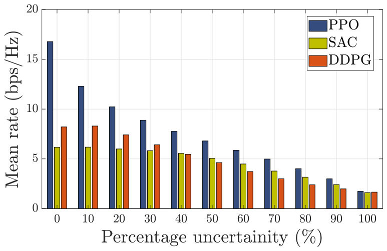

Fig. 4. Performance of DRL algorithms under perfect and imperfect CSI.

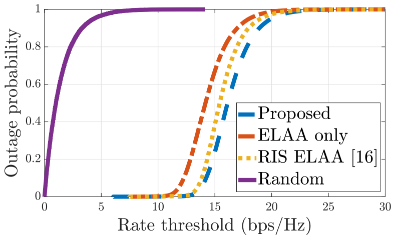

Fig. 5. The CDF of the rates of the users under different system configurations.

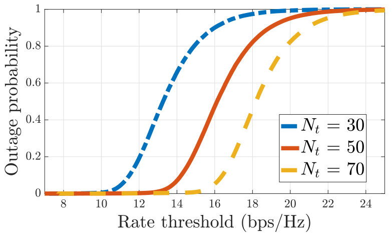

Fig. 6. The CDF of the user rates at different numbers of transmit antennas.

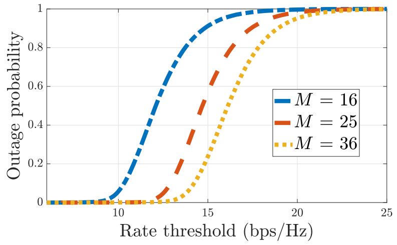

Fig. 7. The CDF of the user rates under different number of BDRIS elements.

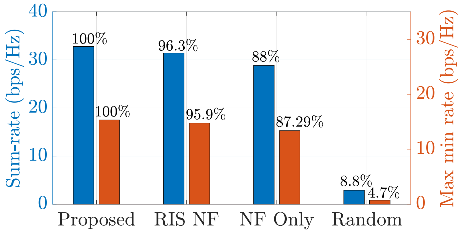

Fig. 8. The sum-rate and max min rate for different system architectures.

5) Sum-rate and Minimum Rate: The mean sum-rate and minimum of rates over multiple system realizations of the proposed and benchmark schemes are illustrated in Fig. 8. The proposed system achieves a mean sum-rate increase of 4.15% and 12% compared to the benchmark RIS and ELAA systems, respectively, while providing a slightly higher minimum rate with the same percentage increase over the benchmark systems, because of the reward chosen for the user fairness and sum-rate maximization of the considered systems.

V. Conclusion

In this paper, we proposed a DRL-based framework for user fairness in a BDRIS-aided ELAA system. We jointly designed the transmit beamformers and the reconfiguration matrix to maximize service fairness. The numerical analysis demonstrates the superior performance of our scheme compared to equivalent benchmark schemes. The proposed system achieves a mean sum-rate increase of 3.86% and 13.6% over the RIS and ELAA schemes. Future works may include the investigation and optimization across multiple quality-of-service domains in BDRIS and ELAA systems with stochastically faded channels in a variety of scattering environments.

References

[1] S. Lv *et al.*, “RIS-aided near-field MIMO communications: Codebook and beam training design,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 12531–12546, 2024.

[2] H. Li, S. Shen, M. Nerini, and B. Clerckx, “Reconfigurable intelligent surfaces 2.0: Beyond diagonal phase shift matrices,” *IEEE Commun. Mag.*, vol. 62, no. 3, pp. 102–108, 2024.

[3] J. Wang *et al.*, “Wideband beamforming for RIS assisted near-field communications,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 16 836–16 851, Nov 2024.

[4] Y. Cheng *et al.*, “Achievable rate optimization of the RIS-aided near-field wideband uplink,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2296–2311, Mar 2024.

[5] M. Nerini, S. Shen, and B. Clerckx, “Discrete-value group and fully connected architectures for beyond diagonal reconfigurable intelligent surfaces,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16 354–16 368, 2023.

[6] K. I. Pedersen *et al.*, “A tutorial on radio system-level simulations with emphasis on 3GPP 5G-Advanced and beyond,” *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2290–2325, 2024.

[7] M. Soleymani *et al.*, “Rate region of RIS-aided URLLC broadcast channels: Diagonal versus beyond diagonal globally passive RIS,” *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 320–324, Feb 2025.

[8] Y. Zhou *et al.*, “Optimizing power consumption, energy efficiency, and sum-rate using beyond diagonal RIS—a unified approach,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7423–7438, 2024.

[9] M. Delbari *et al.*, “Far-versus near-field RIS modeling and beam design,” Jan. 2024. [Online]. Available: http://arxiv.org/abs/2401.08237

[10] M. Anjum, D. Mishra, and A. Seneviratne, “Power-efficient transceiver design for full-duplex dual-function radar communication systems,” in *Proc. IEEE SPAWC*, Sep. 2024, pp. 11–15.

[11] E. Björnson and L. Sanguinetti, “Power scaling laws and near-field behaviors of massive MIMO and intelligent reflecting surfaces,” *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1306–1324, 2020.

[12] S. Shen, B. Clerckx, and R. Murch, “Modeling and architecture design of reconfigurable intelligent surfaces using scattering parameter network analysis,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1229–1243, 2022.

[13] J. Schulman *et al.*, “Proximal policy optimization algorithms,” Aug. 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[14] M. Umer *et al.*, “Deep reinforcement learning for trajectory and phase shift optimization of aerial RIS in CoMP-NOMA networks,” in *Proc. IEEE GLOBECOM*, Dec 2024, pp. 79–84.

[15] M. Monemi *et al.*, “Toward near-field 3D spot beam focusing: Possibilities, challenges, and use cases,” *IEEE Veh. Technol. Mag.*, vol. 20, no. 2, pp. 95–103, Jun 2025.

[16] A. Thakre *et al.*, “Fairness in reconfigurable intelligent surface (RIS) assisted MU-MISO systems using DRL,” in *IEEE Proc. WINTECHCON*, Nov 2024, pp. 1–5.