Aalborg Universitet

Joint User Pairing and Beamforming Design of Multi-STAR-RISs-Aided NOMA in the Indoor Environment via Multi-Agent Reinforcement Learning

Yu Min Park[1], Yan Kyaw Tun[2], and Choong Seon Hong[1]

[1]Department of Computer Science and Engineering, Kyung Hee University, Yongin, 17104, Republic of Korea
[2]Department of Electronic Systems, Aalborg University, . C. Meyers Vænge 15, 2450 København
Email: {yumin0906, cshong}@khu.ac.kr, ykt@es.aau.dk.

Abstract—The development of sixth-generation (6G)/Beyond Fifth-Generation (B5G) wireless networks, which have requirements that go beyond current 5G networks, is gaining interest from academia and industry. However, to increase 6G/B5G network quality, conventional cellular networks that rely on terrestrial base stations are constrained geographically and economically. Meanwhile, Non-Orthogonal Multiple Access (NOMA) allows multiple users to share the same resources, which improves the spectral efficiency of the system and has the advantage of supporting a larger number of users. Additionally, by intelligently manipulating the phase and amplitude of both the reflected and transmitted signals, Simultaneously Transmitting and Reflecting RISs (STAR-RISs) can achieve improved coverage, increased spectral efficiency, and enhanced communication reliability. However, STAR-RISs must simultaneously optimize the amplitude and phase shift corresponding to reflection and transmission, which makes the existing terrestrial networks more complicated and is considered a major challenging issue. Motivated by the above, we study the joint user pairing for NOMA and beamforming design of Multi-STAR-RISs in an indoor environment. Then, we formulate the optimization problem with the objective of maximizing the total throughput of mobile users (MUs) by jointly optimizing the decoding order, user pairing, active beamforming, and passive beamforming. However, the formulated problem is a mixed-integer non-linear programming (MINLP). To address this challenge, we first introduce the decoding order for NOMA networks. Next, we decompose the original problem into two subproblems, namely: 1) MU pairing and 2) Beamforming optimization under the optimal decoding order. For the first subproblem, we employ correlation-based K-means clustering to solve the user pairing problem. Then, to jointly deal with beamforming vector optimizations, we propose Multi-Agent Proximal Policy Optimization (MAPPO), which can make quick decisions in the given environment owing to its low complexity. Finally, simulation results prove that our proposed MAPPO algorithm is superior to Proximal Policy Optimization (PPO) and Advanced Actor-Critic (A2C) by a maximum of 1% and 6%, respectively.

Index Terms—STAR-RIS, NOMA network, indoor environment, reinforcement learning, multi-agent proximal policy optimization.

I. INTRODUCTION

There is a growing interest among both academic and industrial circles regarding the advancement of sixth-generation (6G)/Beyond Fifth-Generation (B5G) wireless networks. The requirements are to address the more stringent demands that surpass those of the existing 5G networks. These requirements include achieving ultra high data rates and energy efficiency, ensuring global coverage and connectivity, and attaining extremely high reliability and low latency [1], [2]. However, Existing cellular networks that rely on terrestrial base stations have economic and geographic limitations for improving network quality. Meanwhile, Reconfigurable Intelligent Surfaces (RISs) are new communication equipment for future next-generation wireless communication network performance improvement [3]. RIS is a plane reflector composed of multiple low-cost reconfigurable passive communication elements. The corresponding element may reconstruct the radio signal propagation by manually adjusting the amplitude and phase appropriately. Therefore, RIS may be deployed in a wireless network concentration area to improve communication quality with economical and low energy consumption. In [4], the authors minimized the latency by improving the communication throughput of ground users through RIS located in buildings in full-duplex communication environments. In addition, the work in [5] provided a study that provides wireless communication for high-speed trains by further maximizing line-of-sight (LoS) by mounting RIS on UAVs. However, the disadvantage of existing RIS is that it only has a reflection function, so the transmitter and receiver must be on the same side. This topological constraint limits the flexibility of employing existing RISs.

To overcome this, unlike the RIS described above, Simultaneously Transmitting and Reflecting (STAR-RIS) can provide communication services to both parties by enabling simultaneous transmission and reflection of incident signals. STAR-RIS is largely classified into three types depending on how transmission and reflection signals are controlled [6]. The types of STAR-RIS are classified into Energy Splitting (ES), which controls the energy for transmission and reflection signals, Mode Switching (MS), which converts the mode of each element to determine the signal method, and Time Switching (TS), which changes the signal method over time. Among these types, ES-type STAR-RIS has high flexibility but has the disadvantage of optimizing energy variables along with the phase variables of each signal. By intelligently manipulating the phase and amplitude of both the reflected and transmitted signals, STAR-RISs can achieve improved coverage, increased spectral efficiency, and enhanced communication reliability [7]. Meanwhile, in [8], a study was conducted on how to optimize a number of STAR-RISs. However, optimization for the Non-Orthogonal Multiple Access (NOMA) networks was left as an assumption without considering it.

NOMA is a multi-access technology that allows multiple users to access the same frequency band simultaneously without the need for orthogonal resource allocation, such as the existing Orthogonal Frequency Division Multiple Access (OFDMA) or Code Division Multiple Access (CDMA) systems [9]. NOMA can separate signals that overlap each other in the same time and frequency domain from the receiver using continuous interference cancellation (SIC) or other signal processing techniques. This allows multiple users to share the same resources, which improves the spectral efficiency of the system and has the advantage of supporting a larger number of users. However, there are currently no studies that have carefully addressed the Multi-STAR-RISs-aided NOMA networks. In this paper, we try to fill this gap and our major contributions may be summed up as follows:

- We propose a novel network architecture for indoor environment wireless communication where the access point provide services to multiple mobile users with the aid of multiple STAR-RIS (Multi-STAR-RISs) in NOMA networks.
- Drawing upon the suggested system architecture, we formulate an optimization problem to address the user pairing, AP active beamforming, as well as passive beamforming included amplitude and phase shift of STAR-RISs.
- To tackle the aforementioned problem, we decompose the main problem into two sub-problems by using the block coordinate descent (BCD) method, and then solve each sub-problem, iteratively.
- Additionally, we propose correlation-based clustering to address the user pairing problem. In addition, the beamforming opimization problem was solved by applying multi-agent reinforcement learning (MARL) to optimize beamforming vectors.
- Finally, we perform in-depth simulations to show that our suggested approach performs better than the baseline algorithms. The simulation results show that the proposed MAPPO performance is better than comparison algorithms to a maximum of 6%.

The subsections of the paper are organized in the following manner. The system model and problem formulation are outlined in Section II. Subsequently, Section III provides a detailed description of the method that has been proposed. The details of the implementation and simulation outcomes are expounded upon in Section IV, while the conclusion of the paper comes in Section V.

II. SYSTEM MODEL & PROBLEM FORMULATION

A. System Model Overview

As illustrated in Fig. 1, we consider a downlink communication in Multi-STAR-RISs-aided NOMA networks for indoor environments, which consists of an access point (AP) with N_b antenna, a set $\mathcal{L}$ of L STAR-RISs, and a set $\mathcal{U}$ of U mobile users (MUs) with a single antenna. We also consider STAR-RISs with $M = M_h M_v$ elements, where M_h and M_v denote
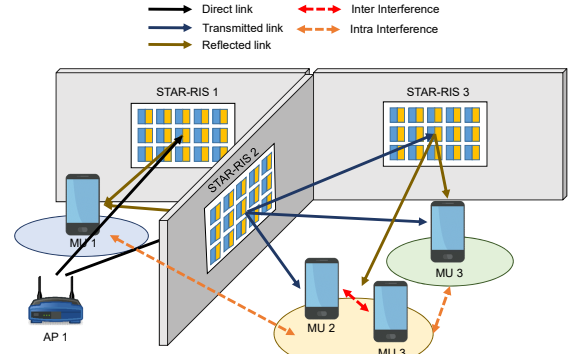

Fig. 1: Illustration of our system model.

the number of elements along the vertical and the horizontal, respectively. The locations of AP, center of STAR-RIS $l \in \mathcal{L}$ and MU $u \in \mathcal{U}$ are $p_0 = [x_0, y_0, h_0]^T$, $p_l = [x_l, y_l, h_l]^T$ and $p_u = [x_u, y_u]^T$. As shown in Fig. 1, we assume that the environment under consideration consists of several rooms, and some of the walls of the rooms are composed of START-RIS. Downlink communication between APs and MUs in different rooms is possible through walls composed of STAR-RISs. In this paper, we assume that perfect channel state information (CSI) is available at the AP to investigate the performance gain of Multi-STAR-RISs. The STAR-RIS adopts an energy splitting (ES) protocol, where each element can operate simultaneous transmission and reflection modes. For given transmission and reflection amplitude coefficients, the signals incident upon each element is split into transmitted and reflected signals having different energy. In a practical implementation, the amplitude and phase shift coefficients of each element for transmission and reflection will be jointly optimized for achieving diverse design objectives in wireless networks.

In our system model, the reflection and transmission surfaces of STAR-RIS are considered differently depending on the location of the AP. Thus, we assume that clockwise surfaces are forward and otherwise backward. Therefore, the forward and backward side passive beamforming vectors of STAR-RIS l are given by

$$\mathbf{\Phi}_l^F = \text{diag}\left\{ \sqrt{\beta_{l_1}^F} e^{j\theta_{l_1}^F}, \sqrt{\beta_{l_2}^F} e^{j\theta_{l_2}^F}, ..., \sqrt{\beta_{l_M}^F} e^{j\theta_{l_M}^F} \right\}, \quad (1)$$

$$\mathbf{\Phi}_l^B = \text{diag}\left\{ \sqrt{\beta_{l_1}^B} e^{j\theta_{l_1}^B}, \sqrt{\beta_{l_2}^B} e^{j\theta_{l_2}^B}, ..., \sqrt{\beta_{l_M}^B} e^{j\theta_{l_M}^B} \right\}, \quad (2)$$

where $\beta_{l_m}^F$ and $\beta_{l_m}^B$ are amplitudes of the forward and backward sides for STAR-RIS l's element m. $\theta_{l_m}^F$ and $\theta_{l_m}^B$ are phase shift coefficients of the forward and backward sides for STAR-RIS l's element m.

By considering the path loss model for the indoor hotspot scenario, as presented in 3GPP TR 38.901 version 16.1.0 Release 16, the path losses, at the reference distance of the LoS link and NLoS link, are given by

$$PL^{\text{LoS}} = 32.4 + 17.3 \log_{10}(d) + 20 \log_{10}(f), \quad (3)$$

$$PL^{\text{NLoS}} = 32.4 + 31.9 \log_{10}(d) + 20 \log_{10}(f), \quad (4)$$

where d is the distance between the transmitter and the receiver, and f is the sub-carrier frequency. The channel between an AP and a MU can be modeled as a Rician channel, which includes one LoS path and a number of NLoS paths. Therefore, the channel gain h_u from AP to MU u can be formulated as

$$h_u = \sqrt{\frac{\kappa}{\kappa+1}} h_u^{\text{LoS}} + \sqrt{\frac{1}{\kappa+1}} h_u^{\text{NLoS}}, \quad (5)$$

where κ is the Rician factor, $h_u^{\text{LoS}} \in \mathbb{C}^{N_b \times 1}$ and $h_u^{\text{NLoS}} \in \mathbb{C}^{N_b \times 1}$ are LoS and NLoS channel gains between AP and MU u, where $\mathbb{C}^{N_b \times 1}$ denotes a complex matrix of size $N_b \times 1$. Similarly, we can define the channel gain g_l from AP to the STAR-RIS l and the channel gain $g_{l,u}$ from STAR-RIS l to MU u as follows:

$$g_l = \sqrt{\frac{\kappa}{\kappa+1}} g_l^{\text{LoS}} + \sqrt{\frac{1}{\kappa+1}} g_l^{\text{NLoS}}, \quad (6)$$

$$g_{l,u} = \sqrt{\frac{\kappa}{\kappa+1}} g_{l,u}^{\text{LoS}} + \sqrt{\frac{1}{\kappa+1}} g_{l,u}^{\text{NLoS}}, \quad (7)$$

where $g_l^{\text{LoS}} \in \mathbb{C}^{N_b \times M}$ and $g_l^{\text{NLoS}} \in \mathbb{C}^{N_b \times M}$ are LoS and NLoS channel gains from AP b to STAR-RIS l, and $g_{l,u}^{\text{LoS}} \in \mathbb{C}^{M \times 1}$ and $g_{l,u}^{\text{NLoS}} \in \mathbb{C}^{M \times 1}$ are LoS and NLoS channel gains from STAR-RIS l to MU u. Hence, the combined channel gain from AP to MU u is given by

$$\hat{h}_u = \begin{cases} h_u + \sum_{l \in \mathcal{L}} \left\{ c_l^b (c_u^{l_F} g_l \boldsymbol{\Phi}_l^F g_{l,u} + c_u^{l_B} g_l \boldsymbol{\Phi}_l^B g_{l,u}) \right\} \\ \qquad\qquad\qquad\qquad\qquad \text{if} \quad c_u^b = 1, \\ \sum_{l \in \mathcal{L}} \left\{ c_l^b (c_u^{l_F} g_l \boldsymbol{\Phi}_l^F g_{l,u} + c_u^{l_B} g_l \boldsymbol{\Phi}_l^B g_{l,u}) \right\} \\ \qquad\qquad\qquad\qquad\qquad \text{if} \quad c_u^b = 0, \end{cases} \quad (8)$$

where $c_u^b, c_l^b, c_u^{l_F}, c_u^{l_B} \in \{0,1\}$ are adjacency indicators between AP and MU u, between AP b and STAR-RIS l, between a forward side of STAR-RIS l_F and MU u, and between a backward side of STAR-RIS l_B and MU u.

In NOMA networks, intra-cluster and inter-cluster interference can be considered, where intra-cluster interference occurs between MUs grouped in the same cluster of AP, and inter-cluster interference occurs between MUs grouped in different clusters of AP. Thus, the MUs associated with AP are further clustered into K groups. Therefore, we define $\gamma_{k,u} \in \{0,1\}$ as a user pairing factor, where $\gamma_{k,u} = 1$ if MU u is involved in cluster k of AP, otherwise $\gamma_{k,u}^b = 0$. Moreover, let $\omega_b = \{w_1, w_2, ..., w_K\}$ be the active beamforming vector of AP. Therefore, the received signal of MU u associated with AP in cluster k can be given by

$$y_{k,u} = \hat{h}_u [\omega_{b,k} (\gamma_{k,u} p_0 s_{k,u} + \sum_{u' \neq u}^{U} \gamma_{k,u'} p_0 s_{k,u'}) \\ + \sum_{k' \neq k}^{K} \sum_{u'}^{U} \gamma_{k',u'} \omega_{k'} p_0 s_{k',u'}] + N_0, \quad (9)$$

where p_0 is the power allocation coefficient of each MU associated with AP. We assume that MUs connected to the AP use power equally. Therefore, the power allocation coefficient for AP satisfies $p_0 = 1/|\mathcal{U}|$. $s_{k,u}$ denotes the signal transmitted by AP for MU u in cluster k, and N_0 is the Additive White Gaussian Noise (AWGN) with variance σ^2. Without loss of generality, for any cluster $k \in \mathcal{K}$, $\delta(u)$ denotes the MU index that corresponds to MU u decoded order in the Successive Interference Cancellation (SIC) procedure. For cluster k, after applying the SIC decoding procedure [10], the intra-cluster and inter-cluster powers of MU u associated with AP on cluster k can be calculated as

$$I_{k,u}^{\text{intra}} = |\hat{h}_u \omega_k|^2 \sum_{\delta_k(u') > \delta_k(u)}^{U} \gamma_{k,u'} p_0, \quad (10)$$

$$I_{k,u}^{\text{inter}} = \sum_{k' \neq k}^{K} \sum_{u'}^{U} \gamma_{k',u'} |\hat{h}_u \omega_{k'}|^2, \quad (11)$$

Accordingly, the received signal-to-interference-plus-noise ratio (SINR) of MU u associated with AP in cluster k is given by

$$\text{SINR}_{k,u} = \frac{|\hat{h}_u \omega_k|^2 \gamma_{k,u} p_0}{I_{k,u}^{\text{intra}} + I_{k,u}^{\text{inter}} + \sigma^2}, \quad (12)$$

For any two MUs v and u with decoding order $\delta_k(v) > \delta_k(u)$ in the same AP and cluster k, the received SINR of the signal $s_{k,u}$ at the MU v is given by

$$\text{SINR}_{k,v \to u} = \frac{|\hat{h}_{b,v} \omega_k|^2 \gamma_{k,v} p_0}{I_{k,v \to u}^{\text{intra}} + I_{k,v \to u}^{\text{inter}} + \sigma^2}, \quad (13)$$

where $I_{k,v \to u}^{\text{intra}} = |\hat{h}_v \omega_k|^2 \sum_{\delta_k(u') > \delta_k(u)}^{U} \gamma_{k,u'} p_0$ is the intra-cluster interference power of the signal $s_{k,u}$ at MU v. $I_{k,v \to u}^{\text{inter}} = \sum_{k' \neq k}^{K} \sum_{u'}^{U} \gamma_{k',u'} |\hat{h}_v \omega_{k'}|^2$ is the inter-cluster interference power of the signal $s_{k,u}$ at MU v. It is worth pointing out that given a decoding order, to guarantee the SIC performed successfully, the condition $\text{SINR}_{k,v \to u} \geq \text{SINR}_{k,u}$ with $\delta_k(v) > \delta_k(u)$ must be guaranteed. Therefore, the achievable data rate of MU u associated with AP in cluster k is calculated as

$$R_{k,u} = \log_2 (1 + \text{SINR}_{k,u}), \quad (14)$$

B. Problem Formulation

In this subsection, we define the detailed problem formulation based on the proposed system model. This work's major goal is to maximize the achievable sum rate of U MUs (considered as a network utility), while jointly optimizing user pairing factor $\boldsymbol{\gamma}$, decoding order $\boldsymbol{\delta}$, active beamforming $\boldsymbol{\omega}$, and passive beamforming $\boldsymbol{\Phi} = \{\Phi^F, \Phi^B\}$ of STAR-Multi-RISs.

Therefore, we can define our optimization problem as follows:

$$\textbf{P1: } \max_{\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\omega},\boldsymbol{\Phi}} \quad \sum_{k=1}^{K}\sum_{u=1}^{U} R_{k,u} \tag{15a}$$

$$\text{subject to} \quad R_{k,u} \geq R_u^{\min}, \ \forall u \in \mathcal{U}, \tag{15b}$$

$$\text{SINR}_{k,v \to u} \geq \text{SINR}_{k,u}, \ \delta_k(v) > \delta_k(u), \tag{15c}$$

$$\gamma_{k,u} \in \{0,1\}, \ \forall k \in \mathcal{K}, \ \forall u \in \mathcal{U}, \tag{15d}$$

$$\sum_{k=1}^{K}\|w_k\|^2 \leq P_{\max}, \tag{15e}$$

$$\sqrt{\beta_{l_m}^t}, \sqrt{\beta_{l_m}^e} \in [0,1], \ \forall l \in \mathcal{L}, \ \forall m \in \mathcal{M}, \tag{15f}$$

$$\beta_{l_m}^t + \beta_{l_m}^r = 1, \ \forall l \in \mathcal{L}, \ \forall m \in \mathcal{M}, \tag{15g}$$

$$\theta_{l_m}^t, \theta_{l_m}^r \in [0,2\pi), \ \forall l \in \mathcal{L}, \ \forall m \in \mathcal{M}, \tag{15h}$$

where $R_u^{\min}$ is the minimum rate requirement of each MU. Constraint (15b) guarantees the QoS requirement of each MU, and constraint (15c) ensures the success of the SIC decoding. Furthermore, constraint (15d) represents the binary variables. Constraint (15e) ensures the power budget constraint of each AP. Finally, Constraints (15f) to (15h) indicate the requirements of each reflecting and transmission element in STAR-RIS. To solve this proposed problem, we provide a solution approach in the next section.

III. SOLUTION APPROACH

As our proposed problem (15) is a mixed-integer non-linear programming (MINLP), which is NP-hard due to complexity. To solve this problem, we first introduce the decoding order for NOMA networks. Next, we decompose the main problem into two sub-problems by using the block coordinate descent (BCD) method, and then solve each sub-problem iteratively until the convergence criteria meet.

A. Decoding Order

Prior to handling the pairing and beamforming optimization problems, the decoding order must be addressed because it is a important one for the Multi-STAR-RIS in NOMA networks. Therefore, we propose a scheme to obtain the optimal decoding order by the following lemma.

Lemma 1. *Given the active beamforming vector $\boldsymbol{\omega}$ and the passive beamforming vector $\boldsymbol{\Phi}$, The decoding order for cluster k with $|C_k|$ MUs in AP is defined as*

$$g_{\delta_k(1)}^k \leq g_{\delta_k(2)}^k \leq \cdots \leq g_{\delta_k(|C_k|)}^k, \tag{16}$$

where $g_{\delta_k(j)}^k = \frac{|\hat{h}_u \omega_k|^2}{\sum_{k' \neq k}^{K}\sum_{u' \in C_{k'}}|\hat{h}_u \omega_{k'}|^2 + \sigma^2}$ is the equivalent-combined channel gain [11], [12].

Lemma 1 indicates that the decoding order for each cluster of the Multi-STAR-RIS in the NOMA system is a function of the active beamforming vectors $\boldsymbol{\omega}$, the passive beamforming vectors $\boldsymbol{\Phi}$.

Algorithm 1 Correlation-based K-means Clustering for MU Pairing

---

1: **Input:** the initial passive beamforming vector $\boldsymbol{\Phi}_0$
2: $C_{\text{pre}}, C = \emptyset$
3: Randomly select χ_k from $u \in \mathcal{U}$, $\forall k = 1,...,K$
4: **while** $C_{\text{pre}} \neq C$ **do**
5: $\quad C_{\text{pre}} \leftarrow C$
6: $\quad$ **for** $u \in \mathcal{U}$ with $u \neq \chi_k$, $\forall k = 1,...,K$ **do**
7: $\quad\quad k^* = \arg\max_{1 \leq k^* \leq K} C_{u,\chi_{k^*}}$
8: $\quad\quad C_{k^*} = C_{k^*} \cup \{u\}$
9: $\quad$ **end for**
10: $\quad$ Update χ_k according to (21), $\forall k = 1,...,K$
11: $\quad C_k = C_k \setminus u, \forall u \neq \chi_k \in C_k, \forall k = 1,...,K$
12: **end while**
13: **Output:** The optimal clustering vector $\mathbf{C}^* \to$ MU pairing vector $\boldsymbol{\gamma}^*$.

Proposition 1. *For any two users u and v belong to cluster k, if the decoding order of the two users satisfies*

$$\delta_k^{-1}(v) > \delta_k^{-1}(u), \tag{17}$$

where $\delta_k^{-1}(\cdot)$ is the inverse of mapping function $\delta_k(\cdot)$. Then, under the optimal decoding order, the following SIC condition is guaranteed:

$$\text{SINR}_{k,v \to u} \geq \text{SINR}_{k,u}. \tag{18}$$

According to **Proposition 1**, the constraint in (15c) can be removed under the optimal decoding order of the NOMA system. This operation will not affect the optimality of the problem (15). Furthermore, **Lemma 1** and **Proposition 2** guarantee that once the association, pairing, and beamforming vectors are determined, the optimal decoding order in each cluster is fixed [10], [11]. Therefore, we develop the optimal beamforming vectors for MUs in STAR-Multi-RIS NOMA system based on this observation.

B. Correlation-Based K-means Clustering for MU Pairing

MUs whose channels are highly correlated should be assigned to the same group to make full use of the multiplexing gain, while MUs whose channels are uncorrelated should be assigned to different groups to decrease the interference. We adapt the K-means clustering algorithm to implement the MU pairing in NOMA networks. K-means clustering is one way to divide given data into multiple partitions [13]. The K-means algorithm determines the cost function as the sum of squares at the center of each group and the group's distance from the data subject. Clustering is also performed by updating the group that each data object belongs to to minimize the value of this cost function. Therefore, we use the channel correlation between each MU for the cost function for MU pairing. The normalized channel correlation between MU i and MU j can be calculated as [14]

$$\text{Cor}_{i,j} = \frac{\hat{h}_i^{\text{H}} \hat{h}_j}{\|\hat{h}_i\|\|\hat{h}_j\|}. \tag{19}$$

Algorithm 2 Learning Process for Multi-Agent Proximal Policy Optimization (MAPPO)

1: **Initialize:** the initial network π_0^ω and π_0^Φ for agent of active and beamforming.
2: **for** episode$= 1, 2, ..., E$ **do**
3:      Initialize randomly each MU's position and calculate the optimal user pairing from III-B solution, i.e., γ^*.
4:      **for** time slot$= 1, 2, ..., N$ **do**
5:          Update observation $\mathcal{S}(n)$.
6:          Run policy $\mathcal{A}^\omega \sim \pi_{\theta_{\text{old}}}^\omega, \mathcal{A}^\Phi \sim \pi_{\theta_{\text{old}}}^\Phi$.
7:          Compute the common reward $\mathcal{R}(n)$.
8:          Save $(\mathcal{S}(n), \mathcal{A}(n), \mathcal{R}(n), \mathcal{S}(n+1))$ in memory of each agent.
9:      **end for**
10:     Compute advantage estimates $\left\langle \hat{A}_1^\omega, ..., \hat{A}_N^\omega \right\rangle$, $\left\langle \hat{A}_1^\Phi, ..., \hat{A}_N^\Phi \right\rangle$ based on (25).
11:     Optimize surrogate L^{PPO} wrt θ^ω, θ^Φ with minibatch from memory based on (26).
12:     $\theta_{\text{old}}^\omega \leftarrow \theta^\omega$, $\theta_{\text{old}}^\Phi \leftarrow \theta^\Phi$
13: **end for**
14: **Output:** Optimal networks $\pi_{\theta_{\text{opt}}}^\omega$, $\pi_{\theta_{\text{opt}}}^\Phi$.

In order to cluster MUs connected to each AP b into K_b, we select randomly K_b MUs assigned to clusters, $C = \{C_1, C_2, ..., C_K\}$, one by one. Then, the channel correlation between the unselected MU u' and the selected MU u is calculated based on (19), and the MU u' having the highest channel correlation is assigned to the cluster to which the MU u belongs. Thenceforth, each representative can be selected from each cluster. The representative of each cluster is updated as the one with the lowest correlation with the other clusters in order to further reduce the correlation of the channels between the various clusters. The correlation between a MU to the other clusters is the total normalized channel correlation between a MU to the MUs of the other clusters. The correlation between a MU u in the cluster C_k to the other clusters is defined as

$$\bar{\text{Cor}}_u = \sum_{l \neq k}^{K} \sum_{u' \in C_l} \text{Cor}_{u,u'}. \tag{20}$$

After that, the representative χ_k of the cluster C_k is updated as

$$\chi_k = \arg \min_{u \in C_k} \bar{\text{Cor}}_u. \tag{21}$$

Following the update of the representative for each cluster, the other MUs are subsequently reassigned to their respective clusters. The iteration is terminated when the representatives of the clusters remain unaltered. Finally, the optimal output $\mathbf{C}^*$ of the clustering vector, is transformed to the MU pairing vector γ^*. The Correlation-based K-means Clustering is described in Algorithm 1.

C. Multi-Agent Reinforcement Learning (MARL) Scheme

At the fixed MU association, pairing, and decoding order, we can rewrite the beamforming vectors optimization problem as follows:

$$\textbf{P1.1:} \underset{\boldsymbol{\omega}, \boldsymbol{\Phi}}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{u=1}^{U} R_{k,u} \tag{22a}$$

$$\text{subject to} \quad (15b), (15e) \sim (15h). \tag{22b}$$

However, the subproblem **P1.1** has the non-convex nature. Thus, it is challenging to solve using the existing optimization techniques. Thus, we propose a proximal policy optimization (PPO)-based multi-agent reinforcement learning method (MARL) to solve problem **P1.1**. PPO is a widely used reinforcement learning technique that is known for its simplicity in implementation and applicability across diverse situations [15]. Furthermore, it has demonstrated consistent and reliable performance. The utilization of the Proximal Policy Optimization (PPO) technique serves to streamline the intricate computational process associated with trust region policy optimization (TRPO). The Trust Region Policy Optimization (TRPO) algorithm aims to optimize a surrogate objective function in the following manner [16].

$$L_n^{\text{TRPO}}(\theta) = \hat{\mathbb{E}}_n \left[\frac{\pi_\theta(\mathcal{A}_n | \mathcal{S}_n)}{\pi_{\theta_{\text{old}}}(\mathcal{A}_n | \mathcal{S}_n)} \hat{A}_n \right] = \hat{\mathbb{E}}_n \left[r_n(\theta) \hat{A}_n \right], \tag{23}$$

where $r_n(\theta)$ denotes the probability ratio, $\mathcal{A}_n$ and $\mathcal{S}_n$ are an action and reward in time step n. The surrogate objective function L^{TRPO} of TRPO has a complicated formula expansion and must calculate the second derivative. Therefore, maximization of the surrogate objective function L^{TRPO} of TRPO would result in an unnecessarily large policy update in the absence of a constraint. Hence, in PPO, the limitations of TRPO were addressed by incorporating an approximation of the first derivative using the clipping technique. The following is the objective function to which the clipping is applied:

$$L_n^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_n \left[\min(r_n(\theta) \hat{A}_n, \text{clip}(r_n(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_n) \right], \tag{24}$$

where ϵ is a hyperparameter and $\hat{A}_n$ is a truncated version of generalized advantage estimation which can be defined as follows:

$$\hat{A}_n = \delta_n + (\gamma\lambda)\delta_{n+1} + \cdots + (\gamma\lambda)^{N-n+1}\delta_{N-1}, \tag{25}$$

where $\delta_n = r_n + \gamma V(s_{n+1}) - V(s_n)$. The function in (24) takes a lower value when comparing the objectives used in the TRPO with the objectives to which clipping is applied. With this clipping method, we only consider the change in the probability ratio if it improves the objective. If it makes the objective worse, we leave it out.

Subsequently, PPO incorporates an actor-critic network architecture, wherein the policy and value functions share the parameters inside the network design. In the context of utilizing a neural network architecture to share parameters between the policy function and the value function, it is

(a) Learning convergence results based on different algorithms.

(b) Learning results based on the number of STAR-RIS elements.

(c) Average throughput of MUs with maximum transmission power changes.
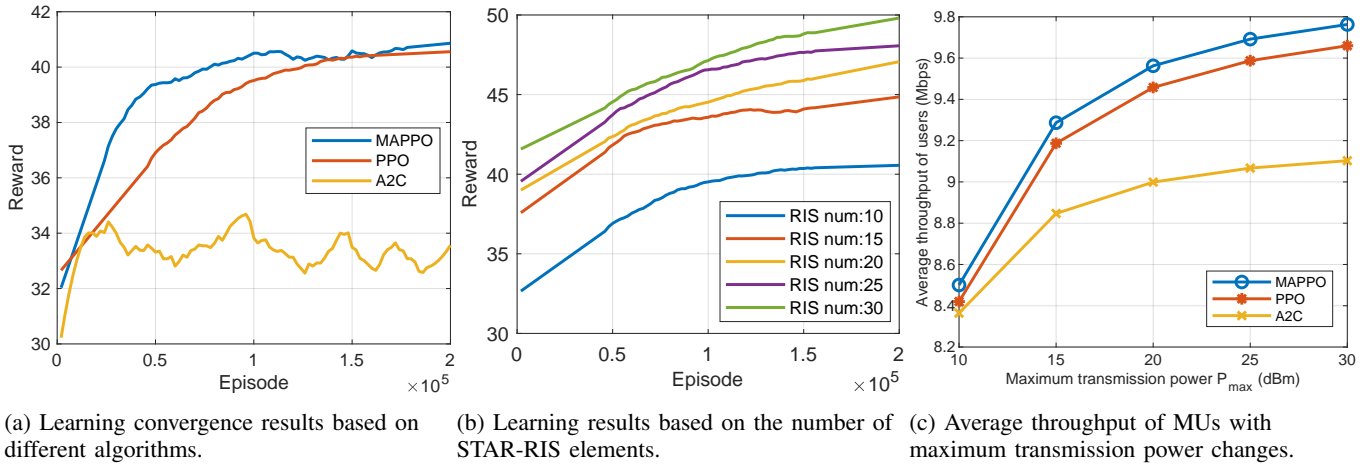
Fig. 2: Simulation results of proposed scheme convergence and comparison with baselines.
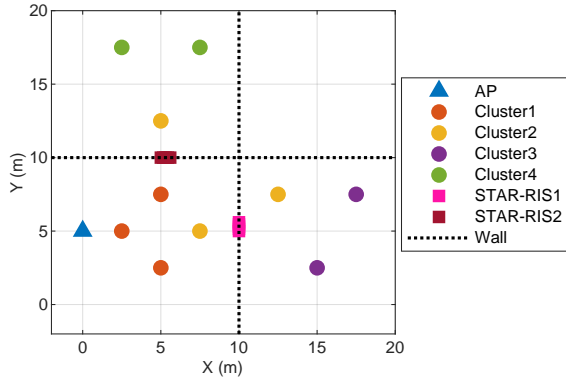

Fig. 3: Deployments of AP, STAR-RISs, MUs.

imperative to deploy a loss function that effectively integrates the policy surrogate and an error term derived from the value function. The objective is achieved by integrating the policy surrogate with a value function error component as follows:

$$L_n^{\text{PPO}}(\theta) = \hat{\mathbb{E}}_n \left[L_n^{\text{CLIP}}(\theta) - c_1 L_n^{\text{VF}}(\theta) + c_2 E[\pi_\theta](s_n) \right], \quad (26)$$

where c_1 and c_2 are coefficients, E denotes an entropy function, and L_n^{VF} is a squared-error loss. In this objective L^{PPO} can further be augmented by adding an entropy bonus E to ensure sufficient exploration.

The proposed MAPPO in this paper is each optimization variable $(\boldsymbol{\omega}, \boldsymbol{\Phi})$ PPO agents. Each agent learns simultaneously in the same environment. Moreover, each agent can determine the optimal action for a common reward. Therefore, we introduce the Markov Decision Process (MDP) of the agents used for learning.

$$\mathcal{S}(n) = \{\boldsymbol{\gamma}^*, \boldsymbol{\omega}(n-1), \boldsymbol{\Phi}(n-1), \{\hat{h}_u\}_{u \in \mathcal{U}}\}, \quad (27)$$

$$\mathcal{A}^\omega(n) = \{\omega_k(n)\}_{k \in \mathcal{K}}, \quad (28)$$

$$\mathcal{A}^\Phi(n) = \{\Phi_l^F, \Phi_l^B\}_{l \in \mathcal{L}}, \quad (29)$$

$$\mathcal{R}(n) = \min_{u \in \mathcal{U}}(R_{k,u}), \quad (30)$$

where $\boldsymbol{\omega}(n-1)$ and $\boldsymbol{\Phi}(n-1)$ denote the active and passive beamforming vector at the last step, $\mathcal{A}^\omega(n)$ and $\mathcal{A}^\Phi(n)$ denote the active and passive beamforming agent. The reward is determined by the lowest throughput of all users. This is to satisfy the minimum throughput of all users for the constraint (15b). Based on the MDP for the proposed optimization problem, the optimal value of the decision variable can be obtained by executing the proposed MAPPO algorithm, such as Algorithm 2.

IV. SIMULATION RESULTS

We consider one AP in a region of 20 m^2, which provides services to the 10 MUs in our simulation setup. Then, the considered region is divided into four rooms, each of which is surrounded by walls. Furthermore, two STAR-RISs are installed in the room where the AP is located. In the learning stage, MUs are randomly placed, and in the verification stage, MUs are placed as in Fig. 3. The AP has four antennas with a frequency of 6 GHz and a noise density of -100 dBm/Hz. Each STAR-RIS is composed of 10 elements, and separations between elements are 0.2 in width and 0.1 in length. One episode has 10 time slots for learning settings, and the learning network model has 2 hidden layers and 256 units for each hidden layer. In our simulation, correlation-based K-means clustering for MU pairing is performed after MUs are assigned at the beginning of each episode. Finally, to evaluate the performance of the proposed algorithm, we use the following three different algorithms as follows:

- *MAPPO (proposed)*: The proposed MAPPO algorithm, in which each beamforming vector variable is thought of as an agent that learns its variable optimization.
- *PPO*: The general PPO algorithm learns all beamforming vector variables in one network at a time.
- *A2C*: The multi-agent advanced actor-critic (MAA2C) algorithm, in which each beamforming vector variable is thought of as an agent that learns its variable optimization.
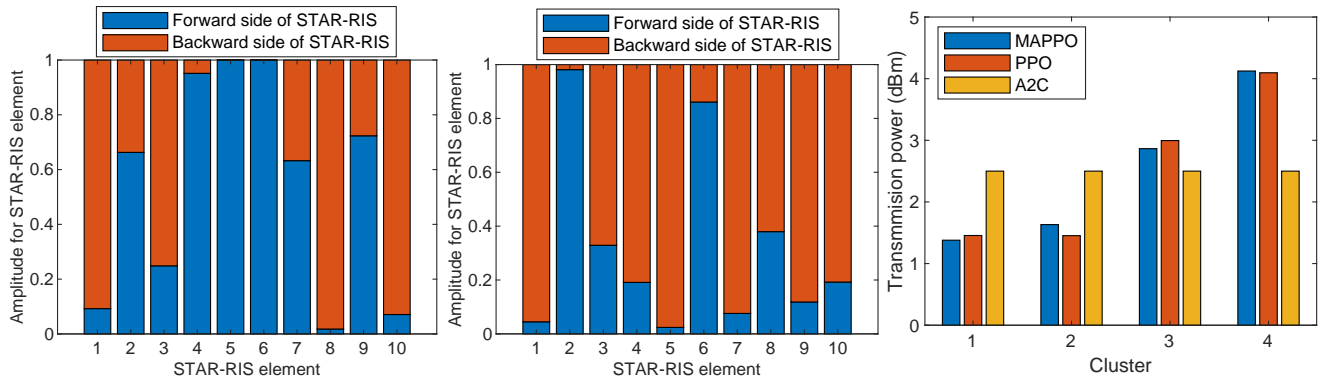
(a) Amplitude by Element for STAR-RIS 1 in Fig. 3.

(b) Amplitude by Element for STAR-RIS 2 in Fig. 3.

(c) Status of active beamforming allocation for clusters by algorithm.

Fig. 4: Analysis of optimal values derived from the learning model.

Fig. 2a shows the results of learning convergence based on different algorithms. Our proposed MAPPO completed the learning with the fastest and highest rewards, followed by a slow PPO but similar rewards. A2C showed learning results that failed to converge and continued to vibrate. Fig. 2b shows the results of learning convergence according to the number of elements in STAR-RIS. It can be seen that the final reward increases as the number of STAR-RIS elements increases, but at the same time, it can be seen that the improvement of the reward becomes smaller and smaller. This shows that the efficiency of learning can vary depending on the requirements of STAR-RIS because learning was done with an artificial neural network of the same size. Fig. 2c compares the average throughput of users by algorithm according to the change in maximum transmission power. As checked in Fig. 2a, it was confirmed that the proposed MAPPO showed the best performance according to the results of learning convergence, followed by PPO and A2C.

Also, we record and present optimal values in a fixed environment for an in-depth understanding of optimization. Fig. 4a and Fig. 4b show the amplitude of STAR-RIS by element in the environment of Fig. 3. It is noteworthy that to optimize the communication throughput of MUs, STAR-RIS assigns large amplitudes where there are no APs. Adding the amplitudes of all elements, STAR-RIS 1 was assigned more amplitudes in the opposite direction of the AP with 5.3997595 : 4.6002405, and likewise, STAR-RIS 2 3.19707643 : 6.802357 was assigned more amplitudes in the opposite direction of the AP. It can be confirmed that this is an optimal choice of STAR-RIS for improving overall communication performance. Finally, Fig. 4c shows how much active beamforming was allocated to the cluster by algorithms in the environment of Fig. 3. The proposed MAPPO allocated relatively little power to Cluster 1 and Cluster 2 because users close to AP can be guaranteed sufficient communication throughput over distance. On the contrary, Cluster 3 and Cluster 4 are users who do not have a direct connection to the AP, so they are optimized to ensure minimum throughput by allocating high power. The

PPO algorithm showed similar allocations to the MAPPO, but the A2C algorithm allocates similar power for all clusters, which can be inferred from the results of Fig. 2c above that the performance was low.

V. CONCLUSIONS

In this paper, we have studied the joint design and optimization of Multi-STAR-RISs-aided NOMA in an indoor environment using MARL. Then, we formulated an optimization problem to maximize the total throughput of MUs by optimizing user pairing, active beamforming vector, and passive beamforming vector while satisfying the resource constraints. We have divided the original problem into two subproblems to address this problem. Firstly, we have employed correlation-based K-means clustering to solve the MU pairing problem. Then, to jointly deal with beamforming vector optimizations, we have proposed the MAPPO, which can make quick decisions in the given environment owing to its low complexity. Based on the proposed MAPPO, by configuring agents for each beamforming vector, it was possible to have faster and higher performance than conventional single-agent-based PPO. In the simulation results, we have shown not only the learning convergence results for neural networks but also the learning convergence results based on the number of different STAR-RIS elements and the performance of each algorithm. In addition, while confirming how the beamforming vectors were actually optimized in one test environment, we analyzed in depth why the proposed MAPPO was high. In the future, this paper can be expanded by applying optimization through multi-cell or other metaheuristic or numerical methods that were not covered in this paper.

REFERENCES

[1] Z. Chen, G. Chen, J. Tang, S. Zhang, D. K. So, O. A. Dobre, K.-K. Wong, and J. Chambers, "Reconfigurable-intelligent-surface-assisted b5g/6g wireless communications: Challenges, solution, and future opportunities," *IEEE Communications Magazine*, vol. 61, no. 1, pp. 16–22, Sep. 2022.

[2] Y. K. Tun, N. H. Tran, D. T. Ngo, S. R. Pandey, Z. Han, and C. S. Hong, "Wireless network slicing: Generalized kelly mechanism-based resource allocation," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1794–1807, Jul. 2019.

[3] W. Khalid, Z. Kaleem, R. Ullah, T. V. Chien, S. Noh, and H. Yu, "Simultaneous transmitting and reflecting-reconfigurable intelligent surface in 6G: Design guidelines and future perspectives," *IEEE Network*, pp. 1–9, Dec. 2022.

[4] Y. M. Park, Y. K. Tun, Z. Han, and C. S. Hong, "Trajectory optimization and phase-shift design in irs-assisted uav network for smart railway," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 10, pp. 11 317–11 321, Jul. 2022.

[5] Y. M. Park, S. S. Hassan, Y. K. Tun, Z. Han, and C. S. Hong, "Joint resources and phase-shift optimization of mec-enabled uav in irs-assisted 6g thz networks," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, Budapest, Hungary, Apr. 2022, pp. 1–7.

[6] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, "Star: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces," *IEEE Wireless Communications*, vol. 28, no. 6, pp. 102–109, Dec. 2021.

[7] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Communications Letters*, vol. 25, no. 9, pp. 3134–3138, Sep. 2021.

[8] Y. Wang, S. Hou, Q. Zhang, and H. Li, "Joint optimization of multi-star-ris aided noma systems," in *2023 8th International Conference on Computer and Communication Systems (ICCCS)*, Punjab, India, Apr. 2023, pp. 221–225.

[9] S. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink noma systems: Key techniques and open issues," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[10] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-noma systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7425–7440, Sep. 2018.

[11] J. Zuo, Y. Liu, Z. Ding, L. Song, and H. V. Poor, "Joint design for simultaneously transmitting and reflecting (star) ris assisted noma systems," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 611–626, Aug. 2022.

[12] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam sic for cellular mimo downlink," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, Las Vegas, USA, Sep. 2013, pp. 1–5.

[13] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, Jan. 1979.

[14] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Millimeter-wave noma with user grouping, power allocation and hybrid beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5065–5079, Aug. 2019.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Jul. 2017.

[16] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *the proc. of International conference on machine learning (ICML)*, Lille, France, Jul. 2015.