A Meta-Learning Approach for Energy-Efficient Resource Allocation and Antenna Selection in STAR-BD-RIS Aided Wireless Networks

Armin Farhadi, Roya Hatami, Mohammad Robat Mili, Christos Masouros, *Fellow, IEEE*, and Mehdi Bennis, *Fellow, IEEE*

***Abstract*—This paper focuses on a wireless network that utilizes beyond diagonal reconfigurable intelligent surfaces (BD-RIS). In this network, multiple BD-RISs assist a multi-antenna base station (BS) with two sectors that simultaneously transmit and reflect signals to single-antenna users. The goal is to maximize energy efficiency by jointly optimizing beamforming at the BS, the BD-RISs' matrix, and antenna selection under the maximum power budget at the BS, BD-RISs' matrix, and antenna selection constraints. The formulated problem is non-convex and challenging to be solved optimally. To address this difficulty, we propose a meta-soft actor critic (Meta-SAC) algorithm, which enables the BS to adjust its beamforming capabilities and BD-RISs' matrix and assign antennas to users. Simulation results demonstrate the superiority of Meta-SAC in comparison with other meta algorithms and a reasonable response compared to the convex optimization benchmark. We also study the influence of system model parameters on the objective function of the proposed optimization problem. In addition, the results show that the multi-BD-RIS system reaches a higher energy efficiency and data rate compared to the provided benchmarks.**

***Index Terms*—beyond diagonal RIS (BD-RIS), meta-learning, soft actor critic (SAC), resource management, antenna selection.**

I. Introduction

Reconfigurable Intelligent Surfaces (RISs) are key to improving spectral and energy efficiency in 6G networks by adjusting signal amplitude and phase, enhancing wireless performance, and reducing energy consumption [1]. Traditional RISs, which only reflect signals, can limit performance if users are outside the reflection area. To address this, Simultaneously Transmitting and Reflecting RISs (STAR-RISs) allow both transmission and reflection, thus improving overall system performance [2].

Former RISs are modeled as diagonal phase shift matrices, with each element connected to its load. Recently, beyond diagonal RIS (BD-RIS) architectures, which are not limited to diagonal matrices, have emerged. BD-RIS can be categorized into three types: group/fully-connected architectures modeled as block diagonal matrices [3], dynamically group-connected architectures adapting to CSI [3], and architectures with non-diagonal phase shift matrices where signals can reflect between elements [3], [4]. Each type offers enhanced performance and broader coverage compared to traditional RIS [4].

Studies on STAR-RIS systems, such as [5], explored minimizing computation errors through joint optimization of power and beamforming. Other research, including [6] and [7], investigated maximizing data rates and minimizing power consumption in IRS-assisted systems using semidefinite programming and alternating optimization. For BD-RIS, [8] introduced new architectures with reconfigurable impedance networks, while [9] and [3] examined discrete reflection coefficients and integrated rate-splitting multiple access (RSMA) to improve coverage and performance.

Existing literature often uses optimization-based algorithms for resource management in RIS or BD-RIS-assisted wireless networks. However, since these problems are usually non-convex, solutions from these algorithms may be suboptimal. Deep reinforcement learning (DRL) offers a promising alternative but is typically suited for static environments, making it less effective for the dynamic nature of next-generation wireless networks. To address this, recent research combines meta-learning with DRL methods [10], [11], which improves convergence speed, performance, and robustness to environmental changes, making it suitable for beyond 5G and 6G networks [12].

Most existing research on BD-RIS focuses on specific aspects of its application. This inspires us to develop a novel system model that addresses broader requirements and emerging challenges in next-generation communication systems. In this letter, we propose a meta-DRL algorithm that jointly optimizes the STAR-BD-RIS matrix, antenna assignment, and beamforming matrix to maximize energy efficiency in a multi-STAR-BD-RIS assisted system. Among DRL methods, we choose soft actor critic (SAC) due to its ability to provide solutions for continuous action spaces with lower complexity. To take advantage of meta-learning, we combine the SAC method with meta-learning and propose a *Meta-SAC* algorithm. To address the optimization problem with both continuous and discrete variables, we employ a quantization strategy. This method allows us to solve the problem efficiently, balancing complexity and computational time. Also, because we have some constraints, we define a hybrid reward function to check constraint satisfaction and construct a proper learning strategy. Simulation results illustrate that the multi-STAR-BD-

A. Farhadi is with the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran (e-mail: armin.farhadi@ut.ac.ir). R. Hatami is with Shahrood University of Technology, Shahrood, Iran (e-mail: roya.hatami@gmail.com). M. R. Mili is with Pasargad Institute for Advanced Innovative Solutions (PIAIS), Tehran, Iran, (email: mohammad.robatmili@gmail.com). Christos Masouros is with the Department of Electronic and Electrical Engineering, University College London, London, UK (e-mail: c.masouros@ucl.ac.uk). M. Bennis is with the Center for Wireless Communications, University of Oulu, Oulu, Finland (email: {mehdi.rasti,mehdi.bennis}@oulu.fi).

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2025.3543780

RIS outperforms the single-STAR-BD-RIS assisted systems in terms of energy efficiency. Moreover, simulation results show the superiority of Meta-SAC under different system model parameters, and we compare it with different meta-learning algorithms and convex optimization benchmark.

II. SYSTEM MODEL

Consider a multi-STAR-BD-RIS assisted system with two sectors to cover the entire space, like STAR-RIS, as illustrated in Fig. 1. In this system model a N-antenna BS transmits information to a set of $\mathcal{M} = \{1, \cdots, M_{ts}, M_{ts}+1, \cdots, M_{ts}+M_{rs}\}$ single-antenna users. The coverage area of the STAR-BD-RISs are divided into two sectors: the transmission sector where M_{ts} users are located, and the reflection sector where M_{rs} users are located. The direct link between the BS and users may have weak quality, for instance, when obstacles such as buildings block the direct link. Hence, it is assumed that a set of $\mathcal{R} = \{1, \cdots, R\}$ STAR-BD-RISs assist the BS in providing service to the end-users. Each sector of STAR-BD-RIS is equipped with K antennas, represented by the set $\mathcal{K} = \{1, \cdots, K\}$. Additionally, we assume that the channel models between the BS, STAR-BD-RISs, and users follow a flat-fading model. Furthermore, it is assumed that the BS and BD-RISs have perfect channel state information.

In this system, $\mathbf{w}_{m_s}^{BU} \in \mathbb{C}^{N\times 1}$ represents the channel vector from the BS to user m_s, where $s \in \mathcal{S}$ and $\mathcal{S} = \{ts, rs\}$, with ts and rs standing for the transmission and reflection sectors, respectively. Besides, $\mathbf{h}_{r,m_s}^{RU} \in \mathbb{C}^{K\times 1}$ denotes the channel vector from rth STAR-BD-RIS to user m_s, and $\mathbf{V}_r^{BR} \in \mathbb{C}^{N\times K}$ represents the channel matrix between the BS and rth STAR-BD-RIS, with K antennas in each sector.

$\mathbf{w}_{m_s}^{BU}$, $\mathbf{h}_{r,m_s}^{RU}$, and $\mathbf{V}_r^{BR}$ are defined as $\mathbf{w}_{m_s}^{BU} = \sqrt{\zeta_0(L)^{-\rho}}\left(\sqrt{\frac{\zeta_1}{1+\zeta_1}}\mathbf{w}_{m_s}^{BU,LOS} + \sqrt{\frac{1}{1+\zeta_1}}\mathbf{w}_{m_s}^{BU,NLOS}\right)$, $\mathbf{h}_{r,m_s}^{RU} = \sqrt{\zeta_0(L)^{-\rho}}\left(\sqrt{\frac{\zeta_2}{1+\zeta_2}}\mathbf{h}_{r,m_s}^{RU,LOS} + \sqrt{\frac{1}{1+\zeta_2}}\mathbf{h}_{r,m_s}^{RU,NLOS}\right)$, and $\mathbf{V}_r^{BR} = \sqrt{\zeta_0(L)^{-\rho}}\left(\sqrt{\frac{\zeta_3}{1+\zeta_3}}\mathbf{V}_r^{BR,LOS} + \sqrt{\frac{1}{1+\zeta_3}}\mathbf{V}_r^{BR,NLOS}\right)$. Here, $\mathbf{w}_{m_s}^{BU,NLOS}$, $\mathbf{h}_{r,m_s}^{RU,NLOS}$, and $\mathbf{V}_r^{BR,NLOS}$ represent the non-line-of-sight (NLOS) components, with elements following a zero-mean complex Gaussian distribution with unit variance. The LOS components, $\mathbf{w}_{m_s}^{BU,LOS}$, $\mathbf{h}_{r,m_s}^{RU,LOS}$, and $\mathbf{V}_r^{BR,LOS}$, are described by $\mathbf{w}_{m_s}^{BU,LOS} = \left[1, e^{-j\frac{2\pi}{\ell}d\cos(\phi^{AoD})}, \ldots, e^{-j\frac{2\pi}{\ell}(N-1)d\cos(\phi^{AoD})}\right]$, $\mathbf{h}_{r,m_s}^{RU,LOS} = \left[1, e^{-j\frac{2\pi}{\ell}d\cos(\phi^{AoD})}, \ldots, e^{-j\frac{2\pi}{\ell}(K-1)d\cos(\phi^{AoD})}\right]$, and $\mathbf{V}_r^{BR,LOS} = \left[1, e^{-j\frac{2\pi}{\ell}d\cos(\phi^{AoD})}, \ldots, e^{-j\frac{2\pi}{\ell}(N-1)d\cos(\phi^{AoD})}\right] \times \left[1, e^{-j\frac{2\pi}{\ell}d\cos(\phi^{AoA})}, \ldots, e^{-j\frac{2\pi}{\ell}(K-1)d\cos(\phi^{AoA})}\right]^H$. In these formulas, ρ and $\{\zeta_i\}_{i=1}^3$ denote the path loss exponent and the Rician factors, respectively. The parameter ζ_0 represents the path loss measured at a reference distance of 1 m. The variable L indicates the distance of the specific link. Additionally, ℓ, d, ϕ^{AoA}, and ϕ^{AoD} denote the wavelength, antenna spacing distance, angle-of-arrival (AoA), and angle-of-departure (AoD), respectively [7]. The two-sector BD-RIS is mathematically defined by two matrices, denoted as $\mathbf{\Psi}_r^s \in \mathbb{C}^{K\times K}$ for each s in the set $\mathcal{S}$. These matrices correspond to the two sectors and are sub-matrices of the larger scattering matrix $\mathbf{\Psi}_r \in \mathbb{C}^{2\times K\times 2\times K}$, which is associated with $2\times K$-port reconfigurable impedance network. Specifically, $\mathbf{\Psi}_r^s = [\mathbf{\Psi}_r]_{K+1:2\times K,1:K}$. They adhere to a combined unitary constraint, expressed as $\sum_s (\mathbf{\Psi}_r^s)^H \mathbf{\Psi}_r^s = \mathbf{I}_K \ \forall r \in \mathcal{R}$. For more comprehensive information on the modeling, design, and performance assessment of the multi-sector BD-RIS, refer to [3].

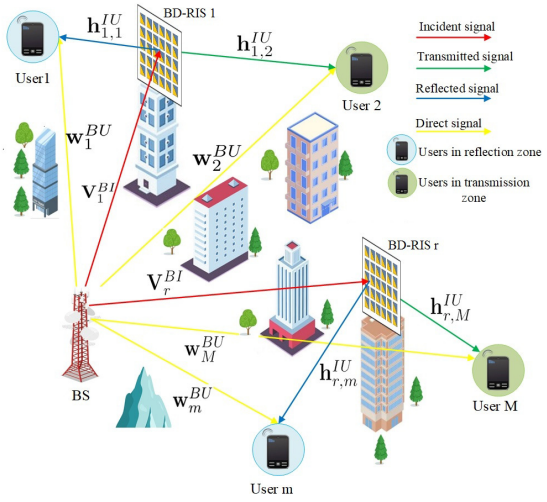


Fig. 1: The STAR-BD-RIS assisted system.

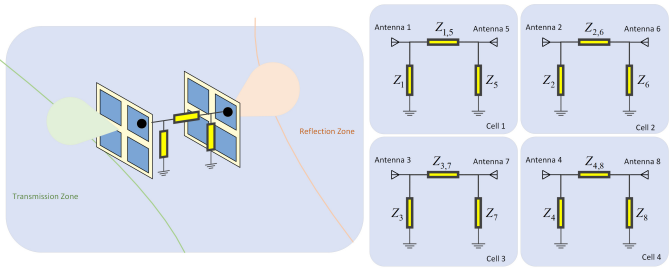


Fig. 2: Structure of a cell-wise single-connected BD-RIS featuring two sectors and four cells, where each cell is equipped with two antennas.

As discussed in [3], [13], reconfigurable impedance networks with different circuit topologies possess scattering matrices with various mathematical properties. In this study, we concentrate on a cell-wise single-connected (CW-SC) architecture of the multi-sector BD-RIS. To simplify understanding, Fig. 2 is provided for our case. In this architecture, the inner-cell antennas are interconnected through reconfigurable impedance components, while the inter-cell antennas remain unconnected. The CW-SC design of the multi-sector BD-RIS leads to diagonal matrices $\mathbf{\Psi}_r^s$ for each $s \in \mathcal{S}$ and $r \in \mathcal{R}$, which can be represented as $\mathbf{\Psi}_r^s = \text{diag}\left(\Psi_r^{s,1}, \ldots, \Psi_r^{s,K}\right)$ with $\Psi_r^{s,k} \in \mathbb{C}$ for all $s \in \mathcal{S}$, $k \in \mathcal{K}$, and $r \in \mathcal{R}$. Thus, the BD-RIS's constraint can be reformulated as $\sum_s \left|\Psi_r^{s,k}\right|^2 = 1, \forall k, r$. For BS antenna selection, we introduce $\mathbf{x_{m_s}} = [x_{m_s}^a] \in \mathbb{Z}^{N\times 1}$, where $x_{m_s}^a$ is a binary variable denoting the assignment of ath antenna to user m_s. That is, $x_{m_s}^a = 1$ if antenna a is assigned to user m_s, $x_{m_s}^a = 0$, otherwise. Let $\mathbf{u}_{m_s} \in \mathbb{C}^{N\times 1}$ denote the beamforming vector of the BS to m_sth user. The signal-to-interference-and-noise ratio (SINR) of user m_s is expressed as $\Gamma_{m_s} =$

$$\frac{\left|\left(\left(\sum_{r\in R}\mathbf{h}_{r,m_s}^{RU\ H}\mathbf{\Psi}_r^s\mathbf{V}_r^{BR^H}\right)+\mathbf{w}_{m_s}^{BU^H}\right)(\mathbf{u}_{m_s}\odot\mathbf{x}_{m_s})\right|^2}{I_{m_s}+\sigma_{m_s}^2},$$

where $\sigma_{m_s}^2$ represents the noise power at user m_s, and I_{m_s} is the interference caused by other users, which is obtained from $I_{m_s} = \sum_{s\in\mathcal{S}}\sum_{\substack{i_s\in\mathcal{M}\\ i_s\neq m_s}}\left|\left(\sum_{r\in R}\mathbf{h}_{r,i_s}^{RU\ H}\mathbf{\Psi}_r^s\mathbf{V}_r^{BR^H}+\mathbf{w}_{i_s}^{BU^H}\right)(\mathbf{u}_{i_s}\odot\mathbf{x}_{i_s})\right|^2$. According to Shannon's capacity formula, the total data rate of the system is given by $R_T = \sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}\log_2(1+\Gamma_{m_s})$. On the other hand, the total power consumption can be stated as $P_T = P_{T_1} + P_{T_2}$, where $P_{T_1} = P_{St}^{BS} + P_{\mathrm{BD-RIS}}$ in which P_{St}^{BS} denotes the static power of the BS, and $P_{\mathrm{BD-RIS}} = P_{St}^{\mathrm{BD-RIS}} + KP_{Dn}^{\mathrm{BD-RIS}}$ in which $P_{St}^{\mathrm{BD-RIS}}$ and $P_{Dn}^{\mathrm{BD-RIS}}$ denote the static power required for the basic operation of the BD-RIS circuits and dynamic power for each reflecting element k, respectively. Furthermore, $P_{T_2} = \sum_{i=1}^{N}(P_i^{Ant}\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}x_{m_s}^i) + \eta^{-1}\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}\|\mathbf{u}_{m_s}\odot\mathbf{x}_{m_s}\|^2$, where $\eta\in(0,1)$ is the efficiency of the transmit power amplifier, and P_i^{Ant} represents the dissipated power of the BS for each antenna i.

III. Problem Formulation

Next, we aim to maximize the energy efficiency (EE) in a multi-STAR-BD-RIS assisted system, in which EE is calculated as $\mathrm{EE} = R_T/P_T$. EE investigates the increase in the ratio of the achievable rate to power consumption. This is a suitable criterion for next-generation communication systems due to the massive number of devices and the increasing demand for higher data rates. Considering the beamforming vector at the BS, antenna selection vector, and phase shift matrix of STAR-BD-RISs as decision variables, the problem of EE maximization is formally stated as

$$\max_{\{\mathbf{u}_{m_s},\mathbf{x}_{m_s},\mathbf{\Psi}_r^s\}} \mathrm{EE} = \frac{R_T}{P_T} \tag{1a}$$

$$\text{s.t.} \quad \sum_{a=1}^{N} x_{m_s}^a \le N_{a_{max}},\ \forall s\in\mathcal{S},\ \forall m_s\in\mathcal{M}, \tag{1b}$$

$$\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}} x_{m_s}^a \le L_{Ant},\ \forall a\in\{1,\cdots,N\}, \tag{1c}$$

$$\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}\|\mathbf{u}_{m_s}\odot\mathbf{x}_{m_s}\|^2 \le P_{max}, \tag{1d}$$

$$x_{m_s}^a\in\{0,1\},\ \forall s\in\mathcal{S},\ \forall m_s\in\mathcal{M}, \forall a\in\{1,\cdots,N\}, \tag{1e}$$

$$\sum_{s\in\mathcal{S}}\left|\Psi_r^{s,k}\right|^2 = 1,\ \forall k\in\mathcal{K}, r\in\mathcal{R},. \tag{1f}$$

In (1), constraint (1b) shows that the number of antennas assigned to each user m_s is limited to $N_{a_{max}}$. Constraint (1c) indicates that the number of users associated with each antenna a should not exceed the maximum number L_{Ant}. Constraint (1d) implies that the total transmit power of the BS should be no larger than the maximum transmit power budget P_{max}. Also, $x_{m_s}^a$ is enforced to give binary values by constraint (1e). Finally, constraint (1f) is for STAR-BD-RIS's matrix.

The optimization problem (1) is a mixed-integer nonlinear programming (MINLP) problem, and there is no polynomial-time algorithm to obtain its optimal solution. To tackle this difficulty, we propose a Meta-SAC algorithm, which combines the benefits of SAC and meta-learning.

IV. Proposed Meta-SAC Algorithm

In this section, we explain the Meta-SAC algorithm in detail. The optimization problem (1) can be expressed as a Markov decision process (MDP), which is defined by state space $\mathcal{S}$, action space $\mathcal{A}$, and reward function R.

1) State space $\mathcal{S}$: the state space $\mathcal{S}$ contains all channels, interference of each user obtained from I_{m_s}, and total data rate R_T. Hence, state space $\mathcal{S}$ is given by $\mathcal{S} = \big\{\{\mathbf{w}_{m_s}^{BU}, I_{m_s}\}_{\forall s\in\mathcal{S},\forall m_s\in\mathcal{M}}, \{\mathbf{h}_{r,m_s}^{RU}\}_{\forall m_s\in\mathcal{M},\forall r\in\mathcal{R}}, \{\mathbf{V}_r^{BR}\}_{\forall r\in\mathcal{R}}, R_T\big\}$.

2) Action space $\mathcal{A}$: we consider all decision variables of the problem (1) as the action space $\mathcal{A}$ as $\mathcal{A} = \{\{\mathbf{u}_{m_s},\mathbf{x}_{m_s}\}_{\forall s\in\mathcal{S},\forall m_s\in\mathcal{M}}, \{\mathbf{\Psi}_r^s\}_{\forall r\in\mathcal{R},\forall s\in\mathcal{S}}\}$.

3) Reward function R: the reward function R is the objective function in (1) adding a penalty term to penalize actions violating the constraints of problem (1). Specifically, the reward function R is defined as

$$\begin{aligned} r =& \nu_1 EE - \nu_2\sum_{s=1}^{\mathcal{S}}\sum_{m_s\in\mathcal{M}}\left(\sum_{a=1}^{N}x_{m_s}^a - N_{a_{max}}\right) \\ &- \nu_3\sum_{a=1}^{N}\left(\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}x_{m_s}^a - L_{Ant}\right) \\ &- \nu_4\left(\sum_{s\in\mathcal{S}}\sum_{m_s\in\mathcal{M}}\|\mathbf{u}_{m_s}\odot\mathbf{x}_{m_s}\|^2 - P_{max}\right) \\ &+ \nu_5\sum_{r\in\mathcal{R}}\sum_{k\in\mathcal{K}}\Big(\sum_{s\in\in\mathcal{S}}\left|\Psi_r^{s,k}\right|^2 - 1\Big) \end{aligned} \tag{2}$$

$$R = \begin{cases} r, & \text{if constraint (1e) is satisfied,} \\ -|r|, & \text{otherwise.} \end{cases} \tag{3}$$

Creating an effective reward function is essential, as it needs to incorporate all constraints from the proposed optimization problem. A hybrid reward function, defined in (2), combines the objective function and constraints, while (3) ensures additional constraint satisfaction. Normalization with weighting factors is used due to varying units of the components, with the sum of these factors equaling one ($\sum_{i=1}^{5}\nu_i = 1 \quad \& \quad \nu_i \in [0,1], \forall i\in\{1,\ldots,5\}$) [13]. The weighting factors can be selected either manually or through a learning process. Typically, in manual selection, the objective of the optimization problem is assigned a higher weighting factor than the constraints, which are given uniform weighting factors. In contrast, with learnable selection, the weighting factors are derived from meta-DRL, taking into account $\nu_1 \ge \sum_{i=2}^{5}\nu_i$ and adding these factors to $\mathcal{A}$. In this paper, we use the second strategy.

Similar to SAC, Meta-SAC consists of actor and critic networks with parameters ϕ^{χ} and ϕ^{Q}, respectively. Additionally, Meta-SAC employs two target critic and target actor networks with parameters $\phi^{\chi'}$ and $\phi^{Q'}$. Let $\chi(s|\phi^{\chi})$ and $Q(s,\alpha|\phi^{Q})$ represent the actor and critic networks, respectively. The objective of the SAC agent is to find an optimal policy $\chi^{\star}$ that maximizes the trade-off between the expected cumulative reward and entropy. This is defined as:

$$\chi^{\star} = \arg\max_{\chi} \mathbb{E}_{(s_n,a_n)\sim \Pr}\Bigg[\sum_{n=0}^{\infty}\gamma^n Q(s_n,a_n|\phi^Q) - \lambda\sum_{n=0}^{\infty}\gamma^n \log\left(\chi(s_n|\phi^\chi)\right) \mid s_n = s_0, a_n = a_0\Bigg],$$

where $\gamma \in [0,1]$ is the discount factor. Pr represents the transition probabilities. $\chi(s_n|\phi^\chi)$ denotes the actor network, parameterized by ϕ^χ, and $Q(s_n,a_n|\phi^Q)$ denotes the critic network, parameterized by ϕ^Q.

In classic SAC, the actor network's parameters are updated to minimize the actor loss. Likewise, the objective of the critic network is to minimize the following loss function.

$$L_{\text{SAC}}^{\text{critic}} = \mathbb{E}_{s\sim p_\pi}\left[\alpha\log\left(\chi(s|\phi^\chi)\right) - Q(s,a|\phi^Q)\mid_{a=\chi(s|\phi^\chi)}\right]. \tag{4}$$

The parameters of the critic network are updated to minimize (4). According to [11], meta-learning can be formulated as a bi-level optimization problem as

$$\begin{aligned} &\kappa = \arg\min_{\kappa} \;\; L^{\text{meta}}\left(B_{\text{val}}\,;\phi^\chi\right) \\ \text{s.t.} \quad &\phi^{\chi^*} = \arg\min_{\phi^\chi}\left(J(B_{\text{trn}};\phi^\chi) + L_\kappa^{\text{meta-critic}}\left(B_{\text{trn}};\phi^\chi\right)\right), \end{aligned} \tag{5}$$

in which κ is the meta-knowledge obtained to minimize the loss function $L^{\text{meta}} = \tanh(L(B_{\text{val}};\phi^\chi_{New}) - L(B_{\text{val}}\,;\phi^\chi_{Old}))$. The values of ϕ^χ_{Old} and ϕ^χ_{New} are respectively updated as

$$\phi^\chi_{Old} \leftarrow \phi^\chi - lr_{\text{actor}}\nabla_{\phi^\chi} J(\phi^\chi), \tag{6}$$

$$\phi^\chi_{New} \leftarrow \phi^\chi_{Old} - lr_{\text{actor}}\nabla_{\phi^\chi} L_\kappa^{\text{meta-critic}}. \tag{7}$$

In contrast to classic SAC, the actor networks' parameters of Meta-SAC are updated to minimize the loss function $J(B_{\text{trn}};\phi^\chi) + L_\kappa^{\text{meta-critic}}\left(B_{\text{trn}};\phi^\chi\right)$, in which $L_\kappa^{\text{meta-critic}}\left(B_{\text{trn}};\phi^\chi\right) = \mathbb{E}\left[\kappa\left(\log\left(1+e^{\chi(s|\phi^\chi)}\right)\right)\right]$. The actor network's parameters are updated by

$$\phi^\chi \leftarrow \phi^\chi - lr_{\text{actor}}(\nabla_{\phi^\chi} J(\phi^\chi) + \nabla_{\phi^\chi} L_\kappa^{\text{meta-critic}}\left(\phi^\chi\right)). \tag{8}$$

Likewise, the parameter of the critic network is obtained from

$$\phi^Q \leftarrow \phi^Q - lr_{\text{critic}}\nabla_{\phi^Q} L(\phi^Q). \tag{9}$$

Meta-knowledge is updated as follows

$$\kappa \leftarrow \kappa - lr_{\text{meta}}\nabla_\kappa L^{meta}(\phi^\chi). \tag{10}$$

At each time step t of Meta-SAC, the agent observes the state s_t, then the agent selects action α_t. After executing the action α_t on the environment, the agent receives immediate reward r_t and transits to a new state s_{t+1}. The transition $(s_t,\alpha_t,r_t,s_{t+1})$ is then stored in the replay buffer. Next, for each gradient descent step n, a mini-batch of B_{trn} is randomly sampled from the replay buffer. Using B_{trn}, parameters ϕ^χ_{Old} and ϕ^χ_{New} are updated. Then, a mini-batch of B_{val} is sampled from the replay buffer. Using B_{val}, parameters of actor-network and meta-knowledge are updated based on (8) and (10), respectively. The proposed Meta-SAC method is summarized in Algorithm 1.

A. Time-Complexity Analysis of Meta-SAC

Inspired by [1], the overall time complexity of Meta-SAC can be expressed as $\mathcal{O}\left(\sum_{\ell=0}^{\mathcal{U}-1}\mu_{\text{actor},\ell}\mu_{\text{actor},\ell+1} + \sum_{k=0}^{\mathcal{P}-1}\mu_{\text{critic},k}\mu_{\text{critic},k+1} + \sum_{m=0}^{\mathcal{W}-1}\mu_{meta-c,m}\mu_{meta-c,m+1}\right)$, where $\mathcal{U}$, $\mathcal{P}$, and $\mathcal{W}$ denote the number of actor, critic, and meta-critic layers, respectively. Also, the number of neurons in each layer is represented by μ.

Algorithm 1: The proposed Meta-SAC algorithm

Input: Maximum number of episodes $E_{\max}$, maximum number of time steps $T_{\max}$, and maximum number of gradient descent steps $G_{\max}$.
Initialization:
Initialize replay buffer $\mathcal{B}$
Initialize the actor and critic networks, $\chi(s|\phi^\chi)$ and $Q(s,\alpha|\phi^Q)$
Initialize the target critic and the target actor networks $\chi'(s|\phi^{\chi'})$ and $Q'(s,\alpha|\phi^{Q\prime})$ with parameters of $\phi^{\chi'}$ and $\phi^{Q'}$
for each episode $e = 1, \cdots, E_{\max}$
  Get the initial state s_0
  for each time step $t = 1, \cdots, T_{\max}$
    Choose action a_t
    Perform action a_t on the environment and receive reward r_t
    The network transits from state s_t to s_{t+1}
    Store transition (s_t, a_t, s_{t+1}, r_t) in $\mathcal{B}$
Meta training:
  for each gradient descent step $n = 1, \cdots, G_{\max}$
    A mini-batch is randomly sampled from $\mathcal{B}$ as $\mathcal{B}_{\text{trn}}$.
    Update parameters of critic network ϕ^Q by (9).
    Obtain ϕ^χ_{Old} and ϕ^χ_{New} respectively using (6) and (7).
    Another mini-batch is randomly sampled from $\mathcal{B}$ as $\mathcal{B}_{\text{val}}$.
    Update parameters of actor network ϕ^χ by (8).
    Update meta-knowledge κ by (10).

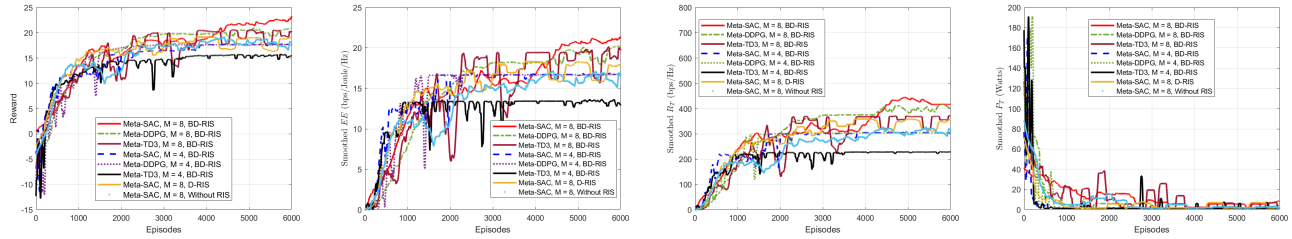
V. Simulation Results

In this section, we evaluate the performance of the proposed Meta-SAC algorithm. To this end, it is assumed that the BS is located at the coordinate origin, and two STAR-BD-RISs are placed at a random distance from it. In addition, all users are located in a circle according to uniform random distribution. All simulation parameters and hyper-parameters of Meta-SAC are given in Table I unless stated otherwise. Moreover, in the following figures, we compare the performance of Meta-SAC for multi-STAR-BD-RIS under different benchmarks.

Learning curves of our proposed system model with different meta benchmarks are provided in 3. In these simulations, to better illustrate the proposed meta strategy, deterministic channel coefficients are assumed [14]. The reward defined in section IV is plotted for different meta-algorithms and different numbers of users. As illustrated, all algorithms converged, and Meta-SAC with $M = 8$ achieved better reward compared to the other algorithms. The performance and convergence of Meta-SAC in terms of EE are shown in Fig. 3b. As can be observed from Fig.3b, Meta-SAC results in a smooth curve. Additionally, these figures present two benchmarks of the system model: one with a conventional RIS and one without RIS. The proposed STAR-BD-RIS demonstrates the

Table I: Simulation Parameters and Meta-SAC Hyper-parameters [3], [7], [11]

Parameter	Value	Parameter	Value
$M_{ts}+M_{rs}$	8	R	2
K	4	N	4
P_{max}	0.5 Watts	σ_m^2	-170 dBm/Hz
L_{Ant}	2	$N_{a_{max}}$	4
P_{St}	30 dBm	P_i^{Ant}	20 dBm
P_{St}^{BD-RIS}	100 mW	P_{Dn}^{BD-RIS}	0.33 mW
ζ_0	10^{-3}	ζ_1,ζ_2,ζ_3	1
ρ	2	ℓ	0.1
d	0.05	mini-batch size	32
Number of episodes	6000	Actor learning rate	0.001
Critic learning rate	0.001	Reward discount	0.99
SAC soft replacement	0.01	Replay memory capacity	1000000



(a) Reward vs episodes (b) EE vs episodes (c) R_T vs episodes (d) P_T vs episodes

Fig. 3: Learning curves of Meta-SAC in terms of (a) reward, (b) EE, (c) R_T, and (d) P_T with different meta-learning algorithms and number of users.

best performance compared to the others. Moreover, Fig. 3b depicts that Meta-SAC performs better when the number of users is increased. Additionally, Meta-SAC with multi-STAR-BD-RISs achieves a higher EE compared to other benchmarks. Also, the same results are obtained in Fig. 3c and Fig.3d. Fig. 4 depicts EE achieved by Meta-SAC versus the number of users. As can be seen, the performance of Meta-SAC grows when the number of users is increased. Additionally, the obtained EE by Meta-SAC in multi-STAR-BD-RIS systems is higher than that in single-STAR-BD-RIS. To demonstrate the effectiveness of the proposed solution, we compared it with the convex optimization approach. The results indicate that the proposed method achieves performance comparable to convex optimization while maintaining reasonable computational time and complexity. In convex benchmark, the one subproblem incorporates $\mathcal{C}_1 = \mathcal{SM} + \mathcal{N} + 2\mathcal{M} + \mathcal{MN} + 2$ convex constraints and involves $V_1 = 2\mathcal{M}N^2 + N\mathcal{M}$ variables. Using the SDP approach, the overall complexity can be expressed as $O_1 = \mathcal{O}\left(J_{\mathrm{SCA}_1} J_{\mathrm{D}} \sqrt{V_1} \log(\frac{1}{\epsilon})(\mathcal{C}_1 V_1^3 + \mathcal{C}_1^2 V_1^2 + \mathcal{C}_1^3)\right)$, where J_{D} and $J_{\mathrm{SCA_i}}, i \in \{1, 2\}$ represent the iterations needed for the Dinkelbach and the SCA algorithms, respectively. Regarding phase shift optimization, the complexity of solving this subproblem is given by $O_2 = \mathcal{O}(J_{\mathrm{SCA}_2})\sqrt{V_2}\log(\frac{1}{\epsilon})(\mathcal{C}_2 V_2^3 + \mathcal{C}_2^2 V_2^2 + \mathcal{C}_2^3)$, where $C_2 = 3\mathcal{M} + \mathcal{K}$ and $V_2 = 2\mathcal{M} + \mathcal{RSK}$. The ϵ represents the accuracy parameter for the SCA technique. The proposed solution outperforms convex optimization due to the high complexity of the convex benchmark caused by iterations. Fig. 5 compares EE obtained in multi-STAR-BD-RIS with other meta DRL algorithms when the number of reflecting elements increases from 4 to 36. As can be observed, EE is improved by increasing the number of elements. Furthermore, the proposed meta DRL achieves better performance in comparison to others.

VI. Conclusion

This paper studied a multi-STAR-BD-RIS assisted system where a multi-antenna BS serves several users with the help of multi-STAR-BD-RISs. We proposed a Meta-SAC algorithm to enable the BS to configure the beamforming matrix, STAR-BD-RIS matrices, and antenna selection to maximize EE by considering system model constraints. The proposed Meta-SAC is a combination of SAC and meta-learning. Simulation results illustrated that the multi-STAR-BD-RIS system achieves higher energy efficiency than single-STAR-BD-RIS systems. Simulation results confirmed that Meta-SAC outperformed other meta-DRL algorithms in terms of energy efficiency. Also, to show system model parameters' effects on EE, different scenarios were investigated.

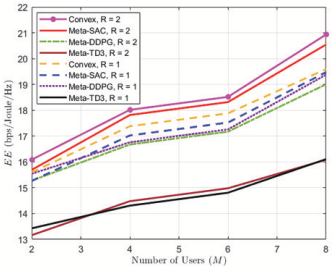


Fig. 4: EE vs. the number of users (M).

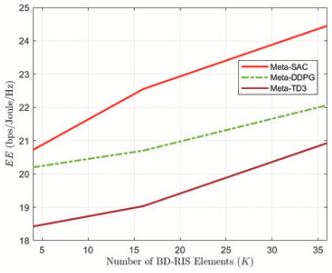


Fig. 5: EE vs. the number of STAR-BD-RISs' reflecting elements.

References

[1] A. Farhadi, M. Moomivand, S. K. Taskou, M. R. Mili, M. Rasti, and E. Hossain, "A Meta-DDPG Algorithm for Energy and Spectral Efficiency Optimization in STAR-RIS-Aided SWIPT," *IEEE Wireless Communications Letters*, 2024.

[2] S. Javadi, A. Farhadi, M. R. Mili, E. Jorswieck, and N. Al-Dhahir, "Meta-learning for resource allocation in uplink multi-active star-ris-aided noma system," *IEEE Wireless Communications Letters*, 2025.

[3] H. Li, S. Shen, and B. Clerckx, "Synergizing Beyond Diagonal Reconfigurable Intelligent Surface and Rate-Splitting Multiple Access," *IEEE Transactions on Wireless Communications*, 2024.

[4] M. Nerini, S. Shen, and B. Clerckx, "Closed-Form Global Optimization of Beyond Diagonal Reconfigurable Intelligent Surfaces," *IEEE Transactions on Wireless Communications*, vol. 23, no. 2, pp. 1037–1051, 2023.

[5] Z. Zhang, Z. Wang, Y. Liu, B. He, L. Lv, and J. Chen, "Security Enhancement for Coupled Phase-Shift STAR-RIS Networks," *IEEE Transactions on Vehicular Technology*, vol. Early Access, 2023.

[6] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum Rate Maximization for IRS-Assisted Uplink NOMA," *IEEE Communications Letters*, vol. 25, no. 1, pp. 234–238, January 2021.

[7] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, November 2019.

[8] S. Shen, B. Clerckx, and R. Murch, "Modeling and Architecture Design of Reconfigurable Intelligent Surfaces Using Scattering Parameter Network Analysis," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1229–1243, 2021.

[9] M. Nerini, S. Shen, and B. Clerckx, "Discrete-Value Group and Fully Connected Architectures for Beyond Diagonal Reconfigurable Intelligent Surfaces," *IEEE Transactions on Vehicular Technology*, 2023.

[10] Y. Yuan, G. Zheng, K. K. Wong, and K. B. Letaief, "Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 8964–8977, September 2021.

[11] W. Zhou *et al.*, "Online Meta-Critic Learning for Off-Policy Actor-Critic Methods," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 662–17 673.

[12] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed Multi-Agent Meta Learning for Trajectory Design in Wireless Drone Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3177–3192, October 2021.

[13] Z. Wang, Y. Wei, F. R. Yu, and Z. Han, "Utility Optimization for Resource Allocation in Edge Network Slicing Using DRL," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[14] Z. Ding, R. Schober, and H. V. Poor, "No-Pain No-Gain: DRL Assisted Optimization in Energy-Constrained CR-NOMA Networks," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5917–5932, 2021.