Joint Beamforming with Extremely Large Scale RIS: A Sequential Multi-Agent A2C Approach

Zhi Chai, *Graduate Student Member, IEEE*, Jiajie Xu, *Member, IEEE*, Justin P. Coon, *Senior Member, IEEE*, Mohamed-Slim Alouini, *Fellow, IEEE*

***Abstract*—Jointly optimizing the base station (BS) precoding matrix and reconfigurable intelligent surface (RIS) phases is challenging in RIS-assisted multi-user multiple-input single-output (MU-MISO) systems, particularly with extremely large RISs. This paper proposes a deep reinforcement learning (DRL) approach based on a sequential multi-agent advantage actor-critic (A2C) framework that accounts for discrete RIS phases, imperfect channel state information (CSI), and inter-user channel correlation. The computational complexity is analyzed, and the proposed method is benchmarked against the zero-forcing (ZF), minimum mean square error (MMSE), and single-agent A2C beamformers in terms of sum spectral efficiency (SE). Simulation results show that the proposed algorithm achieves higher SE and a certain degree of robustness to moderate channel estimation errors.**

***Index Terms*—RIS, joint beamforming, multi-agent deep reinforcement learning**

I. Introduction

Reconfigurable intelligent surfaces (RISs) have emerged as a promising technology for next-generation wireless communications due to their low fabrication cost, ease of deployment, and ability to provide anomalous reflection [1]. A key challenge in integrating RISs into existing systems is the joint design of the BS precoding matrix and RIS element phases to optimize system performance metrics such as energy efficiency, spectral efficiency, or total transmit power in multi-user MISO (MU-MISO) scenarios. Several studies have addressed this joint optimization problem using different approaches. For example, [2] formulated a power minimization problem under signal-to-interference-plus-noise ratio (SINR) constraints, employing a two-layer penalty-based algorithm to separate variables and manifold optimization to handle non-convex unit-modulus constraints. In contrast, [3] jointly optimized BS-RIS-user association and beamforming for sum rate maximization using fractional programming (FP) and block coordinate descent (BCD) to decouple the problem, while applying majorization-minimization (MM) and alternating direction method of multipliers (ADMM) to solve the subproblems.

Z. Chai, and J. P. Coon are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, United Kingdom. (e-mail: zhi.chai@eng.ox.ac.uk, justin.coon@eng.ox.ac.uk). This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-24-2-0102.

J. Xu and M-S. Alouini are with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955, Kingdom of Saudi Arabia. (e-mail: jiajie.xu.1@kaust.edu.sa, slim.alouini@kaust.edu.sa).

In contrast to conventional optimization-based methods [2], [3], deep reinforcement learning (DRL) has emerged as an effective end-to-end approach for joint beamforming design [4]–[6]. Early DRL-based works [4], [5] employ deep deterministic policy gradient (DDPG) algorithms but assume continuous RIS phase control and are limited to relatively small RIS sizes. To improve practicality, [6] considers discrete RIS phase shifts and adopts a multi-agent DRL framework, where one agent learns the BS precoding matrix and multiple agents learn RIS beam patterns for different user clusters. However, the agents operate independently, without explicit cooperation.

It is worth noting that, based on existing RIS prototypes [7], [8] and path loss models for RIS-assisted systems [9], a practical RIS deployment typically requires on the order of 1000 elements. However, most existing studies focus on relatively small RISs [2], [3], [5], [6], leaving the scalability of their approaches to large-scale RISs largely unexplored. Moreover, channel correlation and imperfect channel state information (CSI) are seldom considered, despite their importance in realistic indoor scenarios with spatially clustered users. In this paper, we propose a multi-agent advantage actor-critic (A2C) framework for joint BS–RIS beamforming in extremely large-scale RIS-assisted systems, explicitly accounting for discrete RIS phase adjustment, user channel correlation, and imperfect CSI. Compared with deep deterministic policy gradient (DDPG) and asynchronous advantage actor-critic (A3C), A2C is adopted due to its ability to naturally handle discrete action spaces and its improved learning stability via synchronized updates. The main contributions are summarized as follows: 1) We propose to use the sequential multi-agent A2C to solve the joint optimization problem with extremely large-scale RIS while considering the channel correlations, discrete RIS phase adjustment, and the imperfect CSI, 2) We demonstrate that the proposed method outperforms zero forcing (ZF) beamforming after quantization in terms of sum spectral efficiency (SE) while maintaining lower computational complexity, 3) We validate the robustness of the proposed method under moderate channel estimation errors and inter-user channel correlation.

II. System model and problem statement

We consider an indoor RIS-assisted MU-MISO wireless communication system, where the BS is equipped with N_T antennas, the RIS has N_R elements, and there are K users located in a square room of size a^2. To emphasize the role of the RIS, it is assumed that the line-of-sight (LOS) path is blocked, such that the indirect BS–RIS–user link is the only

viable communication path. To account for different spatial user patterns, two types of user distributions are considered. In the first case, users are uniformly distributed within the room, representing a baseline scenario with spatially independent user locations. In the second case, clustered user distributions are considered to better reflect realistic indoor environments, where users tend to gather around specific areas such as desks, kiosks, or meeting spaces. Specifically, users are uniformly distributed within a circle of radius R_c, forming spatial clusters. We assume that the cascaded BS–RIS–user channel can be estimated, and that the estimates may be subject to errors. Considering that the RIS element length, width, and element spacing are smaller than the sub-wavelength [8], the near-field region typically occupies only a small portion of the overall indoor environment. For example, the Rayleigh distance ranges from approximately 0.5 m to 2 m when the element width, length, and spacing are set to $\lambda/4$, with wavelengths of 10 mm, 6 mm, and 5 mm corresponding to RIS sizes from 900 to 1600 elements. Consequently, the far-field assumption provides an accurate and reasonable approximation in the considered scenarios. Additionally, joint beamforming is carried out based on the estimated CSI, with the objective of maximizing the sum spectral efficiency (SE). The signal-to-interference-plus-noise ratio (SINR) is used for the sum SE calculation. The sum SE is defined as

$$\text{SE} = \sum\nolimits_{k=1}^{K} \log_2(1 + \text{SINR}_k), \tag{1}$$

where SINR_k is the SINR of the kth user, and it is defined as

$$\text{SINR}_k = \frac{\|h_{2,k}^T \Phi H_1 w_k\|^2}{\sum_{j=1, j\neq k}^{K} \|h_{2,k}^T \Phi H_1 w_j\|^2 + \sigma^2}, \tag{2}$$

where $h_{2,k} \in \mathbb{C}^{N_R \times 1}$ is the channel between the RIS and the kth user, $\Phi \in \mathbb{C}^{N_R \times N_R}$ is the diagonal matrix of the RIS phase response, $H_1 \in \mathbb{C}^{N_R \times N_T}$ is the channel between the BS and the RIS, $w_k \in \mathbb{C}^{N_T \times 1}$ is the precoding vector for the kth user, and σ^2 is the noise power. $H_1 = \sqrt{R_1} H_1^{\text{uncorr}}$, where $R_1 \in \mathbb{R}^{N_R \times N_R}$ is the matrix that captures the correlation between elements on the RIS. The i, jth entry of R_1 is given by $R_{1,i,j} = \rho^{|m_x - n_x|} \rho^{|m_y - n_y|}$. ρ is the correlation coefficient, (m_x, m_y) and (n_x, n_y) are the coordinates of the ith and jth RIS elements [10]. The amplitude of each entry in H_1^{uncorr} is modeled as an identical and independently distributed (i.i.d.) Rayleigh random variable. The ith entry in the diagonal matrix Φ belongs to a discrete set $\{0, 2\pi/2^M, \cdots, 2\pi(2^M - 1)/2^M\}$ where M is the resolution for the RIS phase adjustment. We define W as the precoding matrix, where the kth column corresponds to the precoding vector of the kth user. We define H_2 as the RIS-user channel matrix, where the kth column corresponds to the kth RIS-user channel. We define H_2^{uncorr} as the RIS-user channel matrix without channel correlation. Each entry in H_2^{uncorr} is modeled as an i.i.d. complex normal distribution. The relationship between H_2 and H_2^{uncorr} is given as $H_2 = \sqrt{R_2} H_2^{\text{uncorr}}$, where R_2 is the correlation matrix. Assuming the communications are in the far field, the angle separation between users is the dominant factor for the channel correlation. Given the locations of users, RIS, and BS, the i, jth entry of R_2 is defined as $R_{2,i,j} = \exp(-\theta_{i,j}^2/\beta)$ where $\theta_{i,j}$ represents the angle between the ith user and the jth user, and β is a constant [10]. Next, we describe the modeling of channel estimation error. We define the estimated BS-RIS channel and RIS-user channel as $\tilde{H}_1$ and $\tilde{H}_2$. They are given by $\tilde{H}_1 = H_1 + \Delta H_1$ and $\tilde{H}_2 = H_2 + \Delta H_2$, where ΔH_1 and ΔH_2 are the estimation errors for the BS-RIS channel and RIS-user channel. It is assumed that both ΔH_1 and ΔH_2 are modeled as zero-mean complex Gaussian random matrices, i.e., $\Delta H_1 \sim \mathcal{CN}(0, \Sigma_{H_1})$, $\Delta H_2 \sim \mathcal{CN}(0, \Sigma_{H_2})$, [11]. It is assumed that $\tilde{H}_1$ and $\tilde{H}_2$ are available at the BS beamforming design. This assumption is consistent with the conventional optimization-based beamforming algorithms [2], [3], [12], [13]. With the channel estimation error, (2) can be reformulated as

$$\text{SINR}_k = \frac{\|\tilde{h}_{2,k}^T \Phi \tilde{H}_1 w_k\|^2}{\sum_{j=1, j\neq k}^{K} \|\tilde{h}_{2,k}^T \Phi \tilde{H}_1 w_j\|^2 + \sigma^2}. \tag{3}$$

The goal is to select the beamforming vectors and RIS phases such that the sum SE is maximized. Hence, the optimization problem can be formulated as

$$\max_{\Phi, W} \quad \sum\nolimits_{k=1}^{K} \log_2(1 + \text{SINR}_k) \tag{4}$$

$$\text{s.t.} \quad \|W\|_F^2 \leq P_{\text{tx}}, \tag{5}$$

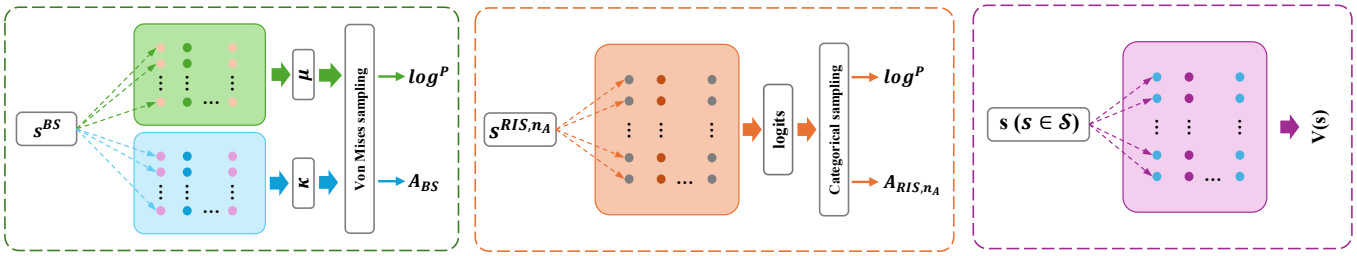
$$\Phi_{ii} \in \{0, 2\pi/2^M, \cdots, 2\pi(2^M - 1)/2^M\}, \quad \forall\, i \tag{6}$$

In the next section, we propose to use a sequential multi-agent A2C algorithm to solve (4).

III. Sequential multi-agent A2C

A. Algorithm Structure

The discrete nature of (4) makes gradient-based optimization inapplicable. To address this challenge, we adopt a DRL framework based on a multi-agent A2C architecture. We define the state space $\mathcal{S}$ as the joint configuration of the RIS phase shifts and the BS precoding matrix, while the action space $\mathcal{A}$ corresponds to adjusting these configurations. The reward function $\mathcal{R}$ is defined as the instantaneous sum spectral efficiency given in (1). At the beginning of training, both the RIS phase configuration and the BS precoder are randomly initialized, and actions are subsequently generated according to the current policy. Due to the extremely large size of the state and action spaces, the RIS is partitioned into N_A sub-arrays, and a separate agent is assigned to each sub-array to learn its local phase configuration. In addition, two agents are employed to learn the BS precoding matrix phases and amplitudes. The resulting architecture follows a centralized-training and decentralized-execution (CTDE) paradigm, in which all agents share a central critic during training, while decisions are executed independently at deployment. To improve training stability under discrete RIS phase constraints, we adopt an A2C-based sequential learning strategy, where only one RIS agent is activated and updated per episode, while the remaining sub-arrays remain fixed [14]. This sequential update mechanism mitigates instability caused by strong coupling among RIS sub-arrays. The overall network architecture consists of a BS actor, multiple RIS actors, and a shared central critic, as illustrated in Fig. 1. The BS actor



(a) Actor network of the BS agent (b) Actor network of the n_Ath RIS agent (c) Central critic network

Fig. 1: (a)-(b) The actor network structure of the BS agent and an RIS agent. (c) Central critic network structure.

Algorithm 1: Sequential Multi-agent A2C

Input : s, σ^2, $\tilde{H}_1$, $\tilde{H}_2$, N_A, N_R, N_T, K, $\mathcal{R}$
Initialize $\mathcal{R} = 0$;
for $i \leftarrow 1$ **to** Max episode **do**
  Calculate the active agent index using i and N_A;
  Reset $\mathcal{R}$ to 0;
  for $j \leftarrow 1$ **to** Max timesteps **do**
    Extract s^{RIS,n_A} from s and predict the action for the activated sub-array through the RIS actor network;
    Extract s^{BS} from s and predict the BS precoding matrix phase and amplitude using two parallel BS actor networks;
    Use s to predict the state value through the central critic network;
    Combine the actions predicted from the activated RIS actor network and the BS actor network, and transit to the next state;
    Update $\mathcal{R}$;
  Update the activated RIS actor network and the BS actor network through the advantage function and the logarithmic probabilities;
  Update the central critic network through the Huber loss;

learns the precoding phases and amplitudes, while each RIS actor learns the phase configuration of its associated sub-array. The BS precoding matrix is learned using two parallel actor networks: one dedicated to phase prediction and the other dedicated to amplitude prediction. This separation is adopted because phase and amplitude lie in different domains and exhibit distinct learning characteristics. By decoupling the two, the proposed framework avoids unnecessary coupling in the action space and improves training stability. The central critic evaluates the global state and provides value estimates that enable coordination among agents.

After introducing the actor and central critic network structures, we then move forward to describe the flow of the proposed algorithm, which is given in Algorithm 1. The activated RIS agent index is defined as: active RIS agent index = i mod N_A. The loss functions for the activated RIS actor network and the BS actor network are given by

$$\mathcal{L}_{RIS,n_A}=\sum_{j=1}^{N_1}-\log(P(A_j^{RIS,n_A}|s_j^{RIS,n_A}))(\mathcal{R}_j-V(s_j)), \tag{7}$$

$$\mathcal{L}_{BS}=\sum_{j=1}^{N_1}-\log(P(A_j^{BS}|s_j^{BS}))(\mathcal{R}_j-V(s_j)), \tag{8}$$

where N_1 denotes the number of max timesteps. In (7) and (8), bootstrapping is disabled in the advantage function (e.g., the term $\gamma V(s_{j+1})$ is not involved) to reduce the bias and improve the learning stability. The loss function for the central critic is defined as

$$\mathcal{L}_{\text{central critic}}=\begin{cases}\sum_{j=1}^{N_1}\frac{1}{2}(R_j-V(s_j))^2 & \text{if } |R_j-V(s_j)|<1\\ \sum_{j=1}^{N_1}\left(|R_j-V(s_j)|-\frac{1}{2}\right) & \text{otherwise.}\end{cases} \tag{9}$$

We move forward to analyze the computational complexity of the proposed algorithm and compare it with the benchmark.

B. Implementation Details

The actor and critic networks are implemented as fully connected multilayer perceptrons (MLPs). Each network consists of one input layer, one hidden layer, and one output layer. For both the actor and critic networks, the hidden-layer dimension is set to twice the dimension of the corresponding input layer, which provides a good balance between representation capability and training stability. ReLU activation is applied to all hidden layers. For the RIS actor network, the output layer produces unnormalized logits corresponding to the discrete phase indices of each RIS element within the activated sub-array. These logits are passed through a softmax function to form a categorical distribution, from which a discrete action index is sampled for each RIS element. The sampled action indices are then converted into physical RIS phase values according to the predefined phase resolution, i.e., each index is mapped to one of the allowable phase shifts in (5). For the BS precoding matrix, two parallel actor networks are employed to separately learn the phase and amplitude of each precoding entry. The BS phase actor network predicts the parameters of a circular probability distribution that is used to model the phase of each entry in the BS precoding matrix. During training, the BS precoder phase action is obtained by sampling from this distribution, ensuring that the resulting phase lies within the interval $[0, 2\pi]$. During deployment, the mean phase values are directly used. The BS amplitude actor network predicts the

amplitude of each entry in the BS precoding matrix. A Softplus activation function is applied at the output layer to guarantee non-negative amplitude values. During training, the amplitude action is sampled from a Gaussian distribution centered at the predicted mean to encourage exploration. The sampled amplitude values are clipped to a predefined range to maintain numerical stability. The log-probability of the sampled amplitude action is incorporated into the policy gradient update jointly with the phase action. During deployment, the predicted mean amplitude values are directly used. The critic network outputs a scalar state-value estimate using a ReLU activation at the output layer. All networks are trained using the Adam optimizer with a learning rate of 0.01 and a discount factor of 0.99. During training, only one RIS agent is activated per episode following the sequential learning strategy described previously, while the remaining RIS agents remain fixed. The central critic network is updated at every timestep using the Huber loss.

C. Complexity Analysis

We adopt the zero-forcing (ZF), minimum mean square error (MMSE), and single-agent A2C algorithms as the benchmarks, since ZF and MMSE are widely used in MIMO beamformer design [4], [12]. The computational complexities of the proposed algorithm, ZF, MMSE, and single-agent A2C are then analyzed. Before presenting the complexity comparison, we briefly describe how ZF (MMSE also in the same way) is applied when N_R is extremely large and each RIS element has a discrete phase. ZF, as a linear precoder, is used to design the precoding matrix. To address the joint beamforming problem, it alternately optimizes the BS precoding matrix and the RIS phase configuration. However, when optimizing the RIS phases, the standard gradient descent method suffers from poor scalability due to the large RIS dimension and discrete phase constraints. To mitigate this issue, we employ block coordinate descent (BCD)[1]. For the discrete phase constraint, the RIS phase configuration is first optimized via BCD under continuous phase assumptions, and the result is subsequently quantized into M bits.

Next, we analyze the computational complexity of the proposed algorithm. We focus on the computational complexity calculation for the forward propagation of the networks and SINR calculation because these happen per timestep, while the back propagation happens per episode. The complexities of the forward propagation for the RIS actor, BS actor, and the central critic are $\mathcal{O}((KN_T)^2)$, $\mathcal{O}((N_R/N_A)^2)$, and $\mathcal{O}((KN_T + N_R/N_A)^2)$, respectively. The complexity of the SINR calculation is $\mathcal{O}(KN_TN_R + K^2N_T)$. We define the number of the maximum episode as N_2. The computational complexity of the proposed algorithm can be summarized as $\mathcal{O}(N_1N_2((KN_T)^2 + (N_R/N_A)^2 + (KN_T + N_R/N_A)^2 + KN_TN_R + K^2N_T))$. We can simplify the expression as N_T and K are much smaller compared to N_R, N_1, and N_2. The complexity expression then becomes $\mathcal{O}(N_1N_2(N_R/N_A)^2)$. As we alternatively optimize the precoding matrix and the RIS phase configuration in the ZF/MMSE algorithm, we define the outer iteration as N_3 and the inner iteration as N_4. The computational complexity of the ZF/MMSE algorithm is $\mathcal{O}(N_3N_4(N_R/N_A))$. The complexity of the single-agent A2C is then given as $\mathcal{O}(N_1N_2N_R^2)$. From the complexity analysis, the proposed algorithm exhibits higher computational complexity than ZF and MMSE owing to the quadratic term, representing a trade-off between performance gains and computational cost.

[1]When optimizing one block (equivalent to a sub-array of the RIS), other blocks remain fixed.

Table I: Parameter setting

Parameter	Value
Learning rate	0.01
Reward discount factor	0.99
N_T, N_R	16, [128, 256, 512 1024,2048]
N_1, N_2, N_3, N_4	2000, 5000 2000, 1000
N_A	[1, 2, 4, 8]
K, M, λ, β, ρ	3, [1,2], 100 mm, 0.1371, 0.9
Cluster radius	[0.5, 2] m
Σ_{H_1}	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]I
Σ_{H_2}	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]I
P_{Tx}, σ^2, a	K mW, 1×10^{-6} mW, 50 m

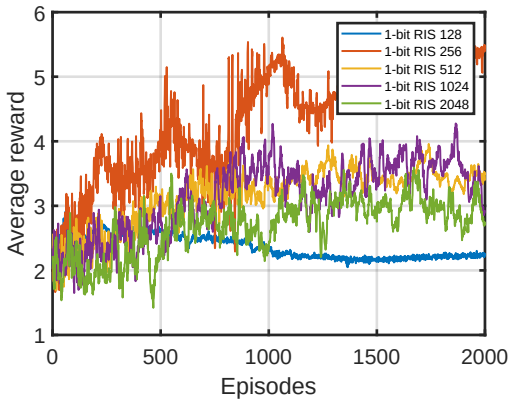
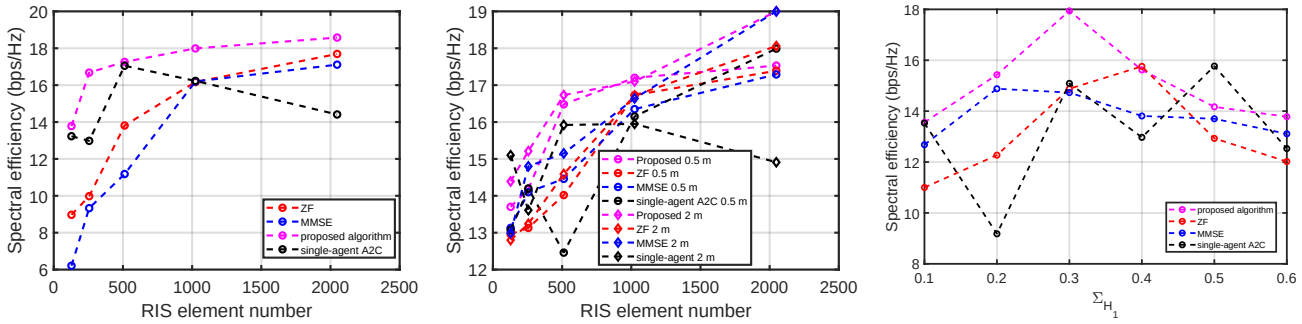


Fig. 2: Learning curve of the proposed algorithm.

IV. Simulation Results

We first give the parameter settings for the proposed algorithm and the ZF algorithm in Table I. In Table I, it is noted that some of the parameters are sequences, such as N_R, N_A, and Σ_{H_1}. This is because different settings are being simulated. Fig. 2 shows the learning curve of the proposed algorithm. It is noted that the average reward for each case gradually converges as the episode number increases. This indicates that the proposed algorithm is stable and gradually learns to reach the optimal precoding matrix and the RIS configuration.

Fig. 3(a) to Fig. 3(c) illustrate the performance of the proposed algorithm with respect to the RIS element number under uniformly distributed users, clustered users, and imperfect CSI conditions. In Fig. 3(a), the proposed algorithm consistently outperforms the ZF and MMSE benchmarks across all RIS element numbers and phase resolutions. However, the performance gap between the proposed algorithm and the



(a) (b) (c)

Fig. 3: (a) Achievable SE at different RIS sizes with 1-bit phase resolution, uniformly distributed user, and perfect CSI. (b) Achievable SE at different RIS sizes with various cluster radii, 1-bit phase resolution, and perfect CSI. (c) Achievable SE at different channel estimation errors with $N_R = 512$, 1-bit phase resolution, and uniformly distributed users.

benchmarks gradually narrows as the number of RIS elements increases. This behavior is mainly attributed to the exponential growth of the state and action spaces, which reduces the learning efficiency of the algorithm. As illustrated in the figure, the proposed algorithm is therefore most suitable for scenarios with 1-bit phase resolution and RIS element numbers up to 2048. Compared with the single-agent A2C, the proposed algorithm achieves superior performance at all RIS element numbers. In contrast, the single-agent A2C gradually fails to learn the optimal configuration within the given N_1 and N_2, confirming the necessity of adopting a multi-agent A2C for extremely large RISs..

Fig. 3(b) illustrates the impact of channel correlation at different RIS element numbers. As expected, higher channel correlation leads to lower achievable SE. In Fig. 3(c), the achievable SE of the proposed algorithm first increases and then decreases as Σ_{H_1} grows. This behavior occurs because moderate channel estimation errors act as artificial noise, alleviating overfitting and enhancing robustness to imperfect CSI. A similar trend is observed for the ZF and MMSE algorithms, where small estimation errors (e.g., $\Sigma_{H_1} = 0.2I$ for MMSE) initially improve SE by partially compensating for the degradation caused by quantization.

V. Conclusion

In this paper, we propose an algorithm that can jointly optimize the precoding matrix and the RIS phase configuration in a MU-MISO scenario when the RIS size is extremely large. We also showed the performance of the proposed algorithm in the presence of channel correlation and imperfect CSI in terms of achievable SE. Throughout simulations, it is shown that the proposed algorithm is also robust to the imperfect CSI. The proposed algorithm also has a better performance compared to the benchmarks.

References

[1] X. Zhang, D. Xu, J. Wang, S. Song, D. W. K. Ng, and M. Debbah, "Fluid antenna meets RIS: Random matrix analysis and two-timescale design for multi-user communications," *IEEE Journal on Selected Areas in Communications*, 2025. Early access.

[2] 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 Trans. Commun.*, vol. 70, no. 4, pp. 2404–2416, 2022.

[3] 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, 2022.

[4] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, 2020.

[5] A. Faisal, I. Al-Nahhal, O. A. Dobre, and T. M. Ngatched, "Deep reinforcement learning for RIS-assisted FD systems: Single or distributed RIS?," *IEEE Commun. Lett.*, vol. 26, no. 7, pp. 1563–1567, 2022.

[6] A. Abdallah, A. Celik, M. M. Mansour, and A. M. Eltawil, "Multi-agent deep reinforcement learning for beam codebook design in RIS-aided systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7983–7999, 2024.

[7] W. Tang, X. Chen, M. Z. Chen, J. Y. Dai, Y. Han, M. Di Renzo, S. Jin, Q. Cheng, and T. J. Cui, "Path loss modeling and measurements for reconfigurable intelligent surfaces in the millimeter-wave frequency band," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6259–6276, 2022.

[8] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "RIS-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8627–8640, 2021.

[9] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 581–585, 2019.

[10] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, 2001.

[11] E. Björnson, L. Sanguinetti, and M. Debbah, "Massive MIMO with imperfect channel covariance information," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 974–978, IEEE, 2016.

[12] X. Ma, S. Guo, H. Zhang, Y. Fang, and D. Yuan, "Joint beamforming and reflecting design in reconfigurable intelligent surface-aided multi-user communication systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3269–3283, 2021.

[13] Q. Sun, Y. Wu, X. Chen, and J. Zhang, "SLNR-based joint RIS-UE association and beamforming design for multi-RIS aided wireless communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8660–8670, 2024.

[14] Y. Zang, J. He, K. Li, H. Fu, Q. Fu, and J. Xing, "Sequential cooperative multi-agent reinforcement learning," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 485–493, 2023.