

# DRL Enabled Coverage and Capacity Optimization in STAR-RIS-assisted Networks

Xinyu Gao, *Member, IEEE*, Wenqiang Yi, *Member, IEEE*,  
Yuanwei Liu, *Senior Member, IEEE*, Jianhua Zhang, *Senior Member, IEEE*,  
and Ping Zhang, *Fellow, IEEE*

**Abstract**—Simultaneously transmitting and reflecting reconfigurable intelligent surfaces (STAR-RISs) is a promising passive device that contributes to full-space coverage via transmitting and reflecting the incident signal simultaneously. As a new paradigm in wireless communications, how to analyze the coverage and capacity performance of STAR-RISs becomes essential but challenging. To solve the coverage and capacity optimization (CCO) problem in STAR-RIS-assisted networks, a multi-objective proximal policy optimization (MO-PPO) algorithm is proposed to handle long-term effects. To strike a balance between each objective, the MO-PPO algorithm provides a set of optimal solutions to approach a Pareto front (PF), where the solution on the approximate PF is regarded as an optimal result. Moreover, in order to improve the performance of the MO-PPO algorithm, two update strategies, i.e., action-value-based update strategy (AVUS) and loss function-based update strategy (LFUS), are investigated. For the AVUS, the improved point is to integrate the action values of both coverage and capacity and then update the loss function. For the LFUS, the improved point is only to assign dynamic weights for both loss functions of coverage and capacity, while the weights are calculated by a min-norm solver at every update. The numerical results demonstrated that the investigated update strategies outperform the fixed weights MO optimization algorithms in different cases, which include a different number of sample grids, the number of STAR-RISs, the number of elements in the STAR-RISs, and the size of STAR-RISs. Additionally, the STAR-RIS-assisted networks achieve better performance than conventional wireless networks without STAR-RISs. Moreover, with the same bandwidth, a millimetre wave is able to provide higher capacity than sub-6 GHz, but at a cost of smaller coverage.

**Index Terms**—Coverage and capacity optimization (CCO), multi-objective proximal policy optimization (MO-PPO), simultaneously transmitting and reflecting reconfigurable intelligent surfaces (STAR-RISs)

## I. INTRODUCTION

For supporting increasing heterogeneous quality-of-service requirements of future wireless networks, e.g., high data rate, low latency, high reliability, massive connectivity, etc., an emerging communication paradigm, i.e., reconfigurable intelligent surfaces (RISs) [2]–[5] has been proposed to smartly control the wireless communication environment. RISs are able to offer line-of-sight (LOS) links to users located in blocked areas via reflection to improve both the coverage and capacity of conventional wireless networks. However,

conventional RISs have maximal 180° coverage, where the ‘blind zone’ still exists at the backside of RISs. To overcome this limitation, a new concept named simultaneously transmitting and reflecting RISs (STAR-RISs) [6] becomes appealing. In contrast to conventional RISs, STAR-RISs are able to transmit and reflect the incident signal simultaneously, which contributes to full-space coverage [7]. As a new communication paradigm, it is an ultra-interesting question how STAR-RISs perform in terms of coverage and capacity. Note that coverage and capacity optimization (CCO) is one of the typical operational tasks mentioned by the 3rd Generation Partnership Project [8]. Since the coverage and capacity have several conflicting relationships, simultaneously optimizing them is important. For example, high transmit power contributes to large coverage but high inter-cell interference that reduces the capacity performance. To this end, multi-objective machine learning (MOML) [9] algorithms can be a potential solution. Compared to single-objective algorithms, MOML algorithms are capable of handling the inherent conflict between objectives to achieve a group of optimal solutions by coordinating and compromising the requirements of objectives.

### A. Related Works

1) *Capacity or Coverage Optimization for STAR-RISs Networks*: Conventional performance optimization for STAR-RIS-assisted networks focuses on a single objective: capacity or coverage. For capacity performance, there are some primary works. In [10], a partitioning algorithm was proposed to determine the proper number of transmitting/reflecting elements that need to be assigned to each user and maximize the system sum-rate while guaranteeing the quality-of-service requirements for individual users. In STAR-RIS-assisted non-orthogonal multiple access (NOMA) systems, the authors in [11] proposed a sub-optimal two-layer iterative algorithm to maximize the achievable sum-rate by jointly optimizing the decoding order, power allocation coefficients, active beamforming, and transmission and reflection beamforming. The sum-rate performance of STAR-RIS-assisted full-duplex communication systems was investigated in [12], where the successive convex approximation technique has been employed to develop efficient algorithms for obtaining sub-optimal solutions. In [13], the authors proposed a sub-optimal block coordinate descent algorithm to maximize the weighted sum-rate for a STAR-RIS-assisted multiple-input multiple-output network. The authors in [14] investigated the resource allocation problem in a STAR-RIS-assisted multi-carrier communication network and proposed location-based matching and semidefinite

Part of this work has been accepted to appear in the IEEE Wireless Communications and Networking Conference, Mar. 26–Mar. 29 March 2023 [1].

X. Gao, W. Yi, and Y. Liu are with the Queen Mary University of London, London E1 4NS, U.K. (e-mail: {x.gao, w.yi, yuanwei.liu}@qmul.ac.uk).

J. Zhang and P. Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (email: {jhzhang, pzhang}@bupt.edu.cn).

programming algorithms to maximize the system sum-rate. To derive the approximated average achievable rates of two users, the authors in [15] investigated the performance of STAR-RIS-assisted downlink NOMA networks by a large array of analysis methods. For coverage performance, only one recent work has discussed its optimization problem. The STAR-RIS-assisted two-user communication networks were studied in [16], where the search-based algorithms were proposed to obtain the optimal one-dimensional (1D) coverage range.

2) *CCO based on MOML algorithms*: There are three main CCO solutions based on MOML algorithms: 1) Keep one objective in the objective function and move the rest objectives to constraints, while the obtained results are sub-optimal [18]. 2) Assign a fixed weight to each objective. This method achieves the optimal results in a single scenario, while it cannot be used in other weight combinations, i.e., other network operation designs [19]. 3) Obtain a set of optimal solutions according to Pareto-based multi-objective optimization algorithms, where one of these solutions can be selected to meet any specific network operation designs [20]. More specifically, for the first method, an reinforcement learning (RL) algorithm-based solution for CCO by optimizing the base station (BS) antenna electrical tilt was proposed in [18], where the coverage objective was considered in the constraint. The proposed sub-optimal solution has the potential to reduce operational costs and complexity, as well as improve the quality of experience for mobile users. For the second method, in [19], minimization of drive tests (MDT)-driven deep RL algorithm was investigated to maximize the coverage and capacity by tuning antennas tilts on a cluster of cells from the cellular network, where the fixed weights were assigned for coverage and capacity. The results showed that the proposed MDT-driven approaches outperform baseline approaches, i.e., deep Q-network and best-first search, in terms of long-term reward and sample efficiency. For the third method, the authors in [20] developed two RL algorithm-based approaches for maximizing coverage and minimizing interference by jointly optimizing the transmit power and antenna down-tilt across cells. The results suggested that data-driven techniques can effectively self-optimize coverage and capacity in cellular networks. There are some other promising MOML methods [21], [22]. A new algorithm was introduced in [21] for multi-objective reinforcement learning (MORL) with linear preferences, with the goal of enabling few-shot adaptation to new tasks. The authors in [22] proposed an upper bound for the multi-objective loss and show that it can be optimized efficiently. However, compared to a simple extension of the vanilla RL approaches to MOML algorithms, a new RL approach named proximal policy optimization (PPO) algorithm is able to provide a more stable training process (e.g., implement small batch updates in multiple training steps) and can be a booster for MOML algorithms.

#### B. Motivations and Contributions

As can be seen from related works, the CCO problem of STAR-RIS-assisted wireless networks is still in its early stage. In this research direction, there are two main challenges:

- **Characterizing Coverage in STAR-RIS-assisted Networks**: STAR-RISs provide a new degree of freedom

for manipulating signal propagation, thus increasing the flexibility of network design. Characterizing the two-dimensional (2D) coverage range for the STAR-RIS-assisted networks is challenging, compared to the one-dimensional coverage range described by the conventional networks. Additionally, the coverage characterization may be affected by the capacity, since the two objectives are conflicts.

- **Designing MORL Algorithms to Solve CCO Problem**: Conventional Pareto-based MO optimization (MOO) solutions mainly aim to find an approximate Pareto front (PF) of objectives within a time step, which ignores the dynamic requirements of temporal correlations in long-term wireless communications. PPO is a policy gradient method where policy updates use a surrogate loss function to avoid catastrophic drops in performance. In addition, the new MOO methods [21], [22] have the capability to dynamically update the weights of objectives. Therefore, how to obtain the Pareto optimal (PO) solution based on the PPO algorithm and these two new MOO methods is challenging.

To solve these challenges and fully reap the advantages of STAR-RISs, in this paper, we propose a new RL approach based on the PPO algorithm, named multi-objective PPO (MO-PPO) algorithm, to provide the maximum coverage and capacity for STAR-RIS-assisted networks. The optimal results obtained by the MO-PPO algorithm are different according to the different update strategies. The main contributions of this paper can be summarized as follows:

- We propose a new model for a narrow-band downlink mode-splitting protocol-based STAR-RIS-assisted network consisting of two single-antenna BSs, where the serving range is defined as a square region. To quantitatively analyze the coverage and capacity, the serving range is discretized into numerous square grids, and the centre point of each grid sets as the evaluating sample point. Based on this framework, we formulate the CCO problem of STAR-RIS-assisted networks by jointly optimizing the transmit power, the reflection phase shift matrix, and the transmission phase shift matrix.
- We investigate an action value-based update strategy (AVUS) for the MO-PPO algorithm to solve the CCO problem. The core point of this strategy is to learn multiple policies for integrating the action values of both coverage and capacity by random sampling preferences, and further invoke a coefficient to update the policy by homotopy optimization. This update strategy with high performance is able to provide the optimal coverage and capacity, while it has to spend a long time to achieve convergence. Therefore, the AVUS has strict requirements on the computation resource, which is suitable for networks with strong computation capability.
- We adopt a loss function-based update strategy (LFUS) for the MO-PPO algorithm to reduce the complexity brought by the AVUS. The improved point is to assign dynamic weights for both loss functions of coverage and capacity and to update the whole MO-PPO policy with

an integrated loss function of coverage and capacity. The dynamic weights are re-calculated by a min-norm solver at every update. Compared to the AVUS, this strategy has slightly worse performance, but it still has acceptable performance gain when compared to the conventional CCO solutions.

- We illustrate that both AVUS and LFUS-based MO-PPO algorithms are capable of striking a balance between the conflicting goals in terms of coverage and capacity. Then, AVUS and LFUS-based algorithms are able to provide the Pareto optimality compared to conventional fixed weights MOO algorithms. With the same bandwidth, a millimetre wave (mmWave) is able to provide better capacity while sub-6 GHz provides better coverage. Next, the coverage and capacity have a positive correlation with the number of STAR-RISs. Finally, when the number of elements in STAR-RISs is fixed, the coverage and capacity have a negative correlation with the physical size of STAR-RISs.

### C. Organization

The rest of this paper is organized as follows. Section II presents the system model for the considered STAR-RIS-assisted networks, and the coverage and capacity optimization problems are formulated. Section III provides the preliminaries, including the principles of the PPO algorithm and the PO solution. In Section IV, we investigate the two updated strategies-based MO-PPO algorithms, i.e., AVUS and LFUS, which are updated for different parts of the algorithm. Section V presents numerical results to verify the effectiveness of the proposed MO-PPO algorithms, by considering the different number of sample grids, the different number of elements in STAR-RISs, the different number of STAR-RISs, and the different physical sizes of STAR-RISs modules. Finally, Section VI concludes this paper.

*Notations:* Scalars, vectors, and matrices are denoted by lower-case, bold-face lower-case, and bold-face upper-case letters, respectively. The conjugate transpose of vector  $\mathbf{a}$  is denoted by  $\mathbf{a}^H$ . The  $\text{diag}(\mathbf{a})$  denotes a diagonal matrix with the elements of vector  $\mathbf{a}$  on the main diagonal. The  $\|\mathbf{a}\|$  denotes the norm of vector  $\mathbf{a}$ . The  $\text{Mod}(a, b)$  denotes the modulus operation between values  $a$  and  $b$ . The  $\lfloor a \rfloor$  denotes the truncated argument of value  $a$ . The  $*$  denotes the dot multiplication operation. The  $\mathbb{E}[\mathbf{A}]$  is the expectation operator of matrix  $\mathbf{A}$ . The  $\log_2(\mathbf{A})$  represents a logarithmic function with a constant base of 2 for matrix  $\mathbf{A}$ . The  $\text{tr}(\mathbf{A})$  denotes the trace of matrix  $\mathbf{A}$ .

## II. SYSTEM MODEL AND PROBLEM FORMULATION

As shown in Fig. 1(a), we consider a narrow-band downlink STAR-RIS-assisted network consisting of two single-antenna BSs and  $N_s$  STAR-RISs of the same size equipped with  $K = K_H K_V$  reconfigurable elements, where  $K_H$  and  $K_V$  denote the number of elements per row and column, respectively. The serving range is defined as a square region with the length of the side  $R_s$ , while the region is discretized into numerous square grids with the length of the side  $R_g$ , while the centre point of each grid acts as the sample point [23]. The BSs are located at the bottom left and bottom right corners with the same height  $h_b$ , while STAR-RISs with the height  $h_{n_s}$  and width  $\omega_{n_s}$  are deployed at designated locations in the square region.

### A. Grid-based Geographic Model

Assuming a three-dimensional (3D) Cartesian coordinate system, where the origin is set at the top-left corner. Here, the locations of two BSs and  $n_s$ -th STAR-RISs are denoted by  $\mathbf{B}_1 = (R_s, 0, h_b)$ ,  $\mathbf{B}_2 = (R_s, R_s, h_b)$ , and  $\mathbf{A}_{n_s} = (x_{n_s}, y_{n_s}, h_{n_s})$ , respectively. Note that  $h_{n_s}$  is the height of the STAR-RISs module, and the thickness of STAR-RISs is ignored. The height  $h_{n_s}$  and width  $\omega_{n_s}$  are depicted in Fig. 1(b). The indicators  $I_{h_{n_s}}$  and  $I_{\omega_{n_s}}$  are invoked to characterize  $h_{n_s}$  and  $\omega_{n_s}$  to further depict whether the direct links between the BSs and sample points exist or not, which can be expressed as follows:

$$I_{h_{n_s}} = \begin{cases} 1, & \text{If } h_{n_s} \leq \frac{R_g h_b}{2R_s - R_g} \\ 0, & \text{If } h_{n_s} > \frac{R_g h_b}{2R_s - R_g} \end{cases}, \quad (1)$$

$$I_{\omega_{n_s}} = \begin{cases} 1, & \text{If } \omega_{n_s} \leq \frac{R_g R_s}{2R_s - R_g} \\ 0, & \text{If } \omega_{n_s} > \frac{R_g R_s}{2R_s - R_g} \end{cases}. \quad (2)$$

If we ensure that there is at least one direct link between BSs and any given sampling point, the indicators  $I_{h_{n_s}}$  and  $I_{\omega_{n_s}}$  need to satisfy the condition:  $I_{h_{n_s}} = 1$  and/or  $I_{\omega_{n_s}} = 1$ . Thus, the indicators  $I_{h_{n_s}}$  and  $I_{\omega_{n_s}}$  can be unified as follows:

$$I_{n_s} = I_{h_{n_s}} \vee I_{\omega_{n_s}}, \quad (3)$$

where  $\vee$  denotes the OR operator.  $I_{n_s} = 1$  denotes that the BSs are able to establish a direct link with the considered receiver from above and/or from the side of the STAR-RISs; Otherwise, there is no direct link between the BSs and the sample points. Additionally, if considering other numbers of BSs, this geographic model needs to be reconstructed according to the actual deployment of BSs.

The coverage and capacity can be accordingly characterized based on the discretized grids. The total number of grids is  $N = \lceil R_s / R_g \rceil^2$ , where the set of sample points can be denoted as  $\mathbf{s} = \{s_1, s_2, \dots, s_N\}$ . In practical networks, in order to characterize the importance of each grid at each time step  $t$ , two time-related weights,  $w_{\text{cov},i}(t)$  and  $w_{\text{cap},i}(t)$ , are assigned for coverage and capacity of each sample points  $s_i$  ( $i \in \{1, 2, \dots, N\}$ ), respectively. Moreover, the weights have been unified, i.e.,  $\sum_{i=1}^N w_{\text{cov},s_i}(t) = 1$  and  $\sum_{i=1}^N w_{\text{cap},s_i}(t) = 1$ . In this system model, we study long-term communication with a time period  $\mathcal{T}$ . For each sample point at any time step, the weighted assignments  $w_{\text{cov},s_i}(t)$  and  $w_{\text{cap},s_i}(t)$  are influenced by the previous network performance and resource allocation strategy. Therefore, the considered problem can be regarded as a Markov Decision Process (MDP).

### B. Spatially Correlated Channel Model

In this section, the fading channels from BSs to STAR-RISs, from STAR-RISs to sample points, and from BSs to sample points are introduced, as well as their spatial channel correlations. There are three different splitting protocols: 1) power-splitting (PS) protocol. In this case, all elements of the STAR-RIS are assumed to operate in the transmission and reflection modes, where the energy of the signal incident on each element is generally split into the energies of the transmitted and reflected signals with a ratio. 2) mode-splitting (MS) protocol. In this case, all elements of the STAR-RIS are divided into two groups. Specifically, one group contains some elements that operate



$$y_{a,n_s,s_i} = \begin{cases} \left( \mathbf{h}_{\delta,n_s,s_i}^H \Phi_{\delta,n_s} \mathbf{h}_{a,n_s} + h_{a,s_i} \right) x + n, & \text{If } I_{n_s} = 1, I_{\omega_{n_s}} = 0, \\ \left( \mathbf{h}_{\delta,n_s,s_i}^H \Phi_{\delta,n_s} \mathbf{h}_{a,n_s} + \bar{a}_a h_{a,s_i} \right) x + n, & \text{If } I_{n_s} = 1, I_{h_{n_s}} = 0, I_{\omega_{n_s}} = 1, \\ \left( \mathbf{h}_{\delta,n_s,s_i}^H \Phi_{\delta,n_s} \mathbf{h}_{a,n_s} \right) x + n, & \text{If } I_{n_s} = 0, \end{cases} \quad (9)$$

shift of STAR-RISs is based on the instantaneous CSI. Denote  $\mathbf{h}_{a,n_s}$ ,  $\mathbf{h}_{\delta,n_s,s_i}$ , and  $h_{a,s_i}$  as the channel from  $a$ -th BS to  $n_s$ -th STAR-RISs with mode  $\delta$ , from  $n_s$ -th STAR-RISs to  $s_i$ -th sample point with mode  $\delta$ , and from  $a$ -th BS to  $s_i$ -th sample point, respectively. Here, the channels  $h_{a,s_i}$  and  $\mathbf{h}_{\delta,u}$ ,  $u \in \{a, n_s; n_s, s_i\}$  can be modelled as Rician fading model, which is expressed as:

$$h_{a,s_i} = \sqrt{L_{a,s_i}} \left( \sqrt{\frac{\alpha_{a,s_i}}{1 + \alpha_{a,s_i}}} h_{a,s_i}^{\text{LOS}} + \sqrt{\frac{1}{1 + \alpha_{a,s_i}}} h_{a,s_i}^{\text{NLOS}} \right), \quad (7)$$

$$\mathbf{h}_{\delta,u} = \sqrt{L_u} \left( \sqrt{\frac{\alpha_u}{1 + \alpha_u}} \mathbf{h}_{\delta,u}^{\text{LOS}} + \sqrt{\frac{1}{1 + \alpha_u}} \mathbf{h}_{\delta,u}^{\text{NLOS}} \right), \quad (8)$$

where  $L_{\bar{u}}, \bar{u} \in \{a, s_i; u\}$ , and  $\alpha_{\bar{u}}, \bar{u} \in \{a, s_i; u\}$  denote the corresponding path loss and Rician factor, respectively. The  $h_{a,s_i}^{\text{LOS}}$  denotes the deterministic LOS component of the channel from  $a$ -th BS to  $s_i$ -th sample point, which can be calculated according to the locations of BS and STAR-RISs.  $\mathbf{h}_{\delta,a,n_s}^{\text{LOS}} = \mathbf{b}(\psi^{\delta,a,n_s}, \theta^{\delta,a,n_s}) = \mathbf{b}\{\arcsin[(h_b - h_{n_s})/d_{a,n_s}], \arccos[(R_s - x_{n_s})/\bar{d}_{\delta,a,n_s}]\}$  and  $\mathbf{h}_{\delta,n_s,s_i}^{\text{LOS}} = \mathbf{b}(\psi^{\delta,n_s,s_i}, \theta^{\delta,n_s,s_i}) = \mathbf{b}\{\arcsin(h_{n_s}/d_{\delta,n_s,s_i}), \arccos[(x_{n_s} - x_{s_i})/\bar{d}_{\delta,n_s,s_i}]\}$  are the deterministic LOS components for the channels from  $a$ -th BS to  $n_s$ -th STAR-RISs, and from  $n_s$ -th STAR-RISs to  $s_i$ -th sample point, respectively. Among them,  $d_{\delta,a,n_s}$  and  $d_{\delta,n_s,s_i}$  denote 3D distance between  $a$ -th BS and  $n_s$ -th STAR-RISs, and 3D distance between  $n_s$ -th STAR-RISs and  $s_i$ -th sample point, while  $\bar{d}_{\delta,a,n_s}$  and  $\bar{d}_{\delta,n_s,s_i}$  denote 2D distance between  $a$ -th BS and  $n_s$ -th STAR-RISs, and 2D distance between  $n_s$ -th STAR-RISs and  $s_i$ -th sample point. The  $x_{n_s}$ ,  $x_{s_i}$  indicate the  $n_s$ -th STAR-RISs, and  $s_i$ -th sample point, respectively.  $\mathbf{h}_{\delta,a,n_s}^{\text{NLOS}} \sim \mathcal{CN}(0, \mathbb{E}[\mathbf{h}_{\delta,a,n_s}^{\text{NLOS}}(\mathbf{h}_{\delta,a,n_s}^{\text{NLOS}})^H])$ ,  $\mathbf{h}_{\delta,n_s,s_i}^{\text{NLOS}} \sim \mathcal{CN}(0, \mathbb{E}[\mathbf{h}_{\delta,n_s,s_i}^{\text{NLOS}}(\mathbf{h}_{\delta,n_s,s_i}^{\text{NLOS}})^H])$ , and  $h_{a,s_i}^{\text{NLOS}} \sim \mathcal{CN}(0, 1)$  are the non-line-of-sight (NLOS) components modeled as Rayleigh fading. Furthermore, for path loss  $L_u$ , it can be modeled as  $L_{\bar{u}} = C_0 d_v^{-\gamma_v}$ ,  $v \in \{\{\delta, a, n_s\}, \{\delta, n_s, s_i\}, \{a, s_i\}\}$ , where  $C_0 = c/(4\pi d_0 f_c)$  denotes the path loss at the reference distance  $d_0 = 1\text{m}$  under frequency  $f_c$ ,  $c$  is the velocity of light, and  $\gamma_v$  represents the path loss factor.

### C. Signal Model

Since the size of the STAR-RISs module affects the direct link, the received signal  $y_{a,n_s,s_i} \in \mathbb{C}$  from the  $a$ -th BS to the  $s_i$ -th sample point via  $n_s$ -th STAR-RISs is determined by  $I_{n_s}$ . Thus, the received signal  $y_{a,n_s,s_i}$  can be written as (9) [27], where the total transmit power  $P_t = |x|^2$  and  $n \sim \mathcal{CN}(0, \sigma^2)$  is the additive white Gaussian noise variance.  $\bar{a}_a$  is a indicator that characterizing the direct link between  $a$ -th BS and  $s_i$ -th sample point.  $\bar{a}_a = 1$  denotes that there is a direct link between  $a$ -th BS and  $s_i$ -th sample point, while  $\bar{a}_a = 0$  denotes the

direct link between  $a$ -th BS and  $s_i$ -th sample point is blocked. Due to the additional backhaul resources, the coordination of BSs needs extra communication requirements and computation resources, hence, we consider the most common practical strategy. based on the received signal power, the reference signal receiving power (RSRP) can be defined as the maximal useful signal power from all possible sources. The RSRP at the sample point  $s_i$  is given by [23]:

$$\text{RSRP}_{s_i} = \max_{a \in \{1,2\}, n_s \in \{1,2,\dots,N_s\}} |y_{a,n_s,s_i} - n|^2. \quad (10)$$

The achievable signal-to-interference-plus-noise ratio (SINR) of  $s_i$ -th sample point is calculated as follows [23]:

$$\text{SINR}_{a,n_s,s_i} = \frac{|y_{a,n_s,s_i} - n|^2}{\sum_{a'=1, a' \neq a}^A \sum_{n'_s=1, n'_s \neq n_s}^{N_s} |y_{a',n'_s,s_i} - n|^2 + n^2}, \quad (11)$$

where  $a = 1, a' = 2$ ; and  $a = 2, a' = 1$ , otherwise. Assume the RSRP threshold for all sample points is  $R_{th}$ , the weighted coverage ratio at time step  $t$  can be written as

$$\text{Coverage}(t) = \frac{\|\mathbf{w}_{\text{cov}, \check{\mathbf{s}}}(t) \cdot \check{\mathbf{s}}(t)\|}{N}, \quad (12)$$

where  $\check{\mathbf{s}}(t) = \{\check{s}_1(t), \check{s}_2(t), \dots, \check{s}_{\check{N}}(t)\}$  is the set of the sample points at time  $t$  that satisfying the condition  $\text{RSRP}_{\check{s}_{\check{n}}(t)} \geq R_{th}$ ,  $\check{s}_{\check{n}}(t) \in \check{\mathbf{s}}(t)$ .  $\mathbf{w}_{\text{cov}, \check{\mathbf{s}}}(t) = \{w_{\text{cov}, \check{s}_1}(t), w_{\text{cov}, \check{s}_2}(t), \dots, w_{\text{cov}, \check{s}_{\check{N}}}(t)\}$  is the normalized corresponding coverage weights for the sample points  $\check{\mathbf{s}}(t)$ . For the network capacity, it is mainly determined by SINR, so at the time step  $t$ , the weighted capacity can be represented by

$$\text{Capacity}(t) = \sum_{s_i=1}^{N_s} w_{\text{cap}, s_i}(t) \cdot B \log_2 (1 + \text{SINR}_{a^*, n_s^*, s_i}(t)), \quad (13)$$

where  $B$  is the system bandwidth and  $a^*, n_s^* = \arg \max_{a \in \{1,2\}, n_s \in \{1,2,\dots,N_s\}} |y_{a,n_s,s_i} - n|^2$ . According to equation (10), when the transmit power  $P_t(t)$  increases, the coverage will also increase with the increases of the RSRP. However, the interference in equation (11) will rise with the growth of the transmit power  $P_t(t)$ , while the performance of capacity will be degraded as the increase of interference. Therefore, there exists a conflict between coverage and capacity.

### D. Problem Formulation

We focus on maximizing the long-term coverage and capacity for the whole serving area by optimizing the transmit power, the reflection phase shift matrix, and the transmission phase shift matrix. The formulated problem can be expressed as follows:

$$\begin{aligned}
& \max_{P_t, \Phi_{\text{Re}, n_s}, \Phi_{\text{Tr}, n_s}} \int_{t=1}^T [\text{Coverage}(t), \text{Capacity}(t)] \quad (14) \\
& \text{s. t. } 0 < P_t(t) \leq P_{\max}, \quad (14a) \\
& 0 < \text{tr}(\Phi_{\delta, n_s}^H \Phi_{\delta, n_s}) < 1, \quad (14b) \\
& 0 < \text{tr}(\Phi_{\text{Re}, n_s}^H \Phi_{\text{Re}, n_s}) + \\
& \quad \text{tr}(\Phi_{\text{Tr}, n_s}^H \Phi_{\text{Tr}, n_s}) \leq 1, \quad (14c)
\end{aligned}$$

where  $P_{\max}$  denotes the permitted maximum transmit power. Constraint (14a) limits the range of the transmit power. According to the energy conservation principle, constraints (14b) and (14c) show that both the energy of different modes and the sum energy of the reflected and transmitted signals is less than one. However, the main difficulty in solving the problem (14) owing to the following reasons. Firstly, the NLOS components for STAR-RIS-assisted links are hard to be determined before the STAR-RISs deployment. Secondly, the distribution weights  $w_{\text{cov}, s_i}(t)$  and  $w_{\text{cap}, s_i}(t)$  at time  $t$  for calculating the coverage and capacity is not a continuous function. Thirdly, with respect to the continuous-time  $t$ , it's difficult to handle infinite variables optimization, since any adjacent time is subjected to the Markov chain. Thus, conventional non-convex optimization methods are not suitable for solving these difficulties. In the next section, the Pareto-based MO-PPO algorithms are invoked to solve this problem.

### III. PO-BASED MO-PPO ALGORITHMS

In this section, we first give a brief introduction to the principles of MDP definition, the PPO algorithm, and PO solution. Then the MDP in the MO-PPO algorithm is exhibited. Finally, the different update strategies of the PO-based MO-PPO algorithm are proposed to obtain the optimal policy applicable to the considered networks.

#### A. Basic Principles

1) *MDP Definition*: For a typical RL problem, it can be defined as an MDP problem. A decision-maker called the agent will execute the action by interacting with the environment. The environment, in return, provides rewards and a new state based on the actions of the agent. In other words, RL presents the agent with rewards whether positive or negative based on its actions instead of teaching the agent how it should do something. Thus, the goal of RL is all about the goal to maximize the reward, where the MDP can be defined as the combination of the Markov reward process (MRP) with values judgement. Mathematically, we define MRP as:

$$R_s = \mathbb{E}[R_{t+1}|S_t], \quad (15)$$

where the total reward  $R_s$  OF MRP denotes the sum of reward  $R_{t+1}$  gets from a particular state  $S_t$ . Then, the MDP can be defined as a tuple  $\langle S, A, p, R \rangle$  with state space  $S$ , action space  $A$ , transition probability  $p$ , and reward  $R$ .

2) *PPO Algorithm*: PPO algorithm is based on trust region policy optimization [28] and utilizes the typical actor-critic architecture. The actor network is to determine the action according to the current state and the critic network, whereas the critic network is to evaluate how well the actor network performs the action. The configuration of the PPO algorithm is shown in Fig. 2. Note that, the design of the architecture is

modularized to separate the cohesion between neural networks, PPO algorithm, and environments. The action-value function in the PPO algorithm is replaced by an advantage function  $\hat{A}_t$  at every  $\bar{T}$  time steps, which is expressed as:

$$\hat{A}_t = \sum_{t=1}^{\bar{T}} \mathbf{Q}_{\pi_{\bar{\theta}}}(S_t, A_t) - V_{\pi_{\bar{\theta}}}(S_t), \quad (16)$$

where  $\mathbf{Q}_{\pi_{\bar{\theta}}}(S_t, A_t)$  is the action-value function at policy  $\pi_{\bar{\theta}}$  with state  $S_t$  and  $A_t$ . The  $V_{\pi_{\bar{\theta}}}(S_t)$  is state-value predicted by the critic network. The update solution of the loss function, i.e., No clipping or penalty (NCP), can be expressed as:

$$\mathcal{L}^{\text{NCP}} = \min_{\bar{\theta}} \mathbb{E}_t \left[ \frac{\pi_{\bar{\theta}^*}(S_t, A_t)}{\pi_{\bar{\theta}}(S_t, A_t)} \hat{A}_t \right], \quad (17)$$

where  $\pi_{\bar{\theta}^*}(\cdot)$  and  $\pi_{\bar{\theta}}(\cdot)$  denote the current policy and old policy. Since a large difference between the new and old policies often leads to destructively large policy [29], there are other two methods invoked for the PPO algorithm, i.e., clipped (CLIP) and Kullback–Leibler (KL) penalty methods. The two methods can be directly expressed as:

$$\begin{aligned}
& \mathcal{L}^{\text{CLIP}} = \\
& \min_{\bar{\theta}} \mathbb{E}_t \left[ \frac{\pi_{\bar{\theta}^*}(S_t, A_t)}{\pi_{\bar{\theta}}(S_t, A_t)} \hat{A}_t, \text{clip} \left( \frac{\pi_{\bar{\theta}^*}(S_t, A_t)}{\pi_{\bar{\theta}}(S_t, A_t)}, 1 - \epsilon, \epsilon \right) \hat{A}_t \right], \quad (18) \\
& \mathcal{L}^{\text{KL}} = \min_{\bar{\theta}} \mathbb{E}_t \left[ \frac{\pi_{\bar{\theta}^*}(S_t, A_t)}{\pi_{\bar{\theta}}(S_t, A_t)} \hat{A}_t - \tilde{\beta} \text{KL}(\pi_{\bar{\theta}^*}(S_t), \pi_{\bar{\theta}}(S_t)) \right], \quad (19)
\end{aligned}$$

where  $\epsilon$  is the probability ratio for the clipped method, and  $\tilde{\beta}$  is a adjustment penalty coefficient for KL method.

3) *PO Solution*: In multi-objective optimization problems, each objective function may have an individual optimal solution, while these solutions usually have significant differences. Therefore, multi-objective optimization with such conflicting objective functions provides a set of optimal solutions, namely, PO solutions [30]. As shown in Fig. 3, considering two conflict objectives, both of which aim to be maximized. The point  $C_1$  represents a solution that  $F_2$  is near-maximum, but  $F_1$  is low, while point  $C_4$  indicates a solution  $F_1$  is near-maximum, but  $F_2$  is small. However, it is difficult to distinguish whether solution  $C_1$  is better than  $C_4$ , or vice versa. In fact, there exist many such solutions belonging to the PO set, which forms a primary PF. Additionally,  $C_5$ ,  $C_6$ , and  $C_7$  are the feasible solutions.  $C_5$  belongs to the second PF, while  $C_6$  and  $C_7$  are the part of the third PF [30].

#### B. MO-PPO Framework

In this work, the locations of STAR-RISs are randomly pre-selected, and the locations of STAR-RISs are not overlapped. The locations stay the same before the training process achieves convergence. Moreover, STAR-RISs are placed along the y-axis direction<sup>3</sup> to ensure that transmit signal from any BS is reflected and transmitted using the same planar of each STAR-RISs.

In the MO-PPO algorithm, the MDP can be represented by a tuple  $\langle \bar{S}, \bar{A}, \mathbf{p}, \bar{R} \rangle$  with state space  $\bar{S}$ , action space  $\bar{A}$ ,

<sup>3</sup>If the STAR-RISs are rotated, the geographic model and system model need to be reconstructed.

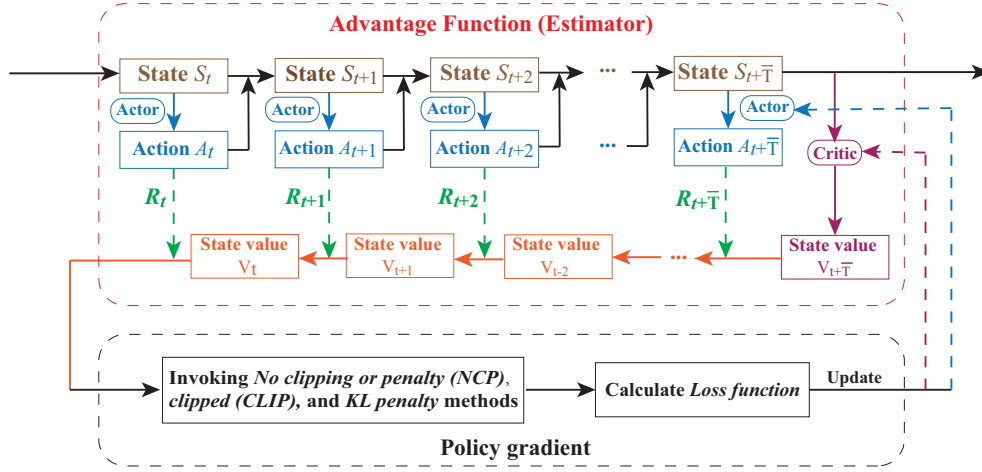


Fig. 2: The framework of PPO algorithm.

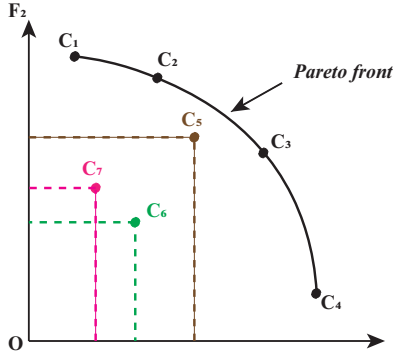


Fig. 3: The Pareto solutions for two objectives.

reward space  $\bar{\mathbf{R}}$ . The  $\mathbf{p}$  is the transition probability matrix indicating the probability of changing the current state to the next state. Define a controller as an agent, which controls both two BSs, to develop the policy from the BSs to sample points via STAR-RISs, i.e., the adjustment policies of phase shifts and transmit power. At each time step  $t$ , the controller observes the state  $\mathbf{S}_t$  from state space  $\bar{\mathbf{S}}$ , and carries out an action  $\mathbf{A}_t$  from action space  $\bar{\mathbf{A}}$ . The received reward  $\mathbf{R} \subseteq \bar{\mathbf{R}}$  is calculated by the current state and action and determines the transition probability to the next state  $\mathbf{S}_{t+1}$ . Additionally, since the locations of STAR-RISs are pre-determined, the distance between any BS and  $n_s$ -th STAR-RISs is fixed. The coverage and capacity are determined by the distance between STAR-RISs and  $s_i$ -th point and the corresponding phase shift of the STAR-RISs, according to the (12) and (13). Thus, the state  $\mathbf{S}_t$  can be defined symbolically as follows:

$$\mathbf{S}_t = [\beta_{\text{Re},n_s}(t), \beta_{\text{Tr},n_s}(t), \Phi_{\text{Re},n_s}(t), \Phi_{\text{Tr},n_s}(t), P_t(t)]. \quad (20)$$

For the action  $\mathbf{A}_t$ , the  $\beta_{\text{Tr},n_s}$  of STAR-RISs is discretized with small step  $z$  as numerous values between  $(0, 1)$ , while the  $\beta_{\text{Re},n_s}$  is determined by  $(1 - \beta_{\text{Tr},n_s})$  based on the energy constraint policy mentioned in [5]. In the MO-PPO algorithm, the category distributions of available locations and phase shifts of STAR-RISs are constructed first. Then, the agent samples phase shifts as an action according to the probability determined by the actor network. The action  $\mathbf{A}_t$  can be expressed as follows:

$$\mathbf{A}_t = [\Delta\beta_{\text{Re},n_s}, \Delta\beta_{\text{Tr},n_s}, \Delta\phi_{\text{Re},n_s}, \Delta\phi_{\text{Tr},n_s}, \Delta P_t]. \quad (21)$$

where  $\Delta\beta_{\text{Re},n_s} \in \{z, 2z, \dots, 1 - z\}$ ,  $\Delta\beta_{\text{Tr},n_s} \in \{1 - z, 1 - 2z, \dots, z\}$ , and  $\Delta\phi_{\delta,n_s} \in \{\phi_{\delta,n_s,1}, \phi_{\delta,n_s,2}, \dots, \phi_{\delta,n_s,K_\delta}\}$  denote the possible values for the transmission amplitude, reflection amplitude, and possible phases for  $n_s$ -th STAR-RISs with mode  $\delta$ , respectively. The  $\Delta P$  is chosen from  $[0, zP_{\max}, 2zP_{\max}, \dots, P_{\max}]$ . For  $k$ -th element, the phase is randomly selected from  $[0, 2\pi)$ . To obtain the maximum transmission coverage and capacity that BSs is able to achieve in the time period  $\mathcal{T}$ , the reward is denoted as the difference of coverage  $\Delta\text{Cov}_{t \rightarrow t+1}$  and capacity  $\Delta\text{Cap}_{t \rightarrow t+1}$  in adjacent time steps, which can be calculated separately and expressed as a vector:

$$\mathbf{R}_t(\mathbf{S}_t, \mathbf{A}_t) = [\Delta\text{Cov}_{t \rightarrow t+1}, \Delta\text{Cap}_{t \rightarrow t+1}]. \quad (22)$$

Additionally, the loss function in the PPO algorithm can be evaluated according to (17), (18), and (19). In this work, we propose a novel framework for the MO-PPO algorithm, where two update strategies, i.e., AVUS and LFUS, are employed for the PO-based MO-PPO algorithms.

### C. AVUS-based MO-PPO Algorithm

In this subsection, we consider the AVUS-based MO-PPO algorithm, where the MO-MDP can be rewritten as  $\langle \bar{\mathbf{S}}, \bar{\mathbf{A}}, \mathbf{p}, \bar{\mathbf{R}}, \Omega, f_\Omega \rangle$ . The  $\Omega$  and  $f_\Omega$  denote the preferences space and the functions of preference, respectively. In this case, a linear preference function is employed, i.e.,  $f_\Omega(\mathbf{R}_t(\mathbf{S}_t, \mathbf{A}_t)) = \bar{\mathbf{w}}^T \mathbf{R}_t(\mathbf{S}_t, \mathbf{A}_t)$ ,  $\bar{\mathbf{w}} \subseteq \Omega$ . All possible returns from MO-MDP are able to form a PF  $\mathcal{F} := \{\hat{\mathbf{R}} \mid \forall \bar{\mathbf{R}} < \hat{\mathbf{R}}, \bar{\mathbf{R}} \subseteq \mathcal{F}^*\}$ , where  $\hat{\mathbf{R}}$  and  $\bar{\mathbf{R}}$ , and  $\mathcal{F}^*$  denote the PO return, non-PO return, and the set of non-PO returns, respectively. For  $\Omega$  in the AVUS, a PF-based convex coverage set  $\mathbf{f}$  can be defined as:

$$\mathbf{f} = \{\hat{\mathbf{R}} \subseteq \mathcal{F} \mid \exists \bar{\mathbf{w}} \subseteq \Omega, \forall \bar{\mathbf{R}} \subseteq \mathcal{F}^*, \bar{\mathbf{w}}^T \hat{\mathbf{R}} \geq \bar{\mathbf{w}}^T \bar{\mathbf{R}}\}. \quad (23)$$

The agent is able to learn a group of policies  $\Pi = \{\pi_{\bar{\theta}^1}, \pi_{\bar{\theta}^2}, \dots\}$  by interacting with the environments. Among them, there exists one linear preference vector  $\bar{\mathbf{w}}$  in policy  $\pi_{\bar{\theta}^*}$  to satisfy:

$$\bar{\mathbf{w}}^T V^{\pi_{\bar{\theta}^*}}(\mathbf{S}_t) \geq \bar{\mathbf{w}}^T V^{\pi_{\bar{\theta}}}(S_t), \quad \exists \bar{\mathbf{w}} \subseteq \Omega, \quad (24)$$



where  $V^{\pi_{\bar{\theta}^*}}(\mathbf{S}_t)$  denote the state-value function with state  $\mathbf{S}_t$ , and  $\pi_{\bar{\theta}}$  denotes other any policy except  $\pi_{\bar{\theta}^*}$ . In the AVUS, the output network policy contains two sub-policies, which are optimized for coverage and capacity over different preferences, respectively. The core point of this strategy is to integrate the action values of all objectives, which are fully based on the convex envelope of the solution front. Here, we provide a theoretical analysis of the AVUS scheme below.

1) *Bellman Operator*: The standard single-objective PPO algorithm [29] utilizes the Bellman expectation operator, where the action value function  $Q_{\pi_{\bar{\theta}}}(S_t, A_t)$  by Bellman optimality operator  $\mathcal{J}$  can be expressed as follows:

$$(\mathcal{J}Q)_{\pi_{\bar{\theta}^*}}(S_t, A_t) = R_t(S_t, A_t) + \gamma \sum_{S' \in \bar{\mathbf{S}}} p(S_t, A_t, S')(\mathcal{H}Q)(S', A'), \quad (25)$$

where  $\gamma$ ,  $p(S_t, A_t, S')$ , and  $(\mathcal{H}Q)(S', A') = \max_{A' \in \bar{\mathbf{A}}} Q_{\pi_{\bar{\theta}^*}}(S', A')$  denote the discount factor, the transition probability from  $S_t$  to  $S'$  by choosing  $A_t$ , and the optimality filter for the next state  $S'$ , respectively. Then, we extend the single-objective PPO algorithm to the MO-PPO algorithm by considering an action-value function space  $\mathcal{Q}$  to estimate expected total rewards under  $\bar{\omega}$  preferences, where  $\mathcal{Q}$  contains all bounded action-value functions  $Q(\mathbf{S}_t, \mathbf{A}_t, \bar{\omega})$ . The corresponding value metric  $\mathcal{D}$  can be defined as follows:

$$\begin{aligned} \mathcal{D}(Q, Q') \\ = \max_{S_t \in \bar{\mathbf{S}}, A_t \in \bar{\mathbf{A}}, \bar{\omega} \in \bar{\Omega}} \|\bar{\omega}^T [Q(S_t, A_t, \bar{\omega}) - Q'(S_t, A_t, \bar{\omega})]\|, \end{aligned} \quad (26)$$

Based on any given policy  $\pi_{\bar{\theta}}$ , the evaluation operator of the action-value function in the MO-PPO algorithm can be defined as follows:

$$\begin{aligned} (\mathcal{J}Q)_{\pi_{\bar{\theta}}}(S_t, A_t, \Omega) &= R_t(S_t, A_t) \\ &+ \gamma \sum_{S' \in \bar{\mathbf{S}}} p(S_t, A_t, S') \sum_{A' \in \bar{\mathbf{A}}} \pi_{\bar{\theta}}(A' | S') Q_{\pi_{\bar{\theta}}}(S', A', \bar{\omega}). \end{aligned} \quad (27)$$

Accordingly, we denote the optimality filter  $\mathcal{H}$  for the MO-PPO action-value function as follows:

$$(\mathcal{H}Q)(S', A', \bar{\omega}) = \max_{A' \in \bar{\mathbf{A}}, \Omega' \in \bar{\Omega}} \Omega'^T Q_{\pi_{\bar{\theta}^*}}(S', A', \bar{\omega}). \quad (28)$$

Intuitively, the filter  $\mathcal{H}$  provides  $Q$  value under given  $S_t$  and  $\bar{\omega}$  while handling the convex envelope of the solution front. The optimality operator  $\mathcal{J}$  for the MO-PPO action function under optimal policy  $\pi_{\bar{\theta}^*}$  can be defined as follows:

$$\begin{aligned} (\mathcal{J}Q)_{\pi_{\bar{\theta}^*}}(S_t, A_t, \Omega) &= \\ R_t(S_t, A_t) &+ \gamma \sum_{S' \in \bar{\mathbf{S}}} p(S_t, A_t, S')(\mathcal{H}Q)_{\pi_{\bar{\theta}^*}}(S', A', \bar{\omega}). \end{aligned} \quad (29)$$

Compared to (25), (29) integrated all the objectives by invoking  $\bar{\omega}$  to update the policy of each objective simultaneously.

**Remark 1.** Compared to the single objective optimization problem, the policy in AVUS contains a preference space  $\bar{\omega}$ , which is utilized to estimate the total rewards under multi-objective and update the whole policy. Note that, each

---

**Algorithm 1** PO-based MO-PPO algorithm, AVUS

---

**Input:**

PPO network structure, preference distribution  $\mathcal{B}$ , path  $\Delta\varpi$  for coefficient  $\varpi$ .

**Output:** The optimal MO-PPO policy network.

**Initialize:** Hyperparameters of PPO network, total epochs  $\bar{U}$  in each update, minibatch size  $M$ , update frequency  $\mathcal{U}$  for MO-PPO algorithm.

**for** iteration = 1, 2,  $\dots$  **do**

Sample a linear preference  $\bar{\omega}$  from  $\mathcal{B}$ .

**for** actor = 1, 2,  $\dots$ ,  $N$  **do**

Run policy  $\pi_{\bar{\theta}}$  in environment for  $T$  time steps.

Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$  for every  $\bar{T}$  updating time.

**end for**

Optimize loss function  $\mathcal{L}$  wrt  $\bar{\theta}$ , with  $\bar{U}$  and  $M \leq \mathcal{U}$ , according to equation (36).

Update parameters  $\bar{\theta} \leftarrow \bar{\theta}^*$ .

Increase  $\varpi$  along the path  $\Delta\varpi$ .

**end for**

---

objective has its own policy instead of sharing a common policy.

2) *Loss Function*: Typically the environment is not known entirely so there is no closed-form solution to obtain optimal action-value and state-value functions. In this case, the advantage estimator can be expressed as (30), where  $\mathbf{R}_t$  and  $V_{\pi_{\bar{\theta}}}(\cdot, \cdot)$  denote the obtained reward at each time step, the output state-value by critic network, respectively. In our proposed strategy, the loss function can be calculated based on the NCP method, CLIP method, and KL penalty method, which are expressed as (31) - (33).

At each update, the optimal method will be selected as follows:

$$\mathcal{L}_1^{\text{optimal}}(\bar{\theta}, \bar{\omega}) = \max\{\mathcal{L}_1^{\text{NCP}}(\bar{\theta}, \bar{\omega}), \mathcal{L}_1^{\text{CLIP}}(\bar{\theta}, \bar{\omega}), \mathcal{L}_1^{\text{KL}}(\bar{\theta}, \bar{\omega})\}. \quad (34)$$

However, owing to a large number of discrete solutions in the optimal PO front, directly optimizing  $\mathcal{L}_1^{\text{NCP/CLIP/KL}}(\bar{\theta}, \bar{\omega})$  in practice is still challenging. To address the difficulty, auxiliary loss functions are invoked and the optimal selection is expressed as follows:

$$\begin{aligned} \mathcal{L}_2^{\text{optimal}}(\bar{\theta}, \bar{\omega}) \\ = \max\{\mathcal{L}_2^{\text{NCP}}(\bar{\theta}, \bar{\omega}), \mathcal{L}_2^{\text{CLIP}}(\bar{\theta}, \bar{\omega}), \mathcal{L}_2^{\text{KL}}(\bar{\theta}, \bar{\omega})\}, \\ = \max\{\bar{\omega}^T \mathcal{L}_1^{\text{NCP}}(\bar{\theta}, \bar{\omega}), \bar{\omega}^T \mathcal{L}_1^{\text{CLIP}}(\bar{\theta}, \bar{\omega}), \bar{\omega}^T \mathcal{L}_1^{\text{KL}}(\bar{\theta}, \bar{\omega})\}. \end{aligned} \quad (35)$$

The  $\mathcal{L}_1^{\text{optimal}}(\bar{\theta}, \bar{\omega})$  is capable of ensuring that predicted action-value closing to any real expected total reward although it may not obtaining the optimal results.  $\mathcal{L}_2^{\text{optimal}}(\bar{\theta}, \bar{\omega})$  is able to pull along the proper direction with better utility. Therefore, to obtain the optimal results, the final loss function can be expressed according to the homotopy optimization [31]:

$$\mathcal{L}^{\text{optimal}}(\bar{\theta}, \bar{\omega}) = \varpi \mathcal{L}_1^{\text{optimal}}(\bar{\theta}, \bar{\omega}) + (1 - \varpi) \mathcal{L}_2^{\text{optimal}}(\bar{\theta}, \bar{\omega}), \quad (36)$$



$$\begin{aligned}\hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}(\bar{\omega}) &= \sum_t^{\bar{T}} \mathbf{Q}_{\pi_{\bar{\theta}}}(S_t, \mathbf{A}_t, \bar{\omega}) - V_{\pi_{\bar{\theta}}}(S_t, \bar{\omega}) \\ &= \mathbf{R}_t + \gamma \mathbf{R}_{t+1} + \gamma^2 \mathbf{R}_{t+2} + \dots + \gamma^{\bar{T}-t+1} \mathbf{R}_{\bar{T}-1} + \gamma^{\bar{T}-t} V_{\pi_{\bar{\theta}}}(\mathbf{S}^{\bar{T}}, \bar{\omega}) - V_{\pi_{\bar{\theta}}}(S_t, \bar{\omega}),\end{aligned}\quad (30)$$

$$\mathcal{L}_1^{\text{NCP}}(\bar{\theta}, \bar{\omega}) = \mathbb{E}_t \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(S_t, \mathbf{A}_t, \bar{\omega})}{\pi_{\bar{\theta}}(S_t, \mathbf{A}_t, \bar{\omega})} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}(\bar{\omega}) \right] \right\}, \quad (31)$$

$$\mathcal{L}_1^{\text{CLIP}}(\bar{\theta}, \bar{\omega}) = \mathbb{E}_t \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(S_t, \mathbf{A}_t, \bar{\omega})}{\pi_{\bar{\theta}}(S_t, \mathbf{A}_t, \bar{\omega})} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}(\bar{\omega}), \text{clip} \left( \frac{\pi_{\bar{\theta}^*}(S_t, \mathbf{A}_t, \bar{\omega})}{\pi_{\bar{\theta}}(S_t, \mathbf{A}_t, \bar{\omega})}, 1 - \epsilon, \epsilon \right) \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}(\bar{\omega}) \right] \right\}, \quad (32)$$

$$\mathcal{L}_1^{\text{KL}}(\bar{\theta}, \bar{\omega}) = \mathbb{E}_t \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(S_t, \mathbf{A}_t, \bar{\omega})}{\pi_{\bar{\theta}}(S_t, \mathbf{A}_t, \bar{\omega})} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}(\bar{\omega}), \tilde{\beta} \text{KL}(\pi_{\bar{\theta}^*}(S_t, \bar{\omega}), \pi_{\bar{\theta}}(S_t, \bar{\omega})) \right] \right\}. \quad (33)$$

where  $\varpi$  is a weight to trade off between  $\mathcal{L}_1^{\text{Optimal}}(\bar{\theta}, \bar{\omega})$  and  $\mathcal{L}_2^{\text{Optimal}}(\bar{\theta}, \bar{\omega})$ . The value of  $\varpi$  is increased from 0 to 1 with step 0.1. The pseudo-code of the AVUS-based MO-PPO algorithm is shown in **Algorithm 1**. To sum up, the AVUS aims to train an agent to recover policies for approaching the entire PF. However, different preference  $\bar{\omega}$  affects the total obtained rewards for coverage and capacity.

#### D. LFUS-based MO-PPO Algorithm

In this subsection, we consider the LFUS-based MO-PPO algorithm, where the multi-task learning (MTL) method is employed. Different from the AVUS, there are multiple gradient policies that need to be updated simultaneously. In the MTL-based MO-PPO problem, the empirical risk minimization formulation is generally followed:

$$\min_{\bar{\theta}} \sum_{m=1}^M \varphi^m \hat{\mathcal{L}}^m(\bar{\theta}), \quad (37)$$

where  $\varphi^m$  and  $\hat{\mathcal{L}}^m(\bar{\theta})$  denote the weights for  $m$ -th task and the empirical loss of  $m$ -th task. Consider two sets of solutions  $\bar{\theta}_1$  and  $\bar{\theta}_2$ , if  $\hat{\mathcal{L}}^1(\bar{\theta}_1) > \hat{\mathcal{L}}^1(\bar{\theta}_2)$  and  $\hat{\mathcal{L}}^2(\bar{\theta}_1) < \hat{\mathcal{L}}^2(\bar{\theta}_2)$ , it is obtained that the two tasks are mutually non-dominated, and therefore belong to the PF. In this case, the MTL problem can be formulated as MO optimization to explore the optimal results for conflicting objectives, where the vector-valued loss  $\mathcal{L}$  are employed as follows:

$$\min_{\bar{\theta}} \mathcal{L}(\bar{\theta}) = \min_{\bar{\theta}} [\hat{\mathcal{L}}^1(\bar{\theta}), \hat{\mathcal{L}}^2(\bar{\theta}), \dots, \hat{\mathcal{L}}^M(\bar{\theta})]^T. \quad (38)$$

Hence, the optimization of equation (38) is to find PO solutions. Define  $\bar{\mathcal{F}} = \{\mathcal{L}(\bar{\theta})\}, \bar{\theta} \in \bar{\Theta}$  as the approximate PF, where  $\bar{\theta}$  and  $\bar{\Theta}$  denote any one set of optimal parameters and all possible sets of optimal parameters. Here, we provide a theoretical analysis of the LFUS scheme below.

1) *Multiple Gradient Descent Algorithm (MGDA)*: To converge to the Pareto stationary (PS) solution problem, the MGDA [32] is a proper method. According to the Karush-Kuhn-Tucker conditions, there exists  $\nu_1, \nu_2, \dots, \nu_M$  such that:

- $\nu_1, \nu_2, \dots, \nu_M \geq 0$ .
- $\sum_{m=1}^M \nu_m = 1$  and  $\sum_{m=1}^M \nu_m \nabla_{\bar{\theta}} \hat{\mathcal{L}}^m(\bar{\theta}) = 0$ .

Before handling the MGDA, the objectives may have values of the different scales, while MGDA is sensitive to the different

ranges. Thus, the following gradient normalization method is invoked to alleviate the value range:

$$\nabla_{\bar{\theta}} \mathcal{L}(\bar{\theta}) = \frac{\nabla_{\bar{\theta}} \mathcal{L}(\bar{\theta})}{\mathcal{L}(\bar{\theta})}, \quad (39)$$

where  $\bar{\theta}'$  is the initial parameters of the model. Consequently, the range of the loss function has been limited to  $[0, 1]$ .

**Definition 1.** A solution  $\bar{\theta}_1$  dominates a solution  $\bar{\theta}_2$  if for all objectives satisfying  $\hat{\mathcal{L}}^m(\bar{\theta}_1) \leq \hat{\mathcal{L}}^m(\bar{\theta}_2)$ , while exists at least one objective satisfying  $\hat{\mathcal{L}}^n(\bar{\theta}_1) < \hat{\mathcal{L}}^n(\bar{\theta}_2)$ ,  $\forall m, n \in \{1, 2, \dots, M\}$ .

**Definition 2.** A solution  $\bar{\theta}_1$  is PO solution while there is no any other solution  $\bar{\theta}_2$  dominates  $\bar{\theta}_1$ .

**Definition 3.** All non-dominated solutions  $\hat{\bar{\theta}}$  are Pareto set.

The solution that satisfies the conditions above is defined as a PS solution, while the PO solution is PS. Thus, the optimization problem can be defined as follows:

$$\min_{\nu_1, \nu_2, \dots, \nu_M} \left\{ \left\| \sum_{m=1}^M \nu_m \nabla_{\bar{\theta}} \hat{\mathcal{L}}^m(\bar{\theta}) \right\|_2^2 \mid \sum_{m=1}^M \nu_m = 1, \nu_m \geq 0 \right\}, \quad (40)$$

where  $\|\cdot\|_2^2$  and  $\nabla_{(\cdot)}$  denote the L2 norm and gradient descent (GD) operator. Define  $\nabla_{\bar{\theta}} \mathcal{L}(\bar{\theta}) = \sum_{m=1}^M \nu_m \nabla_{\bar{\theta}} \hat{\mathcal{L}}^m(\bar{\theta})$ , we have that: if  $\nabla_{\bar{\theta}} \mathcal{L}(\bar{\theta}) = 0$ , the solution is PS; otherwise, it isn't PS and  $\nabla_{\bar{\theta}} \mathcal{L}(\bar{\theta})$  is the general GD vector. Since it has two objectives in problem (14), the equation (40) can be simplified as:

$$\min_{\nu \in [0, 1]} \|\nu \nabla_{\bar{\theta}} \hat{\mathcal{L}}^1(\bar{\theta}) + (1 - \nu) \nabla_{\bar{\theta}} \hat{\mathcal{L}}^2(\bar{\theta})\|_2^2, \quad (41)$$

The optimization problem defined in (41) is equivalent to finding a minimum-norm point in the convex hull, which is a convex quadratic problem with linear constraints. Thus, an analytical solution to equation (41) can be expressed as:

$$\nu = \left\{ \frac{[\nabla_{\bar{\theta}} \hat{\mathcal{L}}^2(\bar{\theta}) - \nabla_{\bar{\theta}} \hat{\mathcal{L}}^1(\bar{\theta})]^T \nabla_{\bar{\theta}} \hat{\mathcal{L}}^2(\bar{\theta})}{\|\nabla_{\bar{\theta}} \hat{\mathcal{L}}^1(\bar{\theta}) - \nabla_{\bar{\theta}} \hat{\mathcal{L}}^2(\bar{\theta})\|_2^2} \right\}_{[0, 1]}, \quad (42)$$

where  $\{\cdot\}_{[0, 1]}$  represents clipping  $\nu$  to  $[0, 1]$ . Alternate optimization of GD vector and  $\nu$  produces different  $\nu$ , which

covers all PO solutions under constraints to approach the PF. According to the system model, it's suitable to select one PO solution as the optimal result.

---

**Algorithm 2** PO-based MO-PPO algorithm, LFUS
 

---

**Input:**

PPO network structure.

**Output:** The optimal MO-PPO policy network.

- 1: **Initialize:** Hyperparameters of PPO network, total epochs  $\bar{U}$  in each update, minibatch size  $M$ , update frequency  $\mathcal{U}$  for MO-PPO algorithm.
  - 2: **for** iteration = 1, 2,  $\dots$  **do**
  - 3:   **for** objective = 1, 2,  $\dots$  **do**
  - 4:    **for** actor = 1, 2,  $\dots$ ,  $N$  **do**
  - 5:      Run policy  $\pi_{\bar{\theta}}$  in environment for  $T$  time steps for each objective.
  - 6:      Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$  for each objective at every  $\bar{T}$  updating time.
  - 7:    **end for**
  - 8:   **end for**
  - 9:   Calculate loss function  $\mathcal{L}$  wrt  $\bar{\theta}$ , with  $\bar{U}$  and  $M \leq \mathcal{U}$ , according to equation (43).
  - 10:   Update  $\bar{\theta}$  by min-norm solver.
  - 11: **end for**
- 

2) *Loss Function:* Our goal is to train one policy containing two sub-policies, where each objective has a specific loss function and shares all parameters. Thus, combining with the PPO algorithm, the loss function for the MO-PPO algorithm based on the NCP method, CLIP method, and KL Penalty method can be expressed as (44) - (46), where  $\hat{A}_t$  is an advantage estimator, it can be expressed as (47). Accordingly, at each update, the optimal method will be selected as follows:

$$\mathcal{L}^{\text{optimal}}(\bar{\theta}) = \max\{\mathcal{L}^{\text{NCP}}(\bar{\theta}), \mathcal{L}^{\text{CLIP}}(\bar{\theta}), \mathcal{L}^{\text{KL}}(\bar{\theta})\}. \quad (43)$$

The pseudo-code of the LFUS-based algorithm is shown in **Algorithm 2**. The proposed algorithm can be applied to other RIS-assisted scenarios, the configuration of networks is constructed according to the input and the training effect by the system model and formulated problem, while the AVUS and LFUS are still available to solve the problem once the formulated problem follows MO-MDP. For example, after providing the input and output of the problem formulated, the number of layers, neurons, and the optimizer can be adjusted for the training effect.

**Remark 2.** According to the theoretical analysis, the LFUS achieves a simpler structure than AVUS by only vectorizing the loss function. For AVUS, two policies are trained, since the action value is parallelly determined according to the preference according to (29). For LFUS, only one policy is trained, since all the objectives share the same loss function in (38). Therefore, the LFUS should have a faster convergence speed than the AVUS.

#### E. Empirical Complexity Analysis

As shown in Tab. I, we analyze the empirical complexity for the AVUS and LFUS, i.e., wall-clock time (time complexity)

and memory utilization (space complexity). For the wall-clock time, AVUS spends 9.852s for 10 episodes and 16.42m for 1000 episodes, while it costs 8.934s for 10 episodes and 14.89m for 1000 episodes in LFUS. For memory utilization, LFUS consumes 108.35MB in total, which saves 7.33MB compared to AVUS. Therefore, the empirical complexity proves that LFUS can achieve less time complexity and space complexity than AVUS.

#### IV. NUMERICAL RESULTS

In this section, we provide numerical results to evaluate the performance of proposed update strategies of MO-PPO algorithms. The simulation is fully performed on the CPU of the Dell Precision 7920 workstation. The configuration of the workstation is listed as follows: 1) CPU: Intel Xeon Bronze 3204 (8.25MB cache, 6 cores, 6 threads, up to 1.90GHz, 85W), 2) GPU: NVIDIA T400, 4GB GDDR6, 3) RAM: 8GB, 1x8GB, DDR4, 2933MHz, ECC, 4) ROM: 256GB, 2.5-inch, SATA, SSD, Class 20. Without loss of generality, a Poisson traffic model is employed to estimate the traffic flows or data sources for the proposed system model. In practice networks, there is a relationship in the traffic load between any adjacent time steps, where the traffic load at the current time step is determined by the previous time step. Based on this traffic model, the normalized coverage probability of observing  $\bar{k}_i$  events and capacity probability of observing  $\hat{k}_i$  events at sample point  $s_i$  at time step 0 can be given by [33]:

$$w_{\text{cov},s_i}(0) = \bar{P}_{s_i}(\bar{k}_i) = \frac{e^{-\bar{\lambda}_i} \frac{\bar{\lambda}_i^{\bar{k}_i}}{\bar{k}_i!}}{\sum_{i=1}^N e^{-\bar{\lambda}_i} \frac{\bar{\lambda}_i^{\bar{k}_i}}{\bar{k}_i!}}, \quad (48)$$

$$w_{\text{cap},s_i}(0) = \bar{P}_{s_i}(\hat{k}_i) = \frac{e^{-\hat{\lambda}_i} \frac{\hat{\lambda}_i^{\hat{k}_i}}{\hat{k}_i!}}{\sum_{i=1}^N e^{-\hat{\lambda}_i} \frac{\hat{\lambda}_i^{\hat{k}_i}}{\hat{k}_i!}}, \quad (49)$$

where  $\bar{\lambda}_n$  and  $\hat{\lambda}_n$  are the average number of events at each sample point  $s_i$ . The parameters of the MO-PPO network and communication network are given in Table. II and Table. III. The construction of both actor and critic networks are the same, the number of layers is three, which consists of the input layer, hidden layer, and output layer. The number of neurons in these three layers is  $2N_s(2K+1)$ , 64,  $2N_s(\frac{P_{\text{max}}+2}{z}+k-1)$ , respectively. The optimizer is Adam while the parameters are randomly initialized. Additionally, there are two benchmarks conceived to evaluate the proposed update strategies:

- **Without STAR-RIS (network performance):** In this benchmark, the BSs serve the whole serving area without the assistance of the STAR-RISs.
- **Fixed weights (algorithm performance):** In this benchmark, the weights of coverage and capacity are fixed as two cases: a) **BM1: weights 0.3 and 0.7;** and b) **BM2: weights 0.6 and 0.4.**

#### A. Approximate PF by Different Proposed Strategies

As shown in Fig. 4, we provide the approximate PF under AVUS and LFUS. Among them, two approximate PFs are depicted are plotted at 3.5GHz and 26GHz signal frequencies using AVUS. Note that, the capacity and coverage are the

$$\mathcal{L}^{\text{NCP}}(\bar{\theta}) = \min_{\nu \in [0,1]} \left\| \nu \mathbb{E}_t^1 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}} \right] \right\} + (1 - \nu) \mathbb{E}_t^2 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}} \right] \right\} \right\|_2^2, \quad (44)$$

$$\begin{aligned} \mathcal{L}^{\text{CLIP}}(\bar{\theta}) = \min_{\nu \in [0,1]} & \left\| \nu \mathbb{E}_t^1 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}, \text{clip} \left( \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)}, 1 - \epsilon, \epsilon \right) \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}} \right] \right\} \right. \\ & \left. + (1 - \nu) \mathbb{E}_t^2 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}, \text{clip} \left( \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)}, 1 - \epsilon, \epsilon \right) \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}} \right] \right\} \right\|_2^2, \end{aligned} \quad (45)$$

$$\begin{aligned} \mathcal{L}^{\text{KL}}(\bar{\theta}) = \min_{\nu \in [0,1]} & \left\| \nu \mathbb{E}_t^1 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}, \tilde{\beta} \text{KL}(\pi_{\bar{\theta}^*}(\mathbf{S}_t), \pi_{\bar{\theta}}(\mathbf{S}_t)) \right] \right\} \right. \\ & \left. + (1 - \nu) \mathbb{E}_t^2 \left\{ \min \left[ \frac{\pi_{\bar{\theta}^*}(\mathbf{S}_t, \mathbf{A}_t)}{\pi_{\bar{\theta}}(\mathbf{S}_t, \mathbf{A}_t)} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}}, \tilde{\beta} \text{KL}(\pi_{\bar{\theta}^*}(\mathbf{S}_t), \pi_{\bar{\theta}}(\mathbf{S}_t)) \right] \right\} \right\|_2^2, \end{aligned} \quad (46)$$

$$\begin{aligned} \hat{\mathbf{A}}_t^{\pi_{\bar{\theta}^*}} &= \sum_t^T \mathbf{Q}_{\pi_{\bar{\theta}}}(\mathbf{S}_t, \mathbf{A}_t) - V_{\pi_{\bar{\theta}}}(\mathbf{S}_t) \\ &= \mathbf{R}_t + \gamma \mathbf{R}_{t+1} + \gamma^2 \mathbf{R}_{t+2} + \dots + \gamma^{T-t+1} \mathbf{R}_{T-1} + \gamma^{T-t} V_{\pi_{\bar{\theta}}}(\mathbf{S}_T) - V_{\pi_{\bar{\theta}}}(\mathbf{S}_t). \end{aligned} \quad (47)$$

TABLE I: Resource footprint for AVUS and LFUS

Policy	Time for 10 episodes	Time for 1000 episodes	Memory utilization in MB
AVUS	~9.852s	~16.42m	~115.68
LFUS	~8.934s	~14.89m	~108.35

TABLE II: Simulation parameters for MO-PPO algorithm

Parameter	Description	Value
$\mathcal{E}$	The maximum number of episodes	10000
$\mathcal{T}$	The maximum of time steps in each episode	5000
$\mathcal{U}$	Update frequency for MO-PPO algorithm	10
$\overline{U}$	The number of epochs in each update	10
$\overline{E}$	Clipped parameter for MO-PPO algorithm	0.2
$\eta$	Discount factor	0.99
$\psi_a$	Learning rate for actor network	0.0001
$\psi_c$	Learning rate for critic network	0.003
$\varpi$	Initial coefficient for updating action-value strategy	0.1
$\Delta \varpi$	Step for the coefficient of updating action-value strategy	0.001

cumulated results in a time period, where the optimized weights for coverage and capacity are dynamic in the proposed strategies. Compared to **BM1** and **BM2**, the coverage and capacity of the solutions on the two fronts both satisfy the PO definition, where at least one of them is better than the benchmarks. It is obtained that a dynamic combination for CCO in a time period is better than the fixed assignment of coverage and capacity. Moreover, the performance of different frequencies on STAR-RIS is an interesting question. When the system bandwidth is the same, mmWave is able to provide better capacity due to channel and frequency characteristics, while sub-6 GHz provides better coverage. Here, the channel model of the mmWave signal only considers the LOS component in (7) - (8). According to the proposed strategy, we randomly

select one result from approximate PF based on AVUS and LFUS for discussion.

### B. Convergence of MO-PPO Algorithm with Proposed Strategies

In Fig. 5, the convergence of the MO-PPO algorithm under proposed update strategies is demonstrated. Note that, to evaluate the performance of proposed algorithms, the learning curves are obtained by ten times repeated training. It can be observed from Fig. 5 that proposed strategies and benchmarks are capable of achieving convergence. Among them, the AVUS converges the slowest, but its cumulative reward is the largest, while the LFUS has a comparable convergence speed, but the cumulative reward is slightly higher. Compared to the benchmarks, both proposed algorithms are able to achieve

TABLE III: Simulation parameters for communication networks

Parameter	Description	Value
$\bar{\lambda}_{\text{cov}}$	Average number of events for coverage	5
$\bar{\lambda}_{\text{cap}}$	Average number of events for capacity	64
$C$	Path loss when $d = 1\text{m}$	-30dB
$n^2$	Noise power variance	$9 \times 10^{-12}\text{mW} \approx -140.46\text{dBW}$
$R_{th}$	Minimal RSRP for all sample points	$0.23\text{mW} \approx -36.38\text{dBW}$
$P_{t,\text{max}}$	Maximum transmit power	$200\text{mW} = 23.01\text{dBm}$
$\alpha_{aR}$	Rician factor for channel from $a$ -th BS to $n_s$ -th STAR-RISs	3dB
$\alpha_{RP}$	Rician factor for channel from $n_s$ -th STAR-RISs to $s_i$ -th sample point	3dB
$\alpha_{aP}$	Rician factor for channel from $a$ -th BS to $s_i$ -th sample point	3dB
$\gamma_{a,n_s}$	Path loss factor for channel from $a$ -th BS to $n_s$ -th STAR-RISs	3.5
$\gamma_{n_s,s_i}$	Path loss factor for channel from $n_s$ -th STAR-RISs to $s_i$ -th sample point	2.8
$\gamma_{a,s_i}$	Path loss factor for channel from $a$ -th BS to $n_s$ -th STAR-RISs	2.2
$z$	Discrete step for amplitude of STAR-RISs	7m
$h_b$	Height of BS	7m
$R_g$	Length of each grid	1m

better performance than the benchmarks in cumulated rewards or convergence speed. Furthermore, compared to other RL algorithms, e.g., deep deterministic policy gradient (DDPG), the DDPG has the fastest convergence speed and achieves the same performance as LFUS. This result proves the correctness of **Remark 2** from practice.

### C. Optimal Coverage and Capacity with Proposed Strategies

In this subsection, we will discuss the impact of the number of sample grids, the number of STAR-RISs, the number of elements in STAR-RISs, and the size of STAR-RISs on the selected optimal coverage and capacity.

1) *Impact of the Number of Sample Grids*: Fig. 6 characterizes the optimized coverage and capacity versus different total grids. In Fig. 6(a), it is observed that the coverage and capacity gains of all cases present decreasing trend with the upgrading of total grids. Specifically, the maximum decreasing gain of coverage among the proposed algorithms and fixed weight-based solutions is 9.23dB, while that capacity can achieve 10.21 dB. The reasons for these results are that compared with other sampling points, the fast fading channel characteristics of sampling points from BS and STAR-RISs make the received RSRP by far grids unable to reach  $R_{th}$ . As a result, both coverage and capacity of the four cases present a downward trend. Additionally, compared to the "Without STAR-RISs" case, the proposed strategies and benchmarks show better performance. This is because the STAR-RISs proactively transmit and reflect the signal to the farther grid with less consumption. To sum up, in the case of only changing the total number of sampling points, the coverage and capacity changes are positively correlated with the total grid changes. Moreover, the

proposed update strategies outperform the benchmarks, while the performance of the AVUS is better than the LFUS.

2) *Impact of the Number of STAR-RISs*: Fig. 7 depicts the optimized coverage and capacity versus the different numbers of STAR-RISs. As shown in Fig. 7(a), the coverage of all cases keeps growing steadily as the number of STAR-RISs increases. When the number of STAR-RISs  $N_s$  reaches 4, the coverage of the **BM1** and **BM2** case can be promoted to over 0.4, and both proposed update strategies can arrive at over 0.6. This is because, with the increase in the number of STAR-RISs, STAR-RISs can help to compensate the received RSRP of some sample points to reach  $R_{th}$ . For the capacity depicted in Fig. 7(b), the gain of capacity between AVUS and **BM2** case achieves 23.48dB, while the gain between LFUS and **BM1** only arrives at 0.31dB. This is because the STAR-RISs can compensate for the severe attenuation of channels from BSs to sample points, which indicates the effectiveness of STAR-RISs. Also, compared to the "Without STAR-RISs" case, the STAR-RISs are able to improve the coverage and capacity of the whole serving area. The gap between any multi-objective optimization solution (fixed weights or proposed strategies) and the "Without STAR-RISs" case keeps enlarging with the increase of the number of STAR-RISs. To sum up, it can be proved that the proposed update strategies also outperform the benchmarks for optimizing coverage and capacity. Since the STAR-RISs have presented their ability to improve spectrum utilization, the "Without STAR-RISs" case will not be discussed in the following subsections.

3) *Impact of the Number of Element in STAR-RISs*: Fig. 8 describes the optimized coverage and capacity versus the different number of elements in STAR-RISs. It can be observed that the coverage shows a slight change in Fig. 8(a).

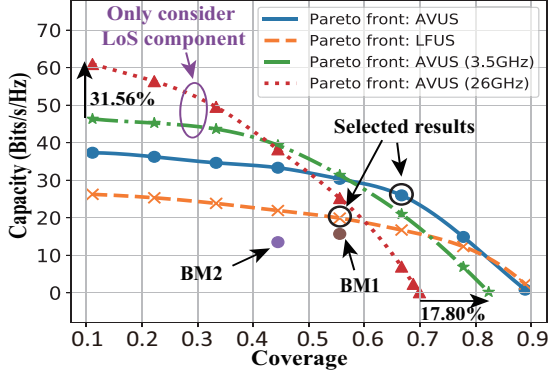


Fig. 4: Approximate PF with different strategies,  $N_s = 3$ ,  $N = 9$ ,  $K = 8 \times 10^2$ ,  $I_{h_{n_s}} = 1$ .

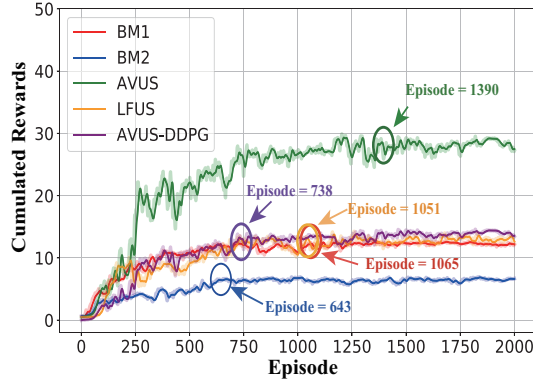
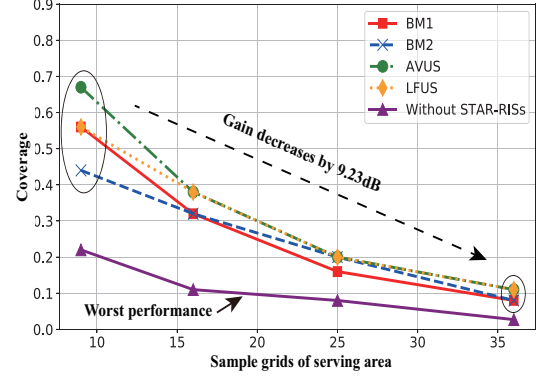
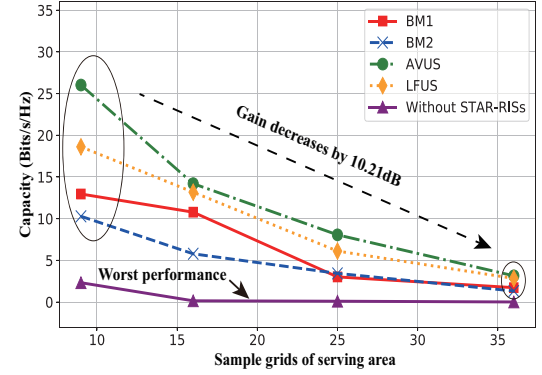


Fig. 5: Learning curves (average convergence for ten times repeated training) for the MO-PPO algorithm with fixed weights, AVUS, and LFUS, initial  $P_t = 2.1\text{mW}$ ,  $N_s = 3$ ,  $N = 9$ ,  $K = 8 \times 10^2$ ,  $I_{h_{n_s}} = 1$ .

The maximum gains among the optimized capacity of four cases in Fig. 8(b) are able to achieve 11.01dB when the number of elements in STAR-RISs increases to 36. It proves that the different number of elements in STAR-RISs bring a huge impact on optimizing capacity. This is because the role of each element is to transmit the BS signal to each sampling point while increasing the number of elements of STAR-RISs is adding multiple links to reduce loss. Compared with increasing the number of STAR-RISs, increasing the number of elements does not change the channel fast-fading characteristics of distant sample points. Also, in order to verify the effectiveness of the mode-splitting protocol in the system considered, we compare it with the PS and TS protocols. As shown in Fig. 8, the mode-splitting-based AVUS case outperformance the "AVUS-TS" case, while the gap between the mode-splitting-based AVUS optimization solution and the "AVUS-TS" case keeps enlarging with the increase in the number of STAR-RISs. This is because MS is able to make full use of the entire available communication time compared with TS. For the "AVUS-PS", there is no big difference between them. This is because the MS can be regarded as a special case PS.



(a) The optimized coverage under different numbers of sample grids of the serving area.



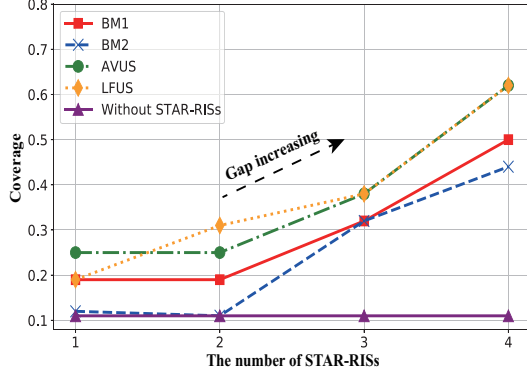
(b) The optimized capacity under different numbers of sample grids of the serving area.

Fig. 6: The optimized coverage and capacity for the MO-PPO algorithm with fixed weights, AVUS, and LFUS with sample grids  $N$  of the serving area,  $N_s = 3$ ,  $K = 8 \times 10^2$ ,  $I_{h_{n_s}} = 1$ .

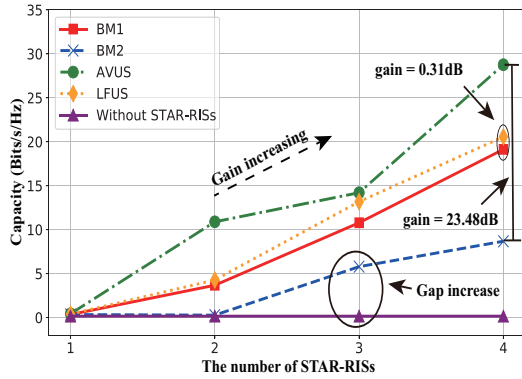
Moreover, for coverage, the LFUS outperforms the AVUS. It proves that when changing elements in STAR-RISs, the LFUS has a priority to be employed for only optimizing coverage. However, for both coverage and capacity optimization, it can be obtained that the AVUS is better than the LFUS. Also, the proposed update strategies both outperform the benchmarks.

4) *Impact of the Physical Size of STAR-RISs*: To evaluate the impact of the physical size of STAR-RISs on optimizing the coverage and capacity, the height  $h_{n_s}$  and width  $\omega_{n_s}$  of the STAR-RISs module are taken out for discussion. In this scenario, the number of STAR-RISs  $N_s$ , the number of total grids  $N$ , and the number elements in STAR-RISs  $K$  are defined as:  $N_s = 2$ ,  $N = 16$ . The number of elements  $K$  are increased linearly with the physical size of the STAR-RISs<sup>4</sup>, and the  $M_a = 6.25 \text{ cm}^2$ . According to the  $h_b$  and  $R_g$ , the threshold of (1) and (2) can be calculated as 1m and 4m. Since  $I_{h_{n_s}} = 1$  has been discussed before, the other three scenarios are further considered as follows:

<sup>4</sup>The effective aperture of each STAR-RIS element keep the same as the area of each element won't be changed.



(a) The optimized coverage under different number of STAR-RISs.

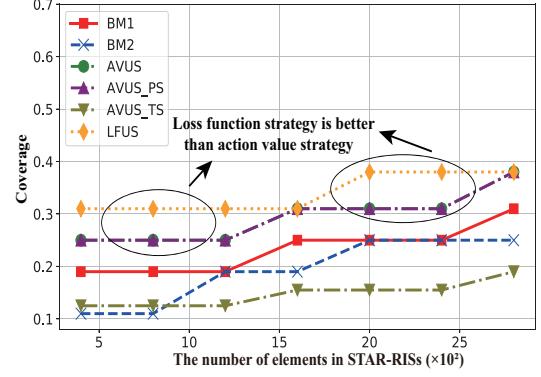


(b) The optimized capacity under different number of STAR-RISs.

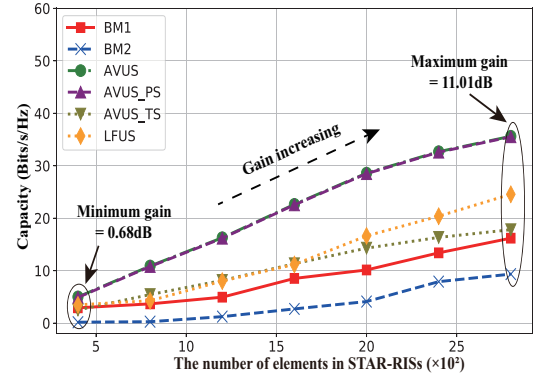
Fig. 7: The optimized coverage and capacity for the MO-PPO algorithm with fixed weights, AVUS, and LFUS under different number  $N_s$  of STAR-RISs,  $N = 16$ ,  $K = 8 \times 10^2$ ,  $I_{h_{n_s}} = 1$ .

- **Case 1:** Width of STAR-RISs are larger than the threshold,  $I_{\omega_{n_s}} = 0$
- **Case 2:** Width of STAR-RISs are smaller than the threshold,  $I_{\omega_{n_s}} = 1$
- **Case 3:** Height of STAR-RISs are smaller than the threshold,  $I_{h_{n_s}} = 0$

Fig. 9 demonstrate the optimized coverage and capacity for the MO-PPO algorithm with fixed weights, AVUS, and LFUS under the different physical sizes of STAR-RISs. Fig. 9(a) provides the changes of **Case 1**. In this case, there is at least one direct link between BS and any given sample point. When the height is also below the threshold, all sample points can have direct links with two BSs. Otherwise, one of the direct links among some sample points and BSs may be blocked. The coverage and capacity sharply fall down while the height of the STAR-RISs module passes over the threshold of 1m. This is because the number of direct links is a significant part to determine the strength of the received RSRP of each sample point. The upgrading number of direct links will increase the probability of reaching  $R_{th}$  at each sampling point. For only considering capacity, the performance of



(a) The optimized coverage with different numbers of elements of STAR-RISs.

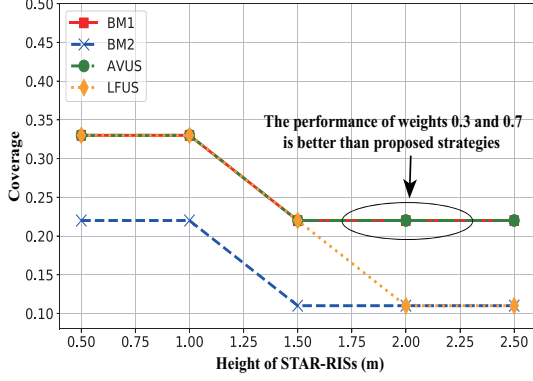


(b) The optimized capacity with different numbers of elements of STAR-RISs.

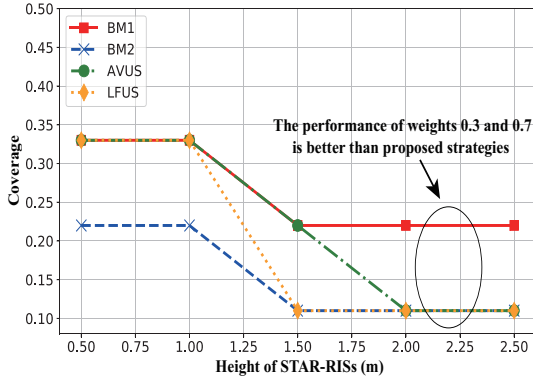
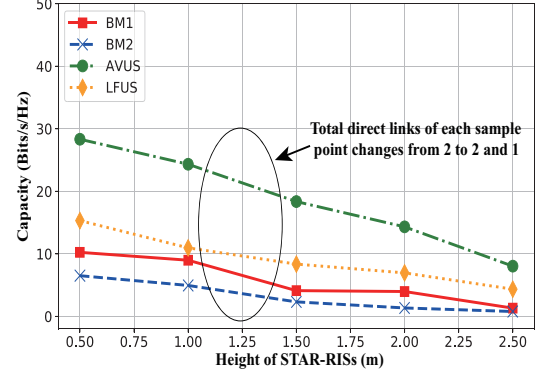
Fig. 8: The optimized coverage and capacity for the MO-PPO algorithm with fixed weights, AVUS, and LFUS with different numbers of elements  $K$  of STAR-RISs,  $N_s = 2$ ,  $N = 16$ ,  $I_{h_{n_s}} = 1$ .

proposed update strategies is better than benchmarks, while the AVUS outperforms the LFUS. For only considering coverage, the performance of proposed strategies cannot present better performance than **BM1**.

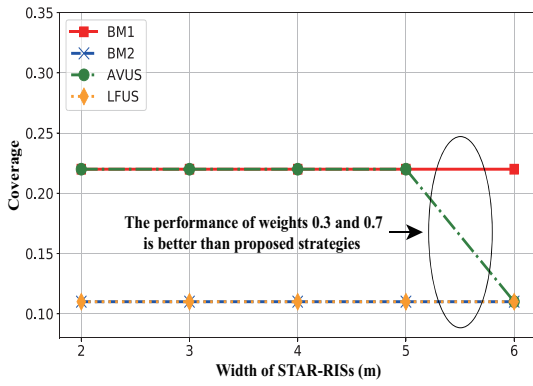
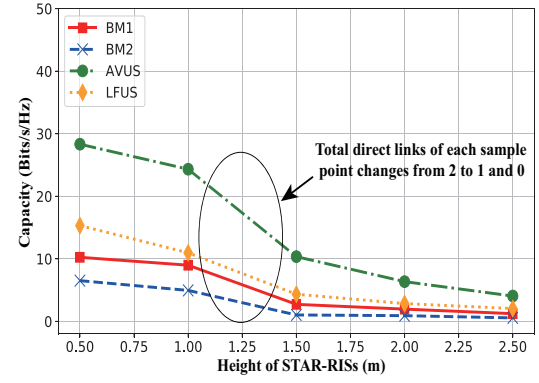
Fig. 9(b) provides the changes of **Case 2**. In this case, the number of direct links between BS and any given sample point can be 0, 1, and 2, which determines by the height of the STAR-RISs module. When the height is also below the threshold, all sample points can have direct links with two BSs. Otherwise, there is at most one direct link between sample points and BSs. The coverage and capacity dramatically decrease while the height of the STAR-RISs module passes over the threshold of 1m. This is because the locations of STAR-RISs determine that the direct links between the sample points and BSs are only 0 or 1. Different from the **Case 1**, the optimized coverage for the proposed update strategies is between benchmarks. It may indicate that the direct links play an important part in receiving RSRP, which needs to be further explored. Additionally, for only considering



(a) Case 1: The optimized coverage and capacity under different height of STAR-RISs,  $\omega_{n_s} = 2\text{m}$ .



(b) Case 2: The optimized coverage and capacity under different height of STAR-RISs,  $\omega_{n_s} = 6\text{m}$ .



(c) Case 3: The optimized coverage and capacity under different width of STAR-RISs,  $h_{n_s} = 2\text{m}$ .

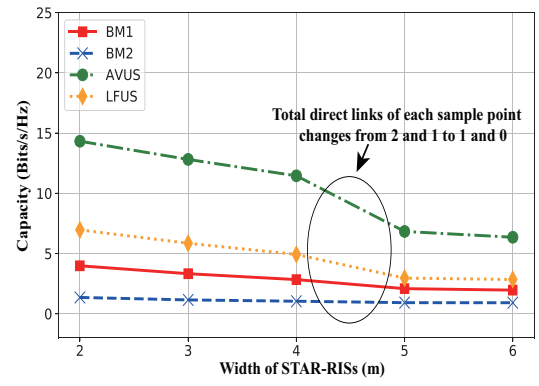


Fig. 9: The optimized coverage and capacity for the MO-PPO algorithm with fixed weights, AVUS, and LFUS under different physical sizes of STAR-RISs,  $N_s = 2$ ,  $N = 16$ .



coverage, the performance of proposed strategies presents worse performance than **BM1**. But considering both coverage and capacity, the proposed update strategies are acceptable in **Case 2**.

Fig. 9(c) provides the optimized coverage and capacity of **Case 3**. In this case, the height of the STAR-RISs module is fixed, which indicates that the direct links between sample points and BSs can be 0, 1, and 2. The number of direct links is 2 and 1, while the width of the STAR-RISs module is below the threshold. Otherwise, the number of direct links is 1 and 0. The capacity also shows a sharp falling down while the width goes over 4m. This is because the locations of STAR-RISs make the direct links between the sample points and BSs 0 or 1. Same with the **Case 2**, for only considering coverage, the performance of proposed strategies presents worse performance than **BM1**. Also, when considering both coverage and capacity, the proposed update strategies can be accepted.

## V. CONCLUSION

In this paper, the coverage and capacity were modelled by considering the geographic property. Based on the model, we proposed a new framework for CCO in STAR-RIS-assisted wireless networks, by optimizing the transmit power, the reflection phase shift matrix, and the transmission phase shift matrix. In order to simultaneously optimize the coverage and capacity, an AVUS for the MO-PPO algorithm was investigated to solve the CCO problem, whose goal was to integrate action value for both coverage and capacity, which shared the same loss function. However, it had strict requirements on the computation resource thereby increasing the cost of the hardware. To handle this problem, another update strategy, i.e., the LFUS, was proposed to update the MO-PPO algorithm with an integrated loss function of coverage and capacity, whose goal was to consider the two-loss function for coverage and capacity. LFUS was able to dynamically assign the weights by a min-norm solver at each update for the MO-PPO algorithms. The numerical results proved that the investigated update strategies were able to provide more efficient solutions than the fixed-weight MOO algorithms. In addition, the coverage and capacity of wireless networks can be enhanced simultaneously with limited energy consumption since STAR-RISs had passive beamforming. In practice, multi-antenna BSs are usually deployed to improve the efficiency of the communication system by joint design active and passive beamforming, as well as considering the effect brought in practical imperfect CSI cases, which can be our future work on STAR-RIS-assisted networks.

## REFERENCES

- [1] X. Gao, W. Yi, A. Agapitos, H. Wang, and Y. Liu, "Coverage and Capacity Optimization in STAR-RISs Assisted Networks: A Machine Learning Approach," *arXiv preprint arXiv:2204.06390*, 2022.
- [2] Y. Liu et al., "Reconfigurable Intelligent Surfaces: Principles and Opportunities," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1546-1577, thirdquarter 2021.
- [3] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously Transmitting And Reflecting (STAR) RIS Aided Wireless Communications," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 5, pp. 3083-3098, May 2022.
- [4] X. Gao, Y. Liu, X. Liu and L. Song, "Machine Learning Empowered Resource Allocation in IRS Aided MISO-NOMA Networks," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 5, pp. 3478-3492, May 2022.
- [5] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. -S. Alouini and R. Zhang, "Wireless Communications Through Reconfigurable Intelligent Surfaces," *IEEE Access*, vol. 7, pp. 116753-116773, 2019.
- [6] J. Xu et al., "Simultaneously Transmitting and Reflecting Intelligent Omni-Surfaces: Modeling and Implementation," *IEEE Veh. Technol. Mag.*, vol. 17, no. 2, pp. 46-54, June 2022.
- [7] C. Zhang, W. Yi, Y. Liu, Z. Ding, and L. Song, "STAR-IOs Aided NOMA Networks: Channel Model Approximation and Performance Analysis," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 9, pp. 6861-6876, Sept. 2022.
- [8] L. Jorgueski, A. Pais, F. Gunnarsson, A. Centonza, and C. Willcock, "Self-organizing networks in 3GPP: Standardization and future trends," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 28-34, Dec. 2014.
- [9] E. Balevi and J. G. Andrews, "Online Antenna Tuning in Heterogeneous Cellular Networks With Deep Reinforcement Learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1113-1124, 2019.
- [10] M. Aldababsa, A. Khaleel, and E. Basar, "Simultaneous Transmitting and Reflecting Intelligent Surfaces-Empowered NOMA Networks," *arXiv preprint arXiv:2110.05311*, 2021.
- [11] J. Zuo, Y. Liu, Z. Ding, L. Song and H. Vincent Poor, "Joint Design for Simultaneously Transmitting And Reflecting (STAR) RIS Assisted NOMA Systems," *IEEE Trans. Wirel. Commun.*, 2022, doi: 10.1109/TWC.2022.3197079.
- [12] P. P. Perera, V. G. Warnasooriya, D. Kudathanthirige and H. A. Suraweera, "Sum Rate Maximization in STAR-RIS Assisted Full-Duplex Communication Systems," *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022.
- [13] H. Niu, Z. Chu, F. Zhou, P. Xiao, and N. Al-Dhahir, "Weighted Sum Rate Optimization for STAR-RIS-Assisted MIMO System," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2122-2127, Feb. 2022.
- [14] 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 Wirel. Commun.*, vol. 28, no. 6, pp. 102-109, December 2021.
- [15] T. Wang, M. -A. Badiu, G. Chen and J. P. Coon, "Performance Analysis of IOS-Assisted NOMA System with Channel Correlation and Phase Errors," *IEEE Trans. Veh. Technol.*, 2022, doi: 10.1109/TVT.2022.3193198.
- [16] C. Wu, Y. Liu, X. Mu, X. Gu, and O. Dobre, "Coverage characterization of STAR-RIS networks: NOMA and OMA," *IEEE Commun. Lett.*, vol. 25, no. 9, pp.3036-3040, 2021.
- [17] A. Asghar, H. Farooq, and A. Imran, "Concurrent Optimization of Coverage, Capacity, and Load Balance in HetNets Through Soft and Hard Cell Association Parameters," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8781-8795, Sept. 2018.
- [18] N. Dandanov, H. Al-Shatri, A. Klein, and V. Poulkov, "Dynamic Self-Optimization of the Antenna Tilt for Best Trade-off Between Coverage and Capacity in Mobile Networks," *Wirel. Pers. Commun.*, vol. 92, pp. 251-278, 2017.
- [19] M. Skocaj, L. Amorosa, G. Ghinamo, G. Muratore, D. Micheli, F. Zabini, and R. Verdone, "Cellular Network Capacity and Coverage Enhancement with MDT Data and Deep Reinforcement Learning," *Comput. Commun.*, vol. 195, pp. 403-415, 2022.
- [20] R. Dreifuerst et al., "Optimizing Coverage and Capacity in Cellular Networks using Machine Learning," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 8138-8142.
- [21] R. Yang, X. Sun, and K. Narasimhan, "A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation," *Adv. Neural Inf. Process. Syst.*, pp. 1-27, 2019.
- [22] O. Sener, and V. Koltun, "Multi-task learning as multi-objective optimization", *Adv. Neural Inf. Process. Syst.*, pp. 31-45, 2018.
- [23] Y. Liu, W. Huangfu, H. Zhang, and K. Long, "An efficient stochastic gradient descent algorithm to maximize the coverage of cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 7, pp. 3424-3436, Jul. 2019.
- [24] Y. Liang, R. Long, Q. Zhang, J. Chen, H. V. Cheng, and H. Guo, "Large Intelligent Surface/Antennas (LISA): Making Reflective Radios Smart," *J. Commun. Netw.*, vol. 4, no. 2, pp. 40-50, Jun. 2019.
- [25] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Joint Deployment and Multiple Access Design for Intelligent Reflecting Surface Assisted Networks," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 10, pp. 6648-6664, Oct. 2021.
- [26] E. Bjornson and L. Sanguinetti, "Rayleigh fading modeling and channel hardening for reconfigurable intelligent surfaces," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 4, pp. 830-834, Apr. 2021.
- [27] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in

- wireless communication,” *IEEE Trans. Wirel. Commun.*, vol. 18, no. 8, pp. 4157-4170, Aug. 2019.
- [28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” *Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, pp. 1889-1897, 2015.
  - [29] L. Watson and R. Haftka, “Modern homotopy methods in optimization,” *Comput. Methods Appl. Mech. Eng.*, vol. 74, no. 3, pp. 289-305, 1989.
  - [30] K. Deb, “Multi-objective evolutionary algorithms: Introducing bias among Pareto-optimal solutions,” *Adv. Evol. Comput.*, pp. 263-292. Springer, Berlin, Heidelberg, 2003.
  - [31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
  - [32] J. Désidéri, “Multiple-gradient descent algorithm (MGDA) for multi-objective optimization,” *Comptes. Rendus. Math.*, vol. 350, no. 5, pp. 313-318, 2012.
  - [33] C. Guerin, H. Nyberg, O. Perrin, S. Resnick, H. Rootzén, C. Starica, “Empirical testing of the infinite source poisson data traffic model,” *Stoch. Models*, vol. 19, no. 2, pp. 151-200, 2003.