

Deep Reinforcement Learning Based Beamforming in RIS-assisted MIMO System Under Hardware Loss

Yuan Sun¹, Zhiquan Bai^{1*}, Jinqu Zhao¹, Dejie Ma¹, Zhaoxia Xian¹, and KyungSup Kwak²

¹Shandong Provincial Key Lab. of Wireless Communication Technologies,

School of Information Science and Engineering, Shandong University, Qingdao, Shandong, China

²Department of Information and Communication Engineering, INHA University, Incheon 22212, Korea

202212666@mail.sdu.edu.cn, zqbai@sdu.edu.cn*, 202020373@mail.sdu.edu.cn,

madj0212@163.com, xianzhaoxia2000@163.com, kskwak@inha.ac.kr

Abstract—Reconfigurable intelligent surface (RIS) is considered as one of the key enabling technologies for future 6G wireless communication by realizing an intelligent radio environment. RIS is used as reflective array to change the transmission and coverage of radio frequency (RF) signals. In this paper, we propose a deep reinforcement learning (DRL) based RIS beamforming design in practical scenarios where RIS may have hardware loss, and the soft actor-critic (SAC)-exploration algorithm is presented to solve the beamforming design. The algorithm reduces the prediction error by introducing a perturbation signal to influence the action prediction. Simulation results show that our proposed SAC-exploration algorithm has significant improvement over the typical SAC algorithm, which verifies the effectiveness of the proposed algorithm.

Index Terms—Reconfigurable intelligent surfaces (RIS), radio frequency, multiple input multiple output (MIMO), soft actor-critic (SAC), time division duplex (TDD).

I. INTRODUCTION

Reconfigurable intelligent surface (RIS) is an emerging technology that is widely recognized as a foundational technology for the forthcoming generation of wireless communications [1]. It comprises multiple reflectors with adjustable impedances to achieve the desired phase shifts of the reflected wave. Consequently, RIS improves radio frequency (RF) signal transmission performance, reduces signal interference, extends signal coverage, and thus improves communication quality.

Recently, the research of RIS has focused on how to perform beamforming design of RIS to achieve the optimal system performance. In [2], passive beamforming is designed for the RIS to maximize the sum rate of the RIS-assisted cellular network considering the frequency selectivity characteristics of the RIS. The signal-to-noise ratios (SNRs) of the target detections are weighted and maximized in [3] by jointly optimizing

the transmit beamforming and the RIS reflection coefficients while satisfying the quality of service (QoS) requirements of communication.

Considering the increasingly intricate radio environment, traditional optimization algorithms have become impractical for their complexities. However, with the continuous advancements in hardware design, computers now possess significantly enhanced computing capability. Consequently, a new alternative algorithm called reinforcement learning (RL) has emerged as a main method for machine learning (ML) based RIS-assisted communication. It gets applications in various areas, such as computing offloading of edge devices and fog computing [4], communication and trajectory planning within the internet of vehicles domain [5], system capacity enhancement in unmanned aerial vehicle (UAV)-assisted nonorthogonal multiple access (NOMA) network [6], and the optimization of beamforming and the phase design of RIS for efficient transmission [7] [8]. The phase design of RIS in previous scenarios has been idealized, and there is a lack of consideration for the hardware loss of the RIS.

In this paper, a deep reinforcement learning (DRL) approach is used for the RIS beamforming design in the RIS-assisted multiple input multiple output (MIMO) communication to maximize the system sum rate in the downlink transmission. The main contributions of this paper are summarized as follows. (1) We consider the hardware loss of RIS and propose a phase shift design for RIS-assisted MIMO communication. (2) We design a DRL based RIS beamforming algorithm for the above scheme, by introducing interference signals and interfering with action selection to prevent from falling into local optimal solutions, and the simulation results show that the proposed algorithm has a significant improvement over the typical algorithm.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

For the RIS-assisted wireless communication system, the system diagram is shown in Figure 1, with k single antenna users and one RIS with L elements, the channel model

The work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2021LZH003, in part by the Regional Innovation and Development Joint Foundation by National Natural Science Foundation of China under Grant U23A20277, in part by the National Defense Preliminary Research Project under Grant 90903030202, in part by the National Key Research and Development Plan under Grant 2022YFF0604900, in part by the National Nature Science Foundation of China under Grant 61771291, and in part by the National Research Foundation of Korea under Grant MSIT-NRF-2020R1A2B5B02002478. *Corresponding author.

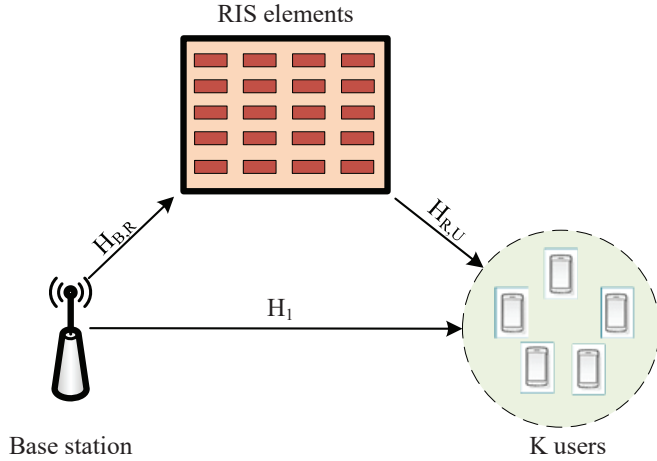


Fig. 1. RIS-assisted MIMO System

between the base station (BS) and the user k can be expressed as

$$\mathbf{h}_k = \mathbf{h}_{bs,k} + \sum_{l=1}^L \mathbf{h}_{bs,l} \phi_l h_{l,k}, \quad (1)$$

where $\mathbf{h}_{bs,k} \in \mathbb{C}^{N_B \times 1}$, $\mathbf{h}_{bs,l} \in \mathbb{C}^{N_B \times 1}$, and $h_{l,k}$ respectively denote the channel vector from the BS to the user k , from the BS to the RIS, and from the RIS to the user k . Considering the conditions of hardware loss, the phase-independent amplitude model of RIS [9] can be obtained as

$$\Phi \triangleq \text{diag}(\phi_1, \dots, \phi_L) \in \mathbb{C}^{L \times L}, \quad (2)$$

where $\phi_l = \zeta(\varphi_l) e^{j\varphi_l}$ is phase-independent amplitude model for each RIS, the phase-shift is $\varphi_l \in [0, 2\pi)$, and the amplitude satisfies

$$\zeta(\varphi_l) = (1 - \zeta_{\min}) \left(\frac{\sin(\varphi_l - \mu) + 1}{2} \right)^\kappa + \zeta_{\min}, \quad (3)$$

where $\zeta_{\min} \in [0, 1]$, $\mu \geq 0$, and $\kappa \geq 0$ are constants depend on the hardware configuration of RIS.

The single-user channel is aggregated to obtain the multi-user channel, which is denoted as

$$\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_{B,R} \Phi \mathbf{H}_{R,U}, \quad (4)$$

with

$$\mathbf{H}_1 = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K],$$

$$\mathbf{H}_{B,U} = [\mathbf{h}_{bs,1}, \mathbf{h}_{bs,2}, \dots, \mathbf{h}_{bs,K}],$$

and

$$\mathbf{H}_{R,U} = \begin{bmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,K} \\ h_{2,1} & h_{2,2} & \dots & h_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ h_{L,1} & h_{L,2} & \dots & h_{L,K} \end{bmatrix}.$$

B. Problem Formulation

The time division duplex (TDD) technique is utilized to implement the wireless transmissions in order to validate the advantages of the proposed algorithm and ensure its applicability and generality [10].

In the uplink stage, the BS simultaneously receives the pilot signals transmitted by the k users. The pilot signal vector received by the BS is expressed as

$$\mathbf{Y}_K = \mathbf{H} \mathbf{P} + \mathbf{N}, \quad (5)$$

where $\mathbf{P} \in \mathbb{C}^{K \times K}$ is the pilot pattern.

The BS uses minimum mean square error (MMSE) to estimate the channel matrix of the received pilot signals, that is

$$\hat{\mathbf{H}} = \mathbf{Y}_U \mathbf{P}^H (\mathbf{P} \mathbf{P}^H + \sigma_U^2 \mathbf{I})^{-1}. \quad (6)$$

When $\mathbf{P}$ is a unitary matrix, the above estimated channel matrix can be transformed.

$$\hat{\mathbf{H}} = \frac{\mathbf{Y}_U \mathbf{P}^H}{1 + \sigma_U^2}. \quad (7)$$

In the downlink transmission, data transmission with zero-forcing (ZF) precoding is performed with the following expression

$$\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K]^H = \mathbf{P}_{norm} (\hat{\mathbf{H}}^H \hat{\mathbf{H}})^{-1} \hat{\mathbf{H}}^H, \quad (8)$$

where $\mathbf{P}_{norm}$ is the normalized power matrix and written as

$$\mathbf{P}_{norm} = \text{diag} \left(\left[\frac{1}{\|\mathbf{c}_1\|_2}, \frac{1}{\|\mathbf{c}_2\|_2}, \dots, \frac{1}{\|\mathbf{c}_K\|_2} \right] \right). \quad (9)$$

The signal received by the k -th user is

$$y_{P,k} = \mathbf{c}_k^H \mathbf{h}_k x_k + \sum_{z \neq k} \mathbf{c}_z^H \mathbf{h}_k x_z + n_k, \quad (10)$$

where x_k is the signal sent to the user k , n_k is additive gaussian white noise. For the RIS-assisted wireless communication MIMO system, we will use the sum downlink rate as an indicator to measure the performance of the system. Thus, the problem can be expressed as

$$\begin{aligned} \max_{\phi} \quad & P_m \\ \text{s.t.} \quad & \varphi \in [0, 2\pi), \\ & \mu \geq 0 \\ & \kappa \geq 0 \end{aligned} \quad (11)$$

where $P_m = \sum_{k=1}^K r_k = \sum_{k=1}^K \log_2(1 + \text{SINR}_k)$, the signal to interference plus noise ratio (SINR) is given as

$$\text{SINR}_k = \frac{|\mathbf{c}_k^H \mathbf{h}_k|^2}{\sum_{z \neq k} |\mathbf{c}_z^H \mathbf{h}_k|^2 + \sigma_k^2}. \quad (12)$$

III. DEEP REINFORCEMENT LEARNING ALGORITHMS

In this section, we propose the deep reinforcement learning framework. We will introduce the principle of soft actor-critic (SAC) algorithm first, and then we present the improved SAC algorithm with the environment parameters.

A. Soft Actor-Critic Algorithm

In contrast to typical DRL algorithms, we use the advanced SAC algorithm [11]. The training of the algorithm involves a trade-off between the maximized expected returns and the maximized entropy. Unlike *on-policy* algorithms, SAC is an *off-policy* algorithm that requires an experience replay buffer for caching experiences to facilitate the subsequent model training. By introducing entropy regularization, the SAC algorithm enables the agent to explore more possible actions while preventing it from getting trapped in local optima.

The SAC agent consists of three networks: two Q networks and a random policy network (or actor network). Each of them is a multi-layer perceptron (MLP). The reason for taking two Q networks is to reduce the overestimation of Q values. The Q networks take the states provided by the environment and actions produced by the actor network as inputs and produce Q-value estimates, which are scalar values. Given a network of actors π_ψ parameterized by ψ , the Q network is trained jointly in the SAC algorithm as

$$\hat{y} \triangleq \tilde{r} + \min_{i=1,2} Q_{\theta'_i}(s', a') |_{a' \sim \pi_\psi(\cdot | s')} - \alpha \log(a' | s'), \quad (13)$$

$$J(\theta_i) = \frac{1}{N} \|\hat{y} - Q_{\theta_i}(s, a)\|_2^2, \quad (14)$$

$$\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} J(\theta_i), \quad (15)$$

where $(s, a, r, s')_{i=1}^N$ is a small batch of experience tuples randomly drawn from the experience replay buffer, N denotes the size of the experience tuple. θ_i is the network parameter of the i th Q-network. α is the entropy regularization term, and $\|\cdot\|_2$ is the L_2 norm.

Similarly, the policy network takes state vectors from the environment as inputs and produces numerical action vectors. The loss function for the policy network in the SAC algorithm is calculated as

$$J(\psi) \triangleq \frac{1}{N} \sum_i \alpha \log \pi_\psi(\hat{a}_i | s_i) - \min_{j=1,2} Q_{\theta_j}(s_i, \hat{a}_{\hat{\beta}, i}) \Big|_{\hat{a} \sim \pi_\psi(\cdot | s)} \quad (16)$$

For the update of the policy network, we use the stochastic gradient algorithm for the policy gradient $\nabla_\psi J(\psi)$ and the gradient ascent algorithm to update the parameter of the policy network as

$$\psi \leftarrow \psi + \eta \nabla_\psi J(\psi). \quad (17)$$

Finally, the regularization factor of entropy, α is a hyper-parameter, and a larger value of α means that the agent will have a larger exploration space, which enables agent to explore more possibilities instead of being limited to the current action.

B. SAC-exploration

Environments for RL are categorized as episodic and non-episodic. When the termination condition is satisfied, the agent stops interacting with the environment, whereas the non-episodic tasks, that donot have a defined termination condition will lead to endless interaction between the agent and the environment. Therefore, we need to set a systematic threshold as a condition for termination.

- **Action:** For the optimization problem we have constructed, the action of the agent is the beamforming design of the RIS. Since the neural networks cannot handle complex numbers, we need to separate the real and imaginary parts of the RIS phases, and the action space contains $2L$ elements.
- **State:** To take full advantage of the information obtained by the agent interacting with the environment, we define the state space as $\mathbf{H}$ and Φ , the equivalent wireless channel matrix and RIS beamforming. In order to reduce the influence of the correlation of the channel state matrix, we perform a whitening operation on the channel matrix after each environmental interaction.
- **Reward:** For each episode, we will use the sum rate of the downlink transmission to determine the reward in order to choose the right action to get a higher reward for the agent.

For the RL, we cannot just focus on the reward that we get instantaneously, future gains are necessary to consider. We can increase or decrease the size of the future influence on the present reward by adjusting γ , and we set initially $\gamma = 1$.

For the non-episode task, we can use average reward instead of the instantaneous reward to allow the agent to learn from the environment as

$$\tilde{r} \triangleq r - \bar{r}, \quad (18)$$

where r is the immediate reward computed by the environment in the current state and $\bar{r}$ is the average of the rewards collected before the current state. For the typical SAC algorithm, it is not a good solution of our proposed problem. Therefore, we propose an SAC-exploration algorithm that adds the explorer network to the typical SAC algorithm to interfere with the actions chosen by the policy, and constantly maximize the prediction error of the Q network. As a result, this causes the agents to enter the state-action space where Q prediction is difficult to perform, thus allowing them to correct the prediction error of the unknown or less-selected actions as

$$a_\zeta \triangleq a \odot \lambda \cdot \xi_\omega(s) \implies \hat{\phi}_\zeta \triangleq [\hat{\zeta}_1 e^{j\varphi_1} \dots \hat{\zeta}_L e^{j\varphi_L}]^\top, \quad (19)$$

where $\odot$ is the Hadamard product and the $\lambda \in (0, 1]$ is used to restrict the explorer network from influencing the policy network to choose the wrong action. At each update step, perturbed actions are sampled from the experience replay buffer instead of the original actions generated by the policy network, and then the loss functions for the Q network and the actor network are as follows

$$\hat{y}_\zeta \triangleq \tilde{r} + \min_{i=1,2} Q_{\theta'_i}(s', a' \odot \lambda \cdot \xi_{\omega'}(s)) - \alpha \log(a' | s'), \quad (20)$$

Algorithm 1 Proposed SAC- exploration Algorithm

- 1: Initialize Q-networks Q_{θ_i} with parameters ψ, θ_1, θ_2 , actor network π_ψ and explorer network ξ_ω with parameters ω
- 2: Initialize target networks: $\hat{\theta}_1 \leftarrow \theta, \hat{\theta}_2 \leftarrow \theta_2, \omega' \leftarrow \omega$ and experience replay buffer M
- 3: **for** each environment step **do**
- 4: Observe state s and select an action $a \sim \pi_\psi(\cdot|s)$
- 5: Perturb the selected action: $a_{\hat{\zeta}} = a \odot \lambda \cdot \xi_\omega(s)$
- 6: Import $a_{\hat{\zeta}}$, and observe reward r and next state s'
- 7: Store $(s, a, a_{\hat{\zeta}}, r, s')$ into the experience replay buffer
- 8: **end for**
- 9: **for** each update step **do**
- 10: Randomly select an empirical tuple $(s, a, a_{\hat{\zeta}}, r, s')_{i=1}^N$
- 11: Update Q-networks by(20)and(16);
 $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} J(\theta_i)$
- 12: Update the actor network by(19);
 $\psi \leftarrow \psi + \eta \nabla_{\psi} J(\psi)$
- 13: Update the explorer network by(23)(24) and (25)
- 14: Update the target networks:
 $\theta'_{i=1,2} \leftarrow (1 - \tau) \theta'_{i=1,2} + \tau \theta_{i=1,2}$
 $\omega' \leftarrow (1 - \tau) \omega' + \tau \omega$
- 15: **end for**

$$J_{\hat{\zeta}}(\theta_i) \triangleq \frac{1}{N} \left\| \hat{y}_{\hat{\zeta}} - Q_{\theta_i}(s, a_{\hat{\zeta}}) \right\|_2^2, \quad (21)$$

and

$$J_{\hat{\zeta}}(\psi) \triangleq \frac{1}{N} \sum_i \alpha \log \pi_\psi(\hat{a}_i | s_i) - \min_{j=1,2} Q_{\theta_j}(s_i, \hat{a}_{\hat{\zeta},i}) \Big|_{\hat{a} \sim \pi_\psi(\cdot|s)}, \quad (22)$$

which $\hat{a}_{\beta,i} = \hat{a} \odot \lambda \cdot \xi_\omega$. We define the loss function of the explorer network by maximizing the sum of the absolute temporal-difference (TD) errors corresponding to the two Q networks as

$$\tilde{\delta}_{\hat{\zeta}_i}(s, a_{\hat{\zeta}}) \triangleq \frac{1}{N} \left\| \hat{y}_{\hat{\zeta}} - Q_{\theta_i}(s, a_{\hat{\zeta}}) \right\|_2^2, \quad (23)$$

$$J_\omega = \tilde{\delta}_{\hat{\zeta}_1}(s, a_{\hat{\zeta}}) + \tilde{\delta}_{\hat{\zeta}_2}(s, a_{\hat{\zeta}}). \quad (24)$$

The deterministic exploration policy gradient is then computed by the deterministic policy gradient algorithm [12] and used to perform the gradient ascent over the TD errors defined by (23),

$$\nabla_\omega J(\omega) = \sum_{i=1}^2 \mathbb{E}[\nabla_{\lambda \cdot \xi_\omega(s)} \tilde{\delta}_{\hat{\zeta}_i}(s, a_{\hat{\zeta}}) \nabla_\omega \xi_\omega(s)], \quad (25)$$

$$\omega \leftarrow \omega + \eta \nabla_\omega J(\omega). \quad (26)$$

Overall, we refer to the resulting algorithm as SAC-exploration and provide the pseudocode in Algorithm 1.

IV. SIMULATION RESULTS

The BS is equipped with a uniform linear array (ULA) that is placed at $[1, 0, 0]$ (i.e., x-axis), and RIS is a ULA, and located at $[0, 1, 0]$ (i.e., y-axis), and the user is equipped with

single antenna and is randomly generated in a circular range. We set the number of the BS antennas as 4, the number of users as 4, and the number of RIS elements as 32. One channel episode includes 20 channel blocks. The parameters of the Q network and the parameters of the RIS hardware loss model are given in Table I.

TABLE I

Hyper-Parameter	Value
hidden layers	2
units in each hidden layer	256
Hidden layers activation	ReLU
actor final layer activation	tanh
explorer final layer activation	tanh
Total time steps per training	20000
Mini-batch size	16
Discount term γ	1
Learning rate for target networks τ	10^{-3}
Initial α	0.2
μ	0
κ	1.5

To verify the effectiveness of the proposed algorithm, we compare the three algorithms, SAC, SAC-exploration and random, under the same configuration. Since the DRL is highly randomized, we conduct 10 experiments and average them in order to facilitate the observation, which makes the results more convincing. In Fig. 2, we analyze the influence

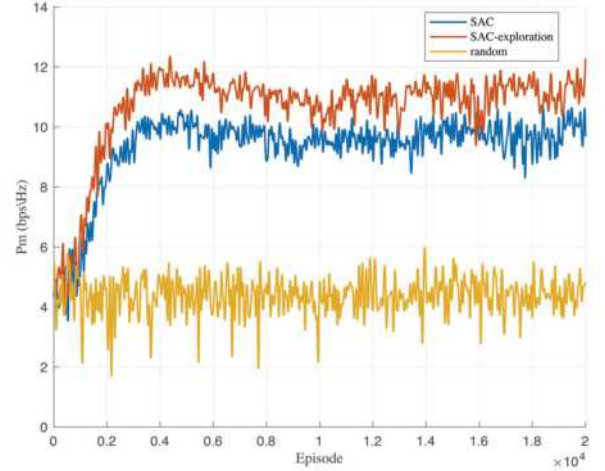


Fig. 2. Effect of different algorithms on P_m at $\zeta = 0.6$ and $L = 32$.

of different algorithms on the system sum rate. Compared to the random algorithm, the RIS beamforming design can significantly improve the system sum rate, and the proposed algorithm achieves significant improvement over the SAC algorithm. This is because, unlike the typical DRL algorithm, the error in the value function for the proposed algorithm can be corrected by introducing a perturbation signal to interfere with the action selection. Thus, the agent may select the optimal action more accurately, and the effectiveness of the

proposed algorithm can be verified by the fact that the system sum rate P_m is significantly improved.

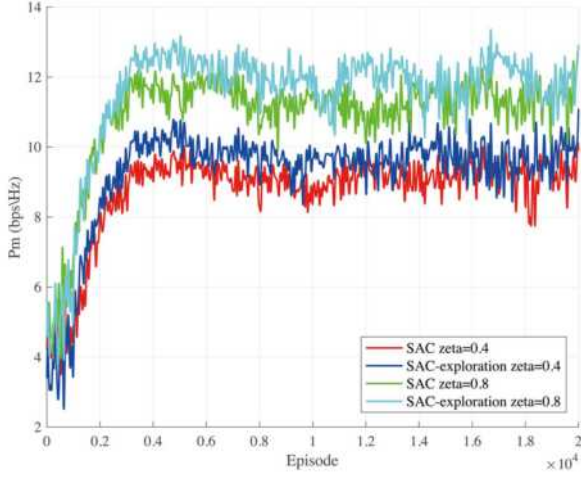


Fig. 3. Effect of different ζ values on P_m at $L = 32$.

The performance of the proposed algorithm for different ζ is investigated in Fig. 3. For the convenience of observation, we set the typical SAC algorithm as a benchmark. Comparing with Fig. 2, it can be seen that the proposed DRL algorithm with $\zeta = 0.6$ has a significant improvement in the system sum rate over the SAC algorithm, and it is also observed that the system sum rate P_m of the proposed algorithm increases as the value of ζ increases, which is due to the fact that the larger the value of ζ , the closer it is to the ideal condition, thus validating the effectiveness of the algorithm.

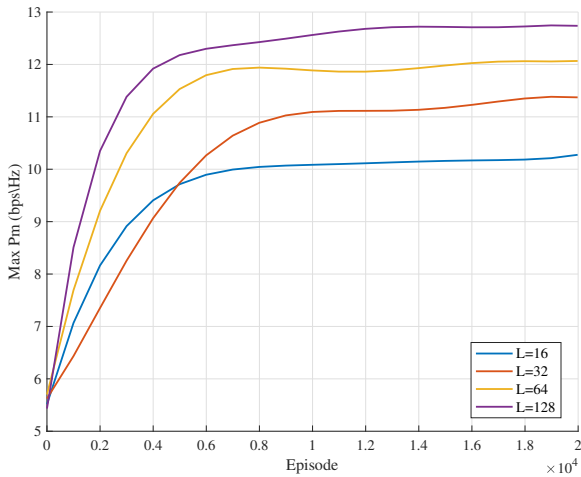


Fig. 4. Effect of different number of RIS elements on the maximum system performance at $\zeta = 0.6$.

Fig. 4 depicts the effect of different number of RIS elements on the system sum rate performance for the proposed algorithm. In order to better observe the experimental results, we take the max P_m within 1000 episodes for the simulation

analysis. We can observe that, as the number of RIS elements L increases, the maximum system sum rate P_m increases, and the increase in the P_m is achieved at the expense of the complexity. The simulation results show that the proposed algorithm has significant advantages in future practical applications as well.

V. CONCLUSION

In this paper, we study the RIS beamforming design in the RIS-assisted MIMO communication and propose a modified SAC-exploration algorithm to solve the RIS beamforming design considering the RIS hardware loss. The simulation results show that the proposed algorithm is able to realize better RIS beamforming and has a significant improvement in the downlink system sum rate compared with the typical SAC algorithm. Considering the case of multiple RISs, we will consider the multi-agent algorithm to solve the beamforming design problem in the future.

REFERENCES

- [1] C. Cai, X. Yuan and Y. -J. A. Zhang, "RIS Partitioning Based Scalable Beamforming Design for Large-Scale MIMO: Asymptotic Analysis and Optimization," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6061-6077, Sept. 2023.
- [2] S. Liu, R. Liu, M. Li, Y. Liu and Q. Liu, "Joint BS-RIS-User Association and Beamforming Design for RIS-Assisted Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6113-6128, May 2023.
- [3] H. Luo, R. Liu, M. Li and Q. Liu, "RIS-Aided Integrated Sensing and Communication: Joint Beamforming and Reflection Design," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9626-9630, July 2023.
- [4] M. Goudarzi, M. Palaniswami and R. Buyya, "A Distributed Deep Reinforcement Learning Technique for Application Placement in Edge and Fog Computing Environments," *IEEE Trans Mob Comput*, vol. 22, no. 5, pp. 2491-2505, 1 May 2023.
- [5] Q. Liu, Y. Zhu, M. Li, R. Liu, Y. Liu and Z. Lu, "DRL-based Secrecy Rate Optimization for RIS-Assisted Secure ISAC Systems," *IEEE Trans. Veh. Technol.*, pp.1-5, 2023.
- [6] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang and K. Long, "Capacity Maximization in RIS-UAV Networks: A DDQN-Based Trajectory and Phase Shift Optimization Approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2583-2591, April 2023.
- [7] Y. Zhu, M. Li, Y. Liu, Q. Liu, Z. Chang and Y. Hu, "DRL-based Joint Beamforming and BS-RIS-UE Association Design for RIS-Assisted mmWave Networks," *IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 345-350.
- [8] C. Huang et al., "Multi-Hop RIS-Empowered Terahertz Communications: A DRL-Based Hybrid Beamforming Design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663-1677, June 2021.
- [9] S. Abeywickrama, R. Zhang, Q. Wu and C. Yuen, "Intelligent Reflecting Surface: Practical Phase Shift Model and Beamforming Optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849-5863, Sept. 2020.
- [10] W. Wang and W. Zhang, "Intelligent Reflecting Surface Configurations for Smart Radio Using Deep Reinforcement Learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2335-2346, Aug. 2022.
- [11] C. Banerjee, Z. Chen and N. Noman, "Improved Soft Actor-Critic: Mixing Prioritized Off-Policy Samples With On-Policy Experiences," *IEEE Trans Neural Netw Learn Syst*, pp. 1-9, 2022.
- [12] X. Yang, H. Zhang and Z. Wang, "Data-Based Optimal Consensus Control for Multiagent Systems With Policy Gradient Reinforcement Learning," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 8, pp. 3872-3883, Aug. 2022.



Yuan Sun is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, deep reinforcement learning, beamforming and signal processing.



Zhiquan Bai received the M.Eng. degree in communication and information system from Shandong University, Jinan, China, in 2003, and the Ph.D. degree (Hons.) from INHA University, Incheon, South Korea, in 2007, under the Grant of Korean Government IT Scholarship. He held a postdoctoral position with INHA University, and was a Visiting Professor with The University of British Columbia, Canada. He is currently a Professor with the School of Information Science and Engineering, Shandong University. His research interests include cooperative

technology and spatial modulation, orthogonal time frequency space modulation, MIMO technology, resource allocation and optimization, and deep-learning based 5G wireless communications. He is a member of the editorial board of Journal of Systems Engineering and Electronics and also an associate editor of the International Journal of Communication Systems.



Jinqiu Zhao received B.E. degree from Shandong Normal University, Jinan, China, in 2020. She is currently pursuing her Ph.D. degree in the School of Information Science and Engineering, Shandong University, Qingdao, China. Her main research interests include reconfigurable intelligent surface and machine learning.



Dejie Ma is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. His research interests include reconfigurable intelligent surface, integrated sensing and communication and signal processing.



Zhaoxia Xian received the B.S. degree from the School of Physical and Electronic Sciences at Shandong Normal University in Jinan, China. She is currently pursuing the M.S. degree in Electronic Information at the School of Information Science and Engineering, Shandong University, Qingdao, China. Her research interests include signal processing, signal recognition and deep learning.



Kyung Sup Kwak received his BS degree from the Inha University, Incheon, Korea, in 1977 and his MS degree from the University of Southern California in 1981 and his PhD degree from the University of California at San Diego in 1988, under the Inha University Fellowship and the Korea Electric Association Abroad Scholarship Grants, respectively. From 1988 to 1989, he was with Hughes Network Systems, San Diego, California. From 1989 to 1990, he was with the IBM Network Analysis Center, North Carolina. Since then, he has been with the School

of Information and Communication Engineering, Inha University, Korea, as a professor. He is the director of UWB Wireless Communications Research Center (UWB-ITRC). Since 1994, he served as a member of the board of directors and the vice president and the president of Korean Institute of Communication Sciences (KICS) in 2006 and the president of Korea Institute of Intelligent Transport Systems (KITS) in 2009. He received many research awards, such as the award of research achievements in UWB radio from the Ministry of Information and Communication and Prime Ministry of Korea in 2005 and 2006, respectively. In 2008, he is elected as Inha Fellow Professor (IFP). In 2010, he received the Korean President official commendation for his contribution to ICT innovation and industrial promotion. He published more than 100 SCI journal papers, 300 conference/domestic papers, obtained 20 registered patents and 35 pending patents, and proposed 21 technical proposals on IEEE 802.15 (WPAN) PHY/MAC. He is one of the members of the IEEE, IEICE, KICS, and KIEE. His research interests include multiple access communication systems, cognitive radio, UWB radio systems and WBAN, WPAN, and sensor networks.