Deep Learning based Energy-Efficient Transmission Control for STAR-RIS aided Cell-Free Massive MIMO Networks

Chihyun Song[a], Donghyun Lee[a], Yunseong Lee[a], Wonjong Noh[b,*], Sungrae Cho[a,*]

[a]*School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea*
[b]*School of Software, Hallym University, Chuncheon 24252, South Korea*

Abstract

Recently, the simultaneous transmitting and reflecting (STAR) reconfigurable intelligent surface (RIS) has been gaining attention as a key enabler for sixth-generation networks, providing additional links with reduction in power consumption. This paper investigates the STAR-RIS's potential in a cell-free (CF) massive multiple-input multiple-output (mMIMO) network, where distributed APs serve user over the same time/frequency. We propose a deep deterministic policy gradient framework satisfying system-specific and per-user spectral efficiency constraints, exploiting a post-normalization and a penalized reward. From the simulations, it is revealed the proposed algorithm provides better energy performance than benchmarks, highlighting the benefits of STAR-RIS in the CF network.

Keywords: Cell-free massive multiple-input multiple-output, Deep reinforcement learning, Energy efficiency, Simultaneous transmitting and reflecting reconfigurable intelligent surface

1. Introduction

In the upcoming sixth-generation (6G) wireless networks, traditional cellular networks face significant challenges in providing uniform services due to inter-cell interference. Consequently, cell-free (CF) massive multiple-input multiple-output (mMIMO) is an emerging network technology expected to replace the cellular architecture, and it has been widely studied as a promising technology for 6G communication [1, 2, 3]. In CF mMIMO, geographically distributed APs jointly serve multiple users on the same time-frequency resource without cell boundaries, similar to a coordinated multi-point system [1]. Compared to cellular systems, providing service without cell boundaries results in better throughput by reducing inter-cell interference for users far from the APs. Nevertheless, high infrastructure costs and power consumption problems remain due to the requirement for large-scale AP deployment to achieve higher throughput.

On the other hand, reconfigurable intelligent surfaces (RIS) are promising techniques to improve the propagation environment by controlling the phase of incident signals [4]. RIS consists of low-cost passive programmable elements, each of which can perform effective passive beamforming [5]. This offers several advantages, such as low power consumption, low deployment cost, and scalability. Thus, the integration of CF mMIMO and RIS has been extensively researched [6, 7, 8, 9, 10].

However, most existing studies on RIS-aided CF mMIMO systems focus on reflecting-only RISs, which require the transmitter and receiver to be located on one side of the RIS. This half-space topology problem prevents taking full advantage of easy deployment and imposes additional topological constraints on the CF network. Recently, simultaneous transmitting and reflecting (STAR) RIS techniques have been developed to overcome the above limitations. This can realize omnidirectional signal coverage and improve the capacity and coverage of wireless networks by establishing cascaded links between transmitters and receivers, allowing the 6G network to meet its demands for high SE and energy-efficient system design [11].

As illustrated in [9], a joint optimization design involving the precoding at the APs and the phase control of the RIS elements is essential to fully utilize above benefits. However, conventional numerical optimization methods rely on complex algorithms or approximations,

*Corresponding author
Email addresses: chsong@uclab.re.kr (Chihyun Song), dhlee@uclab.re.kr (Donghyun Lee), yslee@uclab.re.kr (Yunseong Lee), wonjong.noh@hallym.ac.kr (Wonjong Noh), srcho@cau.ac.kr (Sungrae Cho)

which can lead to high computational costs and performance degradation. These limitations underline the need for more efficient and adaptive optimization techniques.

Recently, the integration of artificial intelligence in wireless network is constantly being studied, among which deep reinforcement learning (DRL) offers an efficient alternative to addressing the overhead of system optimization [12]. For example, DRL can effectively solve complex control problems through trial and error, reducing computational complexity even as the number of network elements grows [13]. Moreover, DRL can adapt to dynamic environments where the state of the network frequently changes [14].

1.1. Related work

The early study of the integration of RIS and CF mMIMO focuses on approximation/heuristic approaches [15, 16, 17, 18, 19, 20, 21, 22]. Qingqing *et al.* investigated the joint beamforming optimization for RIS-aided multi-user MISO system [15]. To handle the nonconvexity of the power minimization problem in the proposed system, a semidefinite relaxation technique was proposed to obtain an approximate solution. The performance of RIS-assisted CF mMIMO system over spatially correlated channels was studied in [16]. The RIS cascaded channel estimation method and optimizing RIS phase shifts control scheme are proposed to minimize the sum of channel estimation errors. In [23], a closed-form solution for the weighted sum rate maximization problem is derived in RIS-aided CF mMIMO system. To realize cooperative hybrid beamforming, the alternating direction method of multipliers and manifold optimization are proposed. Le *et al.* [18] proposed an inner-approximation framework-based joint precoding algorithm to maximize energy efficiency in the RIS-assisted CF mMIMO system under limited backhaul capacity. Zhang *et al.* [19] proposed a hybrid beamforming scheme that integrates digital beamforming in AP and analog beamforming in RIS to improve the energy efficiency of RIS-aided CF mMIMO systems. This study shows that the proposed RIS-assisted CF mMIMO system achieves better energy efficiency than traditional distributed antenna and CF mMIMO systems.

In [22], the cross-entropy-based probability learning method was proposed to optimize phase shifts and t/r ratio in the STAR-RIS-assisted multi-user system. The method incorporates joint parameterized sampling distribution and updating rules for the tilting parameter. Anastasios *et al.* [20] extended the above systems to a STAR-RIS-aided CF mMIMO system and proposed a closed-form expression for downlink achievable spectral efficiency using statistical CSI. Furthermore, Song *et al.* [21] investigated the WSR maximization problem in multiple STAR-RISs-assisted mmWave CF mMIMO systems. To jointly optimize active beamforming of APs and passive beamforming in RISs, a Lagrangian dual transformation and quadratic transformation were proposed to break the highly coupled problem into manageable subproblems.

However, the above-mentioned convex relaxation or heuristic algorithms often require huge computational resources to find the solution [24]. Recently, several DRL-based optimization frameworks have been developed [12, 13, 14, 25, 26, 27]. In [27], the energy consumption optimization problem for MEC offloading under task processing time constraints was derived. To reduce computation costs and adapt the time-varying channel, a game theory-based DRL framework is developed by combining DDQN and distributed LSTM. In [24], authors aimed to optimize the transmit power strategies for anti-jamming game. A novel approach integrating the Stackelberg game and DDPG is developed to solve the formulated problem while reducing the effect of incomplete information. In [28], the AP-user association method was proposed for energy efficiency maximization. To apply the DDPG for the large discrete action space, the action space approximation and the dimension-decreasing approach were proposed.

Huang *et al.* [14] proposed a DDPG-based sum-rate maximization algorithm for the RIS-assisted multi-user MISO system. The beamforming and phase shifts are jointly obtained while reducing the complexity and computation time. In [25], to maximize the achievable data rates while satisfying the QoS and latency constraints for STAR-RIS-assisted V2X communications, spectrum allocation, t/r ratio, phase shift, and power allocation were optimized by using DDQN. In [12], an optimal beamforming problem was formulated to maximize the sum-rate of the CF mMIMO system. The DDPG-based centralized beamforming and distributed deterministic policy gradient (D4PG) based beamforming method were proposed to handle the continuous action space. In [13], the dynamic clustering and beamforming for the CF mMIMO system were obtained via hybrid DRL approach, which utilizes the DDPG to find beamforming and DDQL to find dynamic clustering.

1.2. Contributions

Based on the above motivation, we propose the joint optimization framework based on DDPG for the STAR-RIS-aided CF mMIMO system. The contributions are summarized as follows:

- For a downlink CF network, we design the joint transmit precoding, phase shifts, and t/r ratio optimization method by formulating the energy-efficiency maximization problem under a per-user SE constraint and STAR-RIS system constraints.
- We transform the problem into a Markov decision processes (MDP) framework and apply a reinforcement learning approach to solve the non-convex optimization problem. Post-normalization and a penalized reward function are proposed to satisfy the constraints.
- Through system simulations, we demonstrate that the proposed algorithm converges stably, and it outperforms the conventional CF mMIMO in terms of energy efficiency as the number of network elements increases.

Notation: Vectors are given in lowercase bold (e.g., $\mathbf{a}$), and matrices are given in uppercase bold (e.g., $\mathbf{A}$). The superscripts $^{\mathrm{T}}$, and $^{\mathrm{H}}$ denote the transpose and Hermitian transpose, respectively. In addition, $\exp(v)$ represents a vector with the exponential function applied to each element of $\mathbf{v}$. The elementwise exponential of a matrix $\mathbf{A}$, denoted as $\exp(\mathbf{A})$, applies the exponential function to each element of $\mathbf{A}$. Further, $\|\mathbf{A}\|_F$ denotes the Frobenius norm of matrix $\mathbf{A}$, and $\mathbf{a} \circ \mathbf{b}$ represents the elementwise multiplication of vectors $\mathbf{a}$ and $\mathbf{b}$. Finally, $\mathbf{A} \circ \mathbf{B}$ represents the elementwise multiplication of $\mathbf{A}$ and $\mathbf{B}$.

2. System Model and Problem Formulation

We consider a CF mMIMO system supported by STAR-RIS, where M APs equipped with N antennas and K single antenna UEs are distributed in a coverage area, as illustrated in Fig. 1. With U RIS elements, STAR-RIS is at the center of the area, taking advantage of its full-space coverage. The coverage region can be divided into a front region (f) and a back region (b) depending on the angle of STAR-RIS. Furthermore, UE can be distinguished into front and back users depending on divided regions. That is, $\mathcal{K}_f$ UE is located in the front region and $\mathcal{K}_b$ UE is located in the back region where $\mathcal{K}_f + \mathcal{K}_b = K$.

The APs are assumed to serve UE jointly over the same time and frequency resource blocks. Furthermore, all APs are connected to a CPU via error-free backhaul links to enable channel information exchanges. Assuming the CPU is in the same position as STAR-RIS, it can be directly connected and controls the coordination of the phase shift and the transmission and reflection

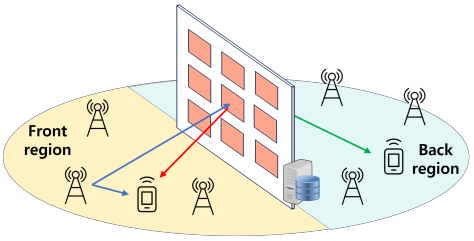

Figure 1. STAR-RIS-assisted Cell-Free mMIMO system.

(t/r) ratio without additional dedicated links. The t/r ratio and phase control can be individually performed on each element. When a signal incidents on the uth STAR-RIS element, it is divided into transmitted and reflected signals through the t/r ratio, β_u^f and β_u^b. The magnitude of the t/r ratio is constrained to satisfy the law of energy conservation as follows:

$$|\beta_u^f|^2 + |\beta_u^b|^2 = 1, \quad \forall u. \tag{1}$$

Furthermore, the divided signals are reconstructed using the phase shift control of the front region ϕ_u^f and back region ϕ_u^b to satisfy the unit-modulus constraint:

$$|\phi_u^f| = |\phi_u^b| = 1, \quad \forall u, \tag{2}$$

where ϕ_u^f, $\phi_u^b \in \mathbb{C}$. The combined matrix representation of the STAR-RIS t/r ratio and the phase shift for $\mathcal{K}_f$ and $\mathcal{K}_b$ is $\mathbf{\Phi}^f = \mathrm{diag}(\beta_1^f \phi_1^f, \ldots, \beta_U^f \phi_U^f)$ and $\mathbf{\Phi}^b = \mathrm{diag}(\beta_1^b \phi_1^b, \ldots, \beta_U^b \phi_U^b)$. For simplicity, the region indicator for UE k is defined as $\mathbf{g}_k \in \{f, b\}$. Thus, the STAR-RIS control parameter for UE k at region ($\mathbf{g}_k$) is represented as $\mathbf{\Phi}^{\mathbf{g}_k}$. Then, the STAR-RIS control parameter $\Phi^{\mathbf{g}}$ can be defined as $\Phi^{\mathbf{g}} = \{\Phi^{\mathbf{g}_1}, ..., \Phi^{\mathbf{g}_K}\}$.

2.1. Channel Model

For a given system model, $\mathbf{H}_m \in \mathbb{C}^{N \times U}$, $\mathbf{q}_k \in \mathbb{C}^{U \times 1}$, and $\mathbf{d}_{m,k} \in \mathbb{C}^{N \times 1}$ represent the AP-RIS channel matrix, RIS-user k channel vector, and AP-user k channel vector, respectively. The channel vector between AP m and UE k is represented as follows:

$$\mathbf{h}_{m,k} = \mathbf{H}_m \mathbf{\Phi}^{\mathbf{g}_k} \mathbf{q}_k + \mathbf{d}_{m,k}, \tag{3}$$

where

$$\mathbf{H}_m = \{\sqrt{\kappa_m}\mathbf{H}_{m,1}, ..., \sqrt{\kappa_m}\mathbf{H}_{m,U}\}, \tag{4}$$

$$\mathbf{q}_k = \sqrt{\kappa_k}\mathbf{h}_k, \tag{5}$$

$$\mathbf{d}_{m,k} = \sqrt{\kappa_{m,k}}\mathbf{h}_{m,k}, \tag{6}$$

where κ_m, κ_k, and $\kappa_{m,k}$ represent the large-scale fading of the AP-RIS, AP-UE k, and RIS-UE k links, respectively. Additionally, $\mathbf{H}_{m,u} \in \mathbb{C}^{N \times 1}$, $\mathbf{h}_{m,k} \in \mathbb{C}^{U \times 1}$, and $\mathbf{h}_k \in \mathbb{C}^{N \times 1}$ represent the small-scale fading vector corresponding components of the system.

2.2. Downlink Data Transmission

The transmit signal vector of AP m is given as follows:

$$\mathbf{x}_m = \sum_{k=1}^{K} \mathbf{w}_{m,k} s_{m,k}, \tag{7}$$

where $s_{m,k} \sim \mathcal{CN}(0,1)$ represents the data symbol transmitted from AP m to UE k, and $\mathbf{w}_{m,k} \in \mathbb{C}^{N\times 1}$ is the precoding vector for UE k from AP m. The transmitted signal power for each AP P_m^t should satisfy following constraint:

$$P_m^t = \sum_{k=1}^{K} |\mathbf{w}_{m,k}|^2 \le P^{\max}. \tag{8}$$

where $P^{\max}$ is the maximum transmit power. The received signal at the kth UE is given by

$$y_k = \sum_{m=1}^{M} \mathbf{h}_{m,k}^H \mathbf{x}_m + n_k \tag{9}$$

where $n_k \sim \mathcal{CN}(0,\sigma^2)$ represents the complex Gaussian noise at UE k. Given that, the signal-to-interference-and-noise ratio of UE k is defined as follows:

$$\gamma_k = \frac{|\sum_{m=1}^{M} \mathbf{h}_{m,k}^H \mathbf{w}_{m,k}|^2}{\sum_{k'\neq k} |\sum_{m=1}^{M} \mathbf{h}_{m,k}^H \mathbf{w}_{m,k'}|^2 + \sigma^2}. \tag{10}$$

Then, the downlink SE of the kth UE is denoted as:

$$R_k = \log_2(1+\gamma_k). \tag{11}$$

2.3. Problem Formulation

Based on the system model, we propose an energy-efficiency maximization problem to optimize the precoding, phase shift, and t/r ratio. Then, the energy-efficiency maximization problem can be expressed as follows:

$$\textbf{(P1)} \quad \max_{\boldsymbol{w},\boldsymbol{\Phi},\boldsymbol{\beta}} \quad EE = \frac{\mathrm{BW}\sum_{k=1}^{K} R_k}{P_{\text{total}}} \tag{12a}$$

$$\text{s.t.} \quad R_k \ge R^{\text{th}}, \quad \forall k \tag{12b}$$

$$\beta_u^t \ge 0, \beta_u^r \ge 0, \quad \forall u \tag{12c}$$

$$(1), (2), (8), \tag{12d}$$

where $\boldsymbol{w} = \{\mathbf{w}_{m,k} | m \in \mathcal{M},\ k \in \mathcal{K}\}$, $\boldsymbol{\Phi} = \{\phi^f,\ \phi^b\}$, $\boldsymbol{\beta} = \{\beta^f,\ \beta^b\}$, BW denotes the system bandwidth, and R^{th} indicates the guaranteed per-user SE. The total power consumption of the system P_{total} is modeled as in [29],[30]:

$$P_{\text{total}} = \frac{1}{\alpha_m}\sum_{m}^{M} P_m^t + \sum_{m=1}^{M} P_{\text{bh},m} + M\cdot P_{\text{ap}} + K\cdot P_{\text{ue}} + U\cdot P_{\text{ris}}, \tag{13}$$

where α_m denotes the power amplifier efficiency, and P_{ap} and P_{ue} represent the circuit static power of the AP and UE, respectively. Moreover, P_{ris} indicates the power consumed by each RIS element. The backhaul power consumption is denoted as:

$$P_{\text{bh},m} = P_{0,m} + \mathrm{BW}\cdot\sum_{k=1}^{K} R_k \cdot P_{\text{bt},m}, \tag{14}$$

where $P_{0,m}$ denotes the fixed power consumption of each backhaul, and $P_{\text{bt},m}$ represents the traffic-dependent backhaul power consumption. In addition, (12b) represents the per-user SE constraints. Moreover, (12c) guarantees that the energy of the divided signals has a positive range.

The precoding vectors $\boldsymbol{w}$, the passive beamforming in the STAR-IRS $\boldsymbol{\Phi}$, and the t/r ratio $\boldsymbol{\beta}$ should be jointly optimized to maximize the total energy efficiency of the system. However, (**P1**) is a non-convex problem with highly coupled variables and constraints. Although optimization methods using convex relaxation or heuristic approaches can be applied, they cannot guarantee a global optimum solution, and the computation cost can be extensive. Therefore, we propose a post-normalization layer and penalized DDPG framework for designing the energy-efficient transmission control.

3. Proposed Approach

We first transform the optimization problem into a task for an RL agent to determine the transmit precoding, phase shift, and t/r ratio for the AP. The agent, which uses the computational capacity of the CPU, observes the environment to determine the appropriate actions and receives a reward at each time step t. The CPU receives environment information through the backhaul, and the agent's decisions are sent to each AP. These learning scenarios are modeled as a Markov decision process (MDP). Based on MDP, the proposed DRL agent collects state, action, reward, and transition pairs to learn the optimal policy.

3.1. Markov Decision Process

The state space, action space, and reward function are defined as follows:

1) *State space*: The state of the system observed by the agent is defined as the CSI of the AP-RIS-user and AP-user paths, which is transmitted to the CPU via the backhaul links.

$$s[t] = [\{\mathbf{H}_m[t] | m\in\mathcal{M}\}, \{q_k[t] | k\in\mathcal{K}\}, \{d_{m,k}[t] | m\in\mathcal{M}, k\in\mathcal{K}\}] \tag{15}$$

2) *Action space*: According to the given state, the agent determines the phase shift, t/r ratio, and AP precoding vector. The action space combines all possible continuous values of these variables.

$$a[t] = \{\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{\beta}\}. \tag{16}$$

3) *Reward function*: During the training session, the agent aims to determine the optimal action to maximize the reward and the energy efficiency is used to assess it. Accordingly, the following definition is used for the instantaneous reward function $r[t]$

$$r[t] = \mathrm{BW}\frac{\sum_{k=1}^{K} R_k[t]}{P_{\text{total}}}. \tag{17}$$

3.2. Post-Normalization Layer and Penalized Reward

The neural network outputs may not satisfy the constraints described in equations (1), (2), (12b), and (12c) because of the added noise and the limited output range of the activation functions. Therefore, we propose a post-normalization layer and penalized reward to satisfy the constraints. The network output is divided into the following components to address the constraints on the optimization variables.

- **Precoding phase**: For each AP m, the precoding direction is represented by the block matrix $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_M\}$. Each element of the matrix $\mathbf{V}_m \in \mathbb{R}^{N\times K}$ indicates the signal phase in the complex space between the corresponding antenna and UE.

- **Precoding amplitude**: For each AP m, the amplitude of precoding is $\mathbf{A} = \{\mathbf{A}_1, ..., \mathbf{A}_M\}$, where each element of $\mathbf{A}_m \in \mathbb{R}^{N\times K}$ indicates the magnitude of the signal between the corresponding antenna and UE.

- **AP power budget**: The power budget for each AP m represents the proportion of maximum power each AP uses and is indicated by the vector $\eta = \{\eta_1, ..., \eta_m\}$.

- **Passive beamforming phase**: The STAR-RIS phase shift is divided into two vectors: $\Psi^f = \{\psi_1^f, ..., \psi_U^f\}$ and $\Psi^b = \{\psi_1^b, ..., \psi_U^b\}$.

- **t/r proportion**: The t/r proportions are described by the vectors $\mathbf{p^f} = \{p_1^f, ..., p_U^f\}$ and $\mathbf{p^b} = \{p_1^b, ..., p_U^b\}$.

Due to the nature of the activation function, these outputs have values between 0 and 1. Thus, the precoding vector $\mathbf{w}_{m,k}$ reformulated as follows:

$$\mathbf{w}_{m,k} = \sqrt{P^{\max}\eta_m}\frac{\mathbf{a}_{m,k} \circ \exp(j2\pi\mathbf{v}_{m,k})}{\|\mathbf{A}_m \circ \exp(j2\pi\mathbf{V}_m)\|_F}, \tag{18}$$

where $\mathbf{a}_{m,k}$ and $\mathbf{v}_{m,k}$ are the kth column of $\mathbf{A}_m$ and $\mathbf{V}_m$, respectively. The above normalized precoding vector is easily proved to satisfy the constraint (8). The STAR-RIS phase shifts, θ_u^f and θ_u^b are as follows:

$$\phi_u^f = \exp\left(j2\pi\psi_u^f\right), \quad \phi_u^b = \exp\left(j2\pi\psi_u^b\right) \tag{19}$$

which satisfy the unit-modulus constraint (2). The t/r ratio of each RIS element β_u^t and β_u^r are rewritten as follows:

$$\beta_u^f = \frac{p_u^f}{\sqrt{{p_u^f}^2 + {p_u^b}^2}}, \quad \beta_u^b = \frac{p_u^b}{\sqrt{{p_u^f}^2 + {p_u^b}^2}}, \tag{20}$$

satisfying constraints (1) and (12c).

Although system-specific constraints can be addressed post-normalization, resolving the per-user SE (12b) via these approaches is challenging. Therefore, we propose a penalized reward, expressed as follows:

$$r^{\text{penalty}}[t] = \mathrm{BW}\frac{\sum_{k=1}^{K} R_k[t]}{P_{\text{total}}[t]} + \lambda\sum_{k=1}^{K}\left(\min(R_k[t] - R^{\text{th}}, 0)\right) \tag{21}$$

where λ is the amplitude of the penalty term that is a positive constant. Setting an appropriate value for λ decreases the reward if the constraint (12b) is not satisfied during training, ensuring maximization of energy efficiency while meeting constraints.

3.3. Proposed DDPG Framework

Due to the continuous action space in the MDP framework, we propose an DDPG algorithm that can manage this continuous space [31]. Two types of neural networks exist: an actor network $\mu(s|\theta^\mu)$ and a critic network $Q(s, a|\theta^Q)$, where θ^μ and θ^Q represent the parameters of the network of the actor and critic networks in DDPG. For off-policy learning, the agent maintains two sets of actor-critic networks: the target and behavior networks. The target networks are denoted by Q' and μ', while the behavior networks are written as Q and μ. The behavior network selects actions based on the current policy to explore the environment. Noise is added to the output of the behavior network to improve exploration as follows:

$$a[t] = \mu(s[t]|\theta^\mu) + \mathcal{N}[t], \tag{22}$$

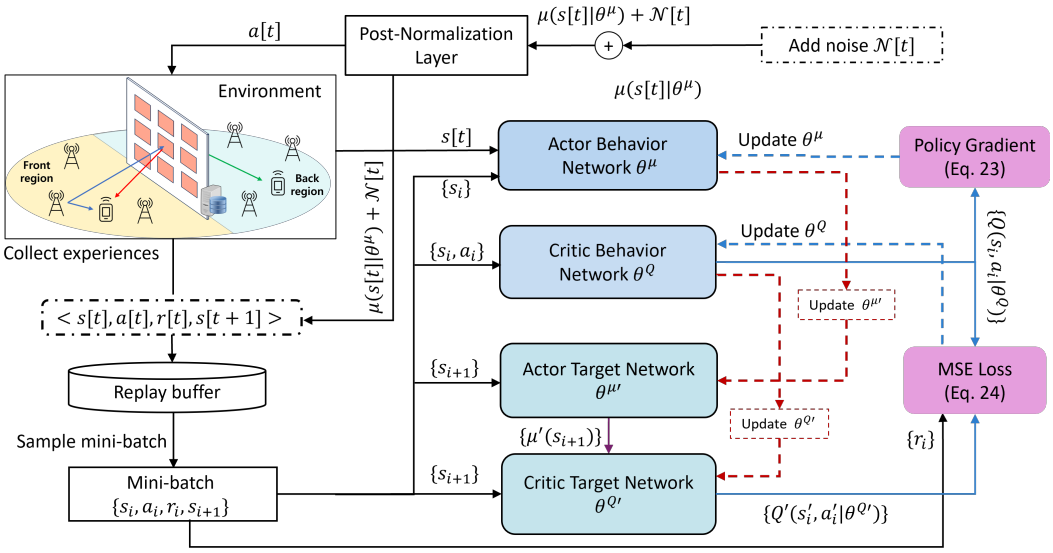

Figure 2. The proposed post-normalization and penalized DRL framework.

Algorithm 1 Proposed DDPG algorithm

Initialize $Q(s,a|\theta^Q)$, $Q'(s,a|\theta^{Q'})$, $\mu(s|\theta^\mu)$, and $\mu'(s|\theta^{\mu'})$
for episode $1,...,E$ **do**
  Initialize noise process $\mathcal{N}$ for exploration
  Observe system state $s[t]$
  for $t = 1,...,S$ **do**
    Receive network outputs: $\mu(s[t]|\theta^\mu)+\mathcal{N}[t]$
    Calculate action a[t] through normalization layer according to (18), (19), and (20)
    Execute $a[t]$, observe $r^{\text{penalty}}[t]$, and state transition $s[t+1]$
    Store transition ($s[t]$,$a[t]$,$r^{\text{penalty}}[t]$,$s[t+1]$) in the replay buffer
    Select a random batch of B samples (s_i, a_i, r_i, s_{i+1}) from the replay buffer
    Update each target network parameter according to (24), (25), (26), (27), and (28)
  end for
end for
Output: actor network μ^*

where $\mathcal{N}[t]$ represents Ornstein-Uhlenbeck noise. Additionally, the target network is a copy of the behavior network used to stabilize the training.

The fundamental objective of the proposed DDPG framework is to learn an optimal policy that maximizes cumulative reward. The optimal policy π^* satisfies the following Bellman optimality equation for all states:

$$\mu^*(s|\theta^\mu) = \arg\max_a \left(Q(s,a,|\theta^Q)\right) \tag{23}$$

In order to achieve it, the proposed DDPG framework updates the behavior network through the policy gradient:

$$\nabla_{\theta_\mu} J = \frac{1}{B}\sum_{i=1}^{B}\left(\nabla_a Q(s,a|\theta^Q)\big|_{s=s_i,a=\mu(s_i|\theta^\mu)}\ \nabla_{\theta_\mu}\mu\left(s_i|\theta^\mu\right)\right), \tag{24}$$

where B denotes the size of the sample batch, and s_i denotes the state in batch i. At the same time, the behavior critic function $Q(s,a)$ is updated by minimizing the following MSE loss between the behavior and target critic values:

$$L = \frac{1}{B}\sum_{i=1}^{B}\left(Q\left(s_i,a_i|\theta^Q\right)-y_i\right)^2, \tag{25}$$

where y_i represents the target critic value, and a_i indicates the action in batch i. The target critic value is defined as follows:

$$y_i = r_i + \gamma Q'\left(s_i', \mu'\left(s_i'|\theta^{\mu'}\right)|\theta^{Q'}\right), \tag{26}$$

where γ denotes the discount factor, r_i and s_i' represent the reward and next state in batch i, respectively.

On the other hand, the target network gradually changes by slowly tracking the behavior network using

Table 1
Hyperparameter Values

Parameters	Value
Learning rate	0.001
Discount factor	0.99
Soft update rate λ	0.001
Steps	500
Episodes	10000
Batch size	128
Critic - Hidden layer	256, Sigmoid
Actor - Hidden layer	256, Sigmoid

the soft update mechanism, making learning more stable. The two target networks can be updated via *Polyak averaging*, as follows:

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{27}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}, \tag{28}$$

where $\tau \in (0, 1)$ indicates the soft update coefficient. Each network hyperparameters are described in Table 1, and the learning process is summarized in Algorithm 1 and Fig.2.

4. Simulation Results

The performance was evaluated in various scenarios by comparing energy efficiency to verify the performance of the proposed algorithm and the STAR-RIS-aided CF system. In the simulation of the proposed system, the M APs and K UEs are uniformly distributed in a region of $D \times D$ m^2. The system environment is modeled from previous work [2]. The large-scale fading coefficients $\kappa_{m,k}, \kappa_m$, and κ_k represent the path loss and shadow fading effects of the respective elements, defined as follows:

$$\kappa_{m,k} = \mathrm{PL}_{m,k} 10^{\frac{\sigma_{\mathrm{sh}} z_{m,k}}{10}}, \tag{29}$$

$$\kappa_{m} = \mathrm{PL}_{m} 10^{\frac{\sigma_{\mathrm{sh}} z_{m}}{10}}, \tag{30}$$

$$\kappa_{k} = \mathrm{PL}_{k} 10^{\frac{\sigma_{\mathrm{sh}} z_{k}}{10}}, \tag{31}$$

where PL indicates the path loss, and the exponential component indicates the log-normal shadow fading with a standard deviation σ_{sh}, and $z \sim \mathcal{N}(0, 1)$. The path loss model is given by following three slope fading model:

$$\mathrm{PL}_{m,k} = \begin{cases} -L - 35\log_{10}(d_{m,k}), & d_{m,k} > d_1 \\ -L - 15\log_{10}(d_1) - 20\log_{10}(d_{m,k}), & d_0 < d_{m,k} \le d_1 \\ -L - 15\log_{10}(d_1) - 20\log_{10}(d_0), & d_{m,k} \le d_0 \end{cases} \tag{32}$$

The small-scale fading is modeled using the Rayleigh distribution. Noise power is determined by multiplying the bandwidth, Boltzmann constant, noise temperature, and noise figure. Table 2 outlines the system parameters employed in the simulations.

Table 2
System parameters

System parameter	Value
Power amplifier efficiency, α_m	0.40
Fixed power consumption of the backhaul, $P_{0,m}$	0.825 W
Traffic-dependent backhaul power, $P_{\mathrm{bt},m}$	0.25 W(Gbits/s)
Fixed power consumption, $P_{\mathrm{ap}}, P_{\mathrm{ue}}, P_{\mathrm{ris}}$	0.2, 0.01, 0.01 mW
Bandwidth, BW	20 MHz
Carrier frequency	1.9 GHz
Noise figure	9 dB
Std of shadow fading, σ_{sh}	8 dB
Antenna height	15 m
User antenna height	1.65 m
RIS height	30 m
D, d_1, d_0	200, 50, 10 m

The STAR-RIS-aided CF system was compared with the following benchmarks.

- **S-CF/MRE** : This scheme employs MR precoding, random phase shifts, and an equal t/r ratio in the STAR-RIS-aided CF mMIMO system.
- **C-CF** : This scheme employs the proposed DDPG algorithm in the conventional CF mMIMO system.
- **C-CF/MRE**: This scheme employs MR precoding, random phase shifts, and an equal t/r ratio in the conventional CF mMIMO system.

First, we evaluated the average reward of the proposed DDPG with various learning rates, which is depicted in Fig. 3. The shaded area represents the 95% confidence interval. Each parameter converges within 5,000 episodes. A learning rate of 1e-3 performs the best, whereas the convergence speed is similar between rates. A lower learning rate tends to make the model more susceptible to becoming trapped in the local optima. A high level of variance exists in the cases of 1e-4 and 5e-4.

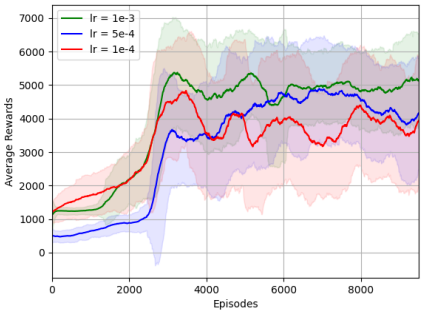

Figure 3. Average rewards versus episodes under various learning rate settings in the STAR-RIS-aided CF mMIMO system, where $M = 10$, $K = 4$, $N = 4$, $U = 16$, and $R^{\text{th}} = 0.5$.

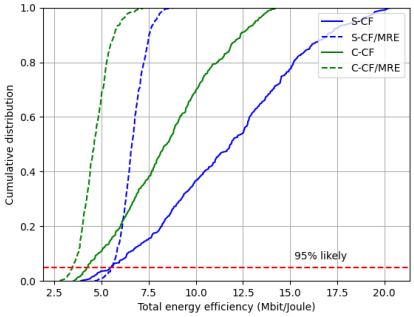

Figure 4. Cumulative distribution of the average total energy efficiency for the STAR-RIS CF mMIMO system for the proposed algorithm and benchmark schemes, where $M = 10$, $K = 4$, $N = 4$, $U = 16$, and $R^{\text{th}} = 0.5$.

Fig. 4 presents the cumulative distribution of the energy efficiency for the DDPG and benchmark schemes, indicating that the STAR-RIS-aided CF mMIMO significantly outperformed the conventional CF mMIMO in the median value. Furthermore, the proposed algorithm performs 79.3% and 83.3% better than the MRE method in the cases of the S-CF and C-CF systems in terms of median value, respectively. However, for the 95% likely, the S-CF and S-CF/MRE have nearly identical values. This result suggests that the proposed algorithm is relatively unstable in poor channel conditions.

Fig. 5 evaluates the influence of the number of AP antennas. The energy efficiency increases with the number of antennas in all cases except for C-CF/MRE because more precise precoding becomes possible as the number of antennas at the AP increases. However, for C-CF/MRE, the precoding gain is initially observed to increase but saturates due to the fixed power consumption of the antennas. Furthermore, the proposed algorithm

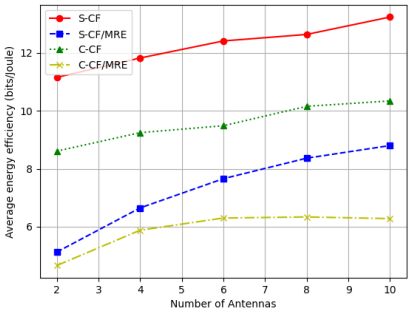

Figure 5. Average total energy efficiency versus the number of antennas, N, where $M = 10$, $K = 4$, $U = 16$, and $R^{\text{th}} = 0.5$.

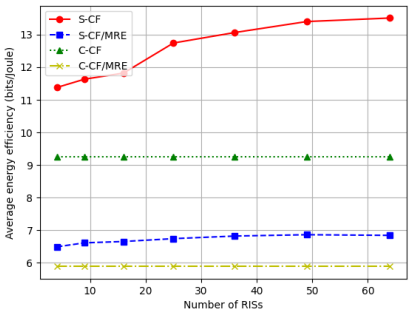

Figure 6. Average total energy efficiency versus the number of RIS elements, U, where $M = 10$, $K = 4$, $N = 4$, and $R^{\text{th}} = 0.5$.

demonstrates an improvement of 67% in the STAR-RIS CF and 62% in the CF compared to the benchmarks.

Next, Fig. 6 represents the influence of the RIS elements. When STAR-RIS was deployed, the proposed algorithm achieved an average increase of 46.1% compared to S-CF, demonstrating the potential of STAR-RIS. Moreover, for S-CF/MRE, the energy efficiency did not increase significantly and even decreased as the number of elements increased. This indicates that when joint precoding optimization is not performed, the static power consumption per element outweighs the gains from RIS.

5. Conclusion

This work demonstrates how STAR-RIS can improve energy efficiency in CF mMIMO networks under diverse system configurations. We propose a novel DDPG approach that addresses the energy-efficiency maximization problem. The proposed post-normalization layer and penalized reward ensure compliance with

system-specific and per-user SE constraints. The simulation results demonstrate that the proposed DDPG-based algorithm learns efficiently from the environment and provides better energy efficiency than the conventional benchmark scheme in CF mMIMO networks.

6. Acknowledgments

This work was supported in part by the IITP (Institute of Information & Coummunications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) (IITP-2025-RS-2022-00156353, 50%) (IITP-2025-RS-2023-00258639, 50%) grants funded by the Korea government (Ministry of Science and ICT) and in part by the Chung-Ang University Research Scholarship Grants in 2023.

References

[1] Özlem Tugfe Demir, E. Björnson, L. Sanguinetti, Foundations of user-centric cell-free massive mimo, Foundations and Trends in Signal Processing 14 (3-4) (2021) 162–472.

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, T. L. Marzetta, Cell-free massive mimo versus small cells, IEEE Transactions on Wireless Communications 16 (3) (2017) 1834–1850.

[3] A. Papazafeiropoulos, H. Q. Ngo, P. Kourtessis, S. Chatzinotas, Star-ris assisted cell-free massive mimo system under spatially-correlated channels, IEEE Transactions on Vehicular Technology 73 (3) (2024) 3932–3948.

[4] Q. Wu, R. Zhang, Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network, IEEE Communications Magazine 58 (1) (2020) 106–112.

[5] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. D. Renzo, C.-B. Chae, L. Hanzo, Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results, IEEE Access 8 (2020) 45913–45923.

[6] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, B. Ottersten, Reconfigurable intelligent surface-assisted cell-free massive mimo systems over spatially-correlated channels, IEEE Transactions on Wireless Communications 21 (7) (2022) 5106–5128.

[7] E. Shi, J. Zhang, R. He, H. Jiao, Z. Wang, B. Ai, D. W. K. Ng, Spatially correlated reconfigurable intelligent surfaces-aided cell-free massive mimo systems, IEEE Transactions on Vehicular Technology 71 (8) (2022) 9073–9077.

[8] E. Shi, J. Zhang, S. Chen, J. Zheng, Y. Zhang, D. W. Kwan Ng, B. Ai, Wireless energy transfer in ris-aided cell-free massive mimo systems: Opportunities and challenges, IEEE Communications Magazine 60 (3) (2022) 26–32.

[9] Z. Zhang, L. Dai, A joint precoding framework for wideband reconfigurable intelligent surface-aided cell-free network, IEEE Transactions on Signal Processing 69 (2021) 4085–4101.

[10] J. Yao, J. Xu, W. Xu, D. W. K. Ng, C. Yuen, X. You, Robust beamforming design for ris-aided cell-free systems with csi uncertainties and capacity-limited backhaul, IEEE Transactions on Communications 71 (8) (2023) 4636–4649.

[11] X. Mu, Y. Liu, L. Guo, J. Lin, R. Schober, Simultaneously transmitting and reflecting (star) ris aided wireless communications, IEEE Transactions on Wireless Communications 21 (5) (2022) 3083–3098.

[12] F. Fredj, Y. Al-Eryani, S. Maghsudi, M. Akrout, E. Hossain, Distributed beamforming techniques for cell-free wireless networks using deep reinforcement learning, IEEE Transactions on Cognitive Communications and Networking 8 (2) (2022) 1186–1201.

[13] Y. Al-Eryani, M. Akrout, E. Hossain, Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design, IEEE Journal on Selected Areas in Communications 39 (4) (2021) 1028–1042.

[14] C. Huang, R. Mo, C. Yuen, Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning, IEEE Journal on Selected Areas in Communications 38 (8) (2020) 1839–1850.

[15] Q. Wu, R. Zhang, Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming, IEEE Transactions on Wireless Communications 18 (11) (2019) 5394–5409.

[16] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, B. Ottersten, Reconfigurable intelligent surface-assisted cell-free massive mimo systems over spatially-correlated channels, IEEE Transactions on Wireless Communications 21 (7) (2022) 5106–5128.

[17] N. T. Nguyen, V.-D. Nguyen, H. V. Nguyen, H. Q. Ngo, S. Chatzinotas, M. Juntti, Spectral efficiency analysis of hybrid relay-reflecting intelligent surface-assisted cell-free massive mimo systems, IEEE Transactions on Wireless Communications 22 (5) (2023) 3397–3416.

[18] Q. N. Le, V.-D. Nguyen, O. A. Dobre, R. Zhao, Energy efficiency maximization in ris-aided cell-free network with limited backhaul, IEEE Communications Letters 25 (6) (2021) 1974–1978.

[19] Y. Zhang, B. Di, H. Zhang, J. Lin, C. Xu, D. Zhang, Y. Li, L. Song, Beyond cell-free mimo: Energy efficient reconfigurable intelligent surface aided cell-free mimo communications, IEEE Transactions on Cognitive Communications and Networking 7 (2) (2021) 412–426.

[20] A. Papazafeiropoulos, H. Q. Ngo, P. Kourtessis, S. Chatzinotas, Star-ris assisted cell-free massive mimo system under spatially-correlated channels, IEEE Transactions on Vehicular Technology 73 (3) (2024) 3932–3948.

[21] Y. Song, S. Xu, R. Xu, B. Ai, Weighted sum-rate maximization for multi-star-ris-assisted mmwave cell-free networks, IEEE Transactions on Vehicular Technology 73 (4) (2024) 5304–5320.

[22] J.-C. Chen, Designing star-ris-assisted wireless systems with coupled and discrete phase shifts: A computationally efficient algorithm, IEEE Transactions on Vehicular Technology 73 (7) (2024) 10772–10777.

[23] X. Ma, D. Zhang, M. Xiao, C. Huang, Z. Chen, Cooperative beamforming for ris-aided cell-free massive mimo networks, IEEE Transactions on Wireless Communications 22 (11) (2023) 7243–7258.

[24] M. Chen, A. Liu, N. N. Xiong, H. Song, V. C. M. Leung, Sgpl: An intelligent game-based secure collaborative communication scheme for metaverse over 5g and beyond networks, IEEE Journal on Selected Areas in Communications 42 (3) (2024) 767–782.

[25] P. S. Aung, L. X. Nguyen, Y. K. Tun, Z. Han, C. S. Hong, Deep reinforcement learning-based joint spectrum allocation and configuration design for star-ris-assisted v2x communications, IEEE Internet of Things Journal 11 (7) (2024) 11298–11311.

[26] Y. Cui, T. Lv, W. Ni, A. Jamalipour, Digital twin-aided learning

for managing reconfigurable intelligent surface-assisted, uplink, user-centric cell-free systems, IEEE Journal on Selected Areas in Communications 41 (10) (2023) 3175–3190.

[27] M. Chen, W. Liu, T. Wang, S. Zhang, A. Liu, A game-based deep reinforcement learning approach for energy-efficient computation in mec systems, Knowledge-Based Systems 235 (2022) 107660.

[28] N. Ghiasi, S. Mashhadi, S. Farahmand, S. M. Razavizadeh, I. Lee, Energy efficient ap selection for cell-free massive mimo systems: Deep reinforcement learning approach, IEEE Transactions on Green Communications and Networking 7 (1) (2023) 29–41.

[29] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, E. G. Larsson, On the total energy efficiency of cell-free massive mimo, IEEE Transactions on Green Communications and Networking 2 (1) (2018) 25–39.

[30] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, C. Yuen, Reconfigurable intelligent surfaces for energy efficiency in wireless communication, IEEE Transactions on Wireless Communications 18 (8) (2019) 4157–4170.

[31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971 (2015).