DRL-Based Transmission Design for Distributed STAR-RIS-aided Communications

Riya Yadav[†], Sravani Kurma[†], Sandeep Kumar Singh[¶], Chih-Peng Li[†], and Mayur Parate[*]
[†]Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[¶]Department of ECE, Motilal Nehru National Institute of Technology Allahabad, India
[*]Department of ECE, Indian Institute of Information Technology, Nagpur, India
E-mail: {riyay533,sravani.phd.nsysu.21}@gmail.com, sksingh@mnnit.ac.in, cpli@mail.nsysu.edu.tw, mparate@iiitn.ac.in

***Abstract*—In this work, we investigate the performance of distributed simultaneous transmitting and reflecting (STAR) reconfigurable intelligent surface (RIS)-assisted downlink multi-user communication system consisting of a base station (BS) serving multiple users (at both front and back ends). In order to provide an energy-efficient transmission design for the considered distributed-STAR-RIS ON/OFF system, we formulate a EE maximisation problem while ensuring a minimum rate quality of service (QoS) at each user under the total transmit power constraint at the BS. One of the critical challenges is to obtain perfect Channel State Information (CSI). Due to the non-convexity of the formulated problem, we propose a deep reinforcement learning (DRL)-based framework that provides optimum solution by incorporating the deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3) algorithms. The efficacy and complexity of the proposed algorithms are validated graphically obtained through the extensive simulations. We provide a demonstration of how the performance of the considered system is affected with the variation in key system parameters, such as the maximum transmit power available at the BS, number of RIS, reflecting elements in each RIS and antenna at the BS. The simulation results demonstrate the effectiveness of the proposed algorithms, with the STAR-RIS approach outperforming the conventional reflecting-only (CR)-RIS.**

***Index Terms*—Deep reinforcement learning (DRL), energy efficiency (EE), simultaneous transmitting and reflecting reconfigurable intelligent surface (STAR-RIS), TD3.**

I. Introduction

The reconfigurable intelligent surface (RIS), an emerging technology in 6G wireless communication, enables an enhanced connection between the transmitter and receiver by providing a significant beamforming gain [1]. It has enhanced various aspects of wireless networks such as multiple-input multiple-output (MIMO) networks, non-orthogonal multiple access (NOMA) technology [2], and unmanned aerial vehicle (UAV) networks [3]. However, the current implementation of the conventional reflecting-only (CR) RIS has a limitation of redirecting incoming signals only towards the area directly in front of it [1], [3]–[5]. Consequently, in recent times, researchers have proposed simultaneous transmitting and reflecting (STAR)-RIS to address the aforementioned limitation [6], [7]. Unlike the reflecting-only RIS, the STAR-RIS is capable of simultaneous signal transmission towards the back end user and reflection towards the front end users, enabling full-space coverage [8], [9]. By manipulating the incident signal into transmission and reflection regions, the STAR-RIS can achieve 360° coverage [10].

Integrating the STAR-RIS into existing wireless networks offers significant advantages, including improved system performance and the ability to dynamically adjust the multiple access channel, ultimately enhancing user satisfaction [9]. Extensive analysis and investigation of the sum rate in STAR-RIS-assisted wireless networks have been conducted by various authors [9], [11]. However, prior studies have not explored the use of distributed STAR-RIS, which is the primary focus of this work.

Additionally, the recent advancements in artificial intelligence, particularly deep reinforcement learning (DRL)-based algorithms, have gained significant attention and recognition as an effective framework to efficiently solve the optimization problems in the field of wireless communications [3], [5], [12]–[14]. DRL-based technologies are well-suited for wireless applications as they enable wireless devices to learn through interaction with the environment [11]. Note that unlike supervised learning which often requires a large number of training labels, DRL-based methods do not rely on training labels and have the availability of online learning and sample generation, making them more efficient.

Note that despite extensive analysis of the impact of STAR-RIS in [6]–[11], the exploration of the DRL-based resource allocation in distributed-STAR-RIS-aided wireless communications remains untapped. Motivated by the aforementioned discussion, this paper aims at providing an unified framework to understand the impact of DRL-based algorithms towards efficient resource allocation for distributed-STAR-RIS ON/OFF assisted energy-efficient systems. In particular, we investigate the performance of a STAR-RIS-assisted downlink multi-user MISO communication system consisting of a base station (BS) serving multiple front end and back end users using multiple STAR-RISs. In order to provide an efficient resource allocation, we formulate an EE maximisation problem while ensuring a minimum rate quality of service (QoS) at each

This work was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 112-2221-E-110-029-MY3, in part by the Sixth Generation Communication and Sensing Research Center funded by the Higher Education SPROUT Project, the Ministry of Education of Taiwan, and in part by MNNIT Allahabad Prayagraj under Cumulative Professional Development Allowance (CPDA).

user under the total transmit power constraint at BS. In order to address the non-convexity of the formulated problem, we propose a DRL-based framework that provides optimum solution by incorporating the deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3) algorithms. The efficacy and complexity of the proposed algorithms are validated graphically obtained through the extensive simulations. We provide a demonstration of how the performance of the considered system is affected with the variation in key system parameters, such as the maximum transmit power available at the BS, reflecting elements in each RIS and antenna at the BS. The advantages of adopting TD3 over DDPG are highlighted. Additionally, we also demonstrate the advantages of using STAR-RIS instead of CR-RIS in terms of extended coverage and enhanced average EE of the system.

Notation: The paper utilises diverse notations for scalars, vectors, and matrices, using lowercase, bold lowercase, and bold uppercase letters, respectively. It uses $\mathcal{CN}(0,\sigma^2)$ for the complex Gaussian distribution with mean 0 and variance σ^2, and diag$(x_1,...,x_N)$ for the diagonal matrix with diagonal components $x_1,...,x_N$. Real and imaginary parts of a complex number x are denoted as $R(x)$ and $I(x)$, respectively. The $[x]_m$ denotes the m^{th} element of the vector x while $[X]_{nm}$ denotes $(n,m)^{th}$ element of the matrix X. $[X]^T$ and $[X]^H$ denote the transpose and conjugate transpose of vector x, respectively.

II. System Model And Preliminaries

We consider a RIS-assisted downlink multi-user MISO communication system consisting of a BS outfitted with a set of M antennas serving K single antenna users on the front and J on the back side (dead zone) of V STAR-RISs (each having N_v reflecting elements).

Firstly, for the ease of understanding, we denote, $\mathcal{V} = \{1,\ldots,V\}$, $\mathcal{N}_v = \{1,\ldots,N_v\}$, $\mathcal{K} = \{1,\ldots,K\}$, and $\mathcal{J} = \{1,\ldots,J\}$. Note that the adjustments to the amplitude and phase of both the reflected and transmitted signal components are determined to characterize the properties of the STAR-RIS element [11], [15]. The elements of the STAR-RIS function in a dual mode, simultaneously operating in both reflection and transmission modes, utilising the Energy Splitting (ES) protocol [9] to maximise flexibility. The reflection and transmission characteristics of the n^{th} element of v^{th} STAR-RIS are defined by $\sqrt{a_{vn}^R}\mathrm{e}^{j\theta_{vn}^R}$ and $\sqrt{a_{vn}^T}\mathrm{e}^{j\theta_{vn}^T}$ where $a_{vn}^R, a_{vn}^T \in [0,1]$ is amplitude and $\theta_{vn}^R, \theta_{vn}^T \in [0,2\pi)$,$v \in \mathcal{V}$ and $n \in \mathcal{N}_v$ denotes phase of the reflected and transmitted signal. While the phase modifications can be set independently, the amplitude coefficients are limited by the principle of energy conservation, $a_{vn}^R + a_{vn}^T = 1$. The signals reflected and transmitted are modelled using the phase shift matrix of the RIS $v \in \mathcal{V}$ and given by

$$\boldsymbol{\Theta}_v^R = \mathrm{diag}\left(\sqrt{a_{v1}^R}\mathrm{e}^{j\theta_{v1}^R},\ldots,\sqrt{a_{vN_v}^R}\mathrm{e}^{j\theta_{vN_v}^R}\right), \qquad (1)$$

$$\boldsymbol{\Theta}_v^T = \mathrm{diag}\left(\sqrt{a_{v1}^T}\mathrm{e}^{j\theta_{v1}^T},\ldots,\sqrt{a_{vN_v}^T}\mathrm{e}^{j\theta_{vN_v}^T}\right). \qquad (2)$$

The superimposed symbol to be transmitted by the BS is given by

$$\boldsymbol{s} = \sum_{i\in\Omega} \boldsymbol{w}_i s_i, \qquad (3)$$

where $\Omega = \{R_1.., R_k....R_K, T_1, ..T_j....T_J\}$, s_i is unit-power information symbol [16] and $\boldsymbol{w}_i \in \mathbb{C}^{M\times 1}$ is the precoding vector of the i^{th} user. The BS transmits $\boldsymbol{s}$ to all users in reflection and transmission spaces simultaneously. As discussed earlier, due to the high cumulative power consumption, enabling all the RISs simultaneously may lead to inefficient energy usages by the considered system. Therefore, similar to [16], we adopt a RIS selection setup, wherein we attempt to deactivate one RIS at a time. For implementing the same, a binary variable $z_v \in \{0,1\}$ is introduced, where 1 means RIS v is ON while 0 means RIS v is OFF. Furthermore, for the ease of analysis, it is assumed that there is perfect channel state information (CSI) available at the respective node [17]. The signal received by users in reflecting and transmitting space can be expressed as follows:

$$y_k^R = \left(\boldsymbol{h}_k^H + \sum_{v=1}^{V} z_v \boldsymbol{g}_{kv}^H \boldsymbol{\Theta}_v^R \boldsymbol{H}_v\right)\boldsymbol{s} + n_k, k \in \mathcal{K}, \qquad (4)$$

$$y_j^T = \left(\sum_{v=1}^{V} z_v \boldsymbol{f}_{jv}^H \boldsymbol{\Theta}_v^T \boldsymbol{H}_v\right)\boldsymbol{s} + n_j, j \in \mathcal{J}, \qquad (5)$$

where $\boldsymbol{H}_v \in \mathbb{C}^{N_v\times M}$ is the channel from BS to v^{th} RIS, $\boldsymbol{h}_k \in \mathbb{C}^M$ and $\boldsymbol{g}_{kv} \in \mathbb{C}^{N_v}$ denote the channel responses from the BS and v^{th} RIS, respectively, to the user k in the reflection zone, and $\boldsymbol{f}_{jv} \in \mathbb{C}^{N_v}$ depicts the channel from RIS v to user j in the transmission zone. Here, $n_k, n_j \sim \mathcal{CN}\left(0,\sigma^2\right)$ represent the additive white Gaussian noise (AWGN). The respective received signal-to-interference-plus-noise ratio (SINR) at users are given by

$$\rho_k^R = \frac{\left|\left(\boldsymbol{h}_k^H + \sum\limits_{v=1}^{V} z_v \boldsymbol{g}_{kv}^H \boldsymbol{\Theta}_v^R \boldsymbol{H}_v\right)\boldsymbol{w}_{r_k}\right|^2}{\sum\limits_{i=1,i\neq k}^{K}\left|\left(\boldsymbol{h}_k^H + \sum\limits_{v=1}^{V} z_v \boldsymbol{g}_{kv}^H \boldsymbol{\Theta}_v^R \boldsymbol{H}_v\right)\boldsymbol{w}_{r_i}\right|^2 + \sigma^2},$$

$$\rho_j^T = \frac{\left|\left(\sum\limits_{v=1}^{V} z_v \boldsymbol{f}_{jv}^H \boldsymbol{\Theta}_v^T \boldsymbol{H}_v\right)\boldsymbol{w}_{t_j}\right|^2}{\sum\limits_{i=1,i\neq j}^{J}\left|\left(\sum\limits_{v=1}^{V} z_v \boldsymbol{f}_{jv}^H \boldsymbol{\Theta}_v^T \boldsymbol{H}_v\right)\boldsymbol{w}_{t_i}\right|^2 + \sigma^2}.$$

The achievable sum rate of the system is given by

$$R_{tot} = B\sum_{k=1}^{K}\log_2\left(1+\rho_k^R\right) + B\sum_{j=1}^{J}\log_2\left(1+\rho_j^T\right),$$

where B denotes the total bandwidth of the channel. Further, the total energy consumed by the considered system comprises the dynamic transmission power at the BS, the static circuit power consumed by the BS, users and all RISs. Consequently, the total power consumed by the system is given by

$$P_{\text{tot}} = \sum_{k=1}^{K} \iota \boldsymbol{w}_{R_k}^H \boldsymbol{w}_{R_k} + \sum_{j=1}^{J} \gamma \boldsymbol{w}_{T_j}^H \boldsymbol{w}_{T_j} + P_{\text{B}} \\ + \sum_{k=1}^{K} P_{R_k} + \sum_{j=1}^{J} P_{T_j} + \sum_{v=1}^{V} z_v N_v P_{\text{r}}, \qquad (6)$$

where P_{B} is the circuitry power of the BS and $\iota = \nu_0^{-1}$ account for the efficiency of the transmit power amplifier adopted at users and P_{R_k}, P_{T_j} refers to the circuit power consumption of reflected and transmitted users, and P_r represents the power consumption of each reflecting element in the RIS.

III. Optimisation Problem Formulation

Given the system under consideration, the main objective of this study is to maximise the EE of the system by jointly optimising the precoding vector at the BS, RIS activation vector $\boldsymbol{z}$ and the phase shift matrix at the respective RISs, while ensuring a minimum rate QOS at each user under the total transmit power constraint at BS. Thus, we formulate a EE maximisation problem as follows

$$\max_{\boldsymbol{\theta^R}, \boldsymbol{\theta^T}, \boldsymbol{w}, \boldsymbol{z}} \quad R^{tot}/P_{tot}, \qquad (7a)$$

$$\text{s.t. } B\log_2\left(1+\rho_k^R\right) \geq R_{k,\min}, \forall k \in \mathcal{K}, \qquad (7b)$$

$$B\log_2\left(1+\rho_j^T\right) \geq R_{j,\min} \forall j \in \mathcal{J}, \qquad (7c)$$

$$\boldsymbol{w}^H \boldsymbol{w} \leq P_{\max}, \qquad (7d)$$

$$\theta_{vn}^R, \theta_{vn}^T \in [0, 2\pi], \forall v \in \mathcal{V}, n \in \mathcal{N}_v, \qquad (7e)$$

$$0 \leq a_{vn}^T, a_{vn}^R \leq 1, a_{vn}^T + a_{vn}^R = 1, \qquad (7f)$$

$$z_v \in \{0,1\}, \quad \forall v \in \mathcal{V}, \qquad (7g)$$

where R^{tot} represents the sum rate of the system, $\boldsymbol{\theta^R} = \left[\left(\theta_{11}^R, \ldots, \theta_{1N_1}^R, \ldots, \theta_{VN_V}^R\right)\right]^T$, $\boldsymbol{\theta^T} = \left[\left(\theta_{11}^T, \ldots, \theta_{1N_1}^T, \ldots, \theta_{VN_V}^T\right)\right]^T$, $\boldsymbol{w} = [\boldsymbol{w}_{R_1}; \ldots; \boldsymbol{w}_{R_K}, \boldsymbol{w}_{T_1}; \ldots; \boldsymbol{w}_{T_J}]$, $\boldsymbol{z} = [z_1, \ldots, z_V]^T$, $R_{k,\min}$ and $R_{j,\min}$ are the minimum data rate requirement of transmitted and reflected users, respectively, and $P_{\max}$ represents the maximum available transmit power at the BS. Note that the constraint (7b) ensures that the minimum data rate requirement is met by the respective user. (7d) represents the total power constraint, describing that the transmission power is limited at the BS, whereas (7e) and (7f) ensure that amplitudes and phase shift coefficients at the STAR-RIS will be adjusted within reasonable ranges. Moreover, (7g) enforces the ON/OFF condition for all RIS.

Evidently, (7a) is a non-convex optimization problem. The traditional mathematical method is not suitable for finding the optimal solution due to the problem's large number of optimization variables. In contrast, DRL-based algorithms can easily be employed to efficiently solve the optimization problem without the need for complex mathematical derivations.

IV. DRL Based Proposed Solution

Reinforcement Learning (RL) agents acquire optimal actions by interacting with the environment and learning from experience via the Markov Decision Process (MDP) framework to handle state transitions and rewards. It is characterized by its components: $(\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{P}_{s \to s^{t+1}}})$. Here, $\mathcal{S}$ represents the set of environment states, $\mathcal{A}$ denotes the set of possible actions, $\mathcal{R}$ is the reward function, and $\mathcal{P}_s \to s^{t+1}$ represents the transition probabilities from state s to s^{t+1} for all s and s^{t+1} in $\mathcal{S}$.

The Markov property states that the probability of the next state (future state) is independent of the past, given the present state. The key elements of RL learning are state, action, and immediate reward which are briefly discussed below:

State space: The state represents a series of observations that describe the environment, with the state at time step t denoted as $s_t \in S$. $\mathcal{S} \in \{\mathcal{S}_1, \mathcal{S}_2\}$ is defined for both users as follows:

$$\mathcal{S}_1 = \left\{h_1, H_1, g_{11}, \ldots, h_k, H_v, g_{kv} \ldots, h_K, H_V, g_{KV}\right\}, \quad (8)$$

$$\mathcal{S}_2 = \left\{H_1, f_{11}, \ldots, H_v, f_{jv} \ldots, H_V, f_{JV}\right\}. \quad (9)$$

Action space: During training, the phase shift matrix $\boldsymbol{\theta}$, precoding vector $\boldsymbol{w}$, and RIS activation vector $\boldsymbol{z}$ create an action vector which is given by

$$a_t = \left[\boldsymbol{\theta}^R, \boldsymbol{\theta}^T, \boldsymbol{w}, \boldsymbol{z}\right]. \qquad (10)$$

The environment transitions from the current state s_t to the subsequent s' when the agent takes action a_t at time t, according to a policy π.

Reward: Reward used to assess the performance of actions in a specific state. Note that to optimise EE, a reward function r_t is adopted.

In brief, DRL requires the agent to interact with the environment to optimise a reward function. The agent's goal is to choose the most appropriate action a_t based on its observations and the current state. RL algorithms can be applied to environments that are either fully observable, with the agent having direct access to all environmental information, or partially observable. The primary objective of the agent is to discover an optimal policy, denoted as π^*, which maximises the cumulative and discounted reward function over successive time-steps. In other words, the agent aims to find a mapping from the set of states $\mathcal{S}$ to the set of actions $\mathcal{A}$, expressed as π^*: $\mathcal{S} \to \mathcal{A}$. Let us define the action-value function as $Q_\pi(s,a)$. This function is maximised by the optimal policy denoted as π^*. The function can be expressed as

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \alpha^t r_{(t+u+1)} | S_t = s, A_t = a\right],$$

where α is a variable that is confined within the range $0 \leq \alpha \leq 1$ and serves as the discount factor, accounting for the uncertainty associated with future rewards. Here, r_i denotes the reward procured at the i-th step.

$$\mathcal{L}\left(\zeta^{\text{crit}},\mathcal{R}\right)= \mathbb{E}\Bigg[\Big(Q\left(s,a;\zeta^{\text{crit}}\right)- \overbrace{\Big(r+\alpha\max_{a'}Q\left(s',a';\zeta^{\text{crit}}\right)\Big)}^{\text{target value}}\Big)^2\Bigg]. \tag{11}$$

DDPG is an efficient model-free and off-policy actor-critic method for continuous action-spaces. It utilises four deep neural networks (DNNs): two as actor-critic networks and the other two as target networks. The actor network, driven by ζ^{act}, directly provides actions based on states using a deterministic policy denoted as $\gamma(s,\zeta^{act})$, here, γ signifies a deterministic policy with single-valued outputs, rather than probability distributions. The critic network, using ζ^{crit} weights, evaluates the action-value function with the policy network's action and the current state. Target networks generate target action-values to minimise the mean-squared Bellman error (MSBE).

The experience replay memory $\mathcal{R}$ stores tuples $\left(s,a,r,s^{t+1}\right)$ of states, actions, rewards, and next states from previous steps. From equation (10), the next optimal action a^{t+1} is obtained using the target actor network with parameter set $\zeta^{targ-act}$, where $a^{t+1}=\gamma(s^{t+1},\zeta^{targ-act})$. The corresponding action-value $Q(s^{t+1},a^{t+1},\zeta^{targ-crit})$ is then evaluated using the target critic network with weights $\zeta^{targ-crit}$). Typically, the weights of the two networks are updated by copying them over from the main network using Polyak averaging, which is

$$\boldsymbol{\zeta}^{\text{targ-act}}\leftarrow\alpha\boldsymbol{\zeta}^{\text{act}}+(1-\tau)\boldsymbol{\zeta}^{\text{targ-act}}, \tag{12}$$

$$\zeta^{\text{targ-crit}}\leftarrow\tau\zeta^{\text{crit}}+(1-\tau)\zeta^{\text{targ-crit}}, \tag{13}$$

where the soft update hyperparameter, denoted as $\tau << 1$, regulates the updating procedure. In the learning process, both the true action-value function, denoted as $Q_\pi(s,a)$, and the approximated function, represented by $Q(s,a;\zeta^{\text{crit}})$, adhere to the subsequent inequality.

$$\begin{aligned}&\mathbb{E}\left[Q\left(s,a=\gamma\left(s;\boldsymbol{\zeta}^{act}\right);\boldsymbol{\zeta}^{crit}\right)\right]\\&\quad\geq\mathbb{E}\left[Q_\pi\left(s,a=\gamma\left(s;\boldsymbol{\zeta}^{act}\right)\right)\right].\end{aligned} \tag{14}$$

In general, the DDPG algorithm, which relies on typical Q-learning methods, tends to overestimate Q-values during training. This overestimation propagates to subsequent states and episodes, leading to a degradation of the policy network's performance, and consequently, poor policy updates. To address these challenges, TD3 (Twin Delayed Deep Deterministic) introduces the following improvements (as shown in Algorithm 1): In TD3, two deep neural networks (DNNs) are utilised to estimate the action-value function in the Bellman equation, and the minimum value of the Q-values is used. Additionally, the target and policy networks are updated less frequently compared to the critic networks. A regularization technique is used to prevent the policy network from selecting actions that can cause large peaks or Q-value failures. Instead, actions are determined by introducing a slight amount of clipped random noise to the chosen action, as indicated by the algorithm.

$$\begin{aligned}a^{t+1}=&\operatorname{clip}(\gamma\left(s^{t+1};\zeta^{\text{targ}-\text{act}}\right)\\&+\operatorname{clip}(\chi,-b,+b),d_{\text{Low}},d_{\text{High}}).\end{aligned} \tag{15}$$

In this context, $\chi\sim\mathcal{N}\left(0,\tilde{\sigma}^2\right)$ represents the normal Gaussian noise added to the selected action. The lower and upper limits for the action at the RIS elements are defined as $d_{Low}=-\pi$ and $d_{High}=+\pi$, respectively. These limits ensure that the action remains feasible even with the added noise. Additionally, a constant "c" is utilised to truncate the added noise during the first stage, preserving the proximity of the target action to the original action. Algorithm 1 outlines the proposed TD3-based algorithm, which involves taking tuples out of the experience replay buffer R and then giving the target and current networks the extracted tuple as input.

V. Numerical Case Studies

This section showcases and discusses the outcomes of the simulations to validate the efficacy of the proposed algorithm. Due to LOS dominating nature, the small-scale fading is modelled as a Rician distribution for the channels from BS to RIS and RIS to users follows the rician fading channel. For example channel from BS to RIS is given by

$$\mathbf{H_v}=F\left(\sqrt{\frac{K_v}{1+K_v}}\mathbf{H_v}^{LoS}+\sqrt{\frac{1}{1+K_v}}\mathbf{H_v}^{NLoS}\right),$$

where $\mathbf{H_v}^{LoS}$ represents the deterministic LoS component, $\mathbf{H_v}^{NLoS}$ represents the random non-LoS (NLoS) component which is modelled as Rayleigh fading. Furthermore, F denotes

Algorithm 1 Proposed TD3 based Algorithm

```
Input: All the channels $\boldsymbol{h}_k$, $\boldsymbol{H}_v$, $\boldsymbol{g}_{kv}$ and $\boldsymbol{f}_{jv}$;
Initialize: Actor and critic networks $\zeta^{act}$, $\zeta_1^{critic}$, $\zeta_2^{critic}$ for training in the TD3 network.
Initialize: Replay buffer $\mathcal{R}$
Output: The current EE as the result of optimal action $a=\left[\boldsymbol{\theta}^R,\boldsymbol{\theta}^T,\boldsymbol{w},\boldsymbol{z}\right]$;
for each episodes do
    Collect and preprocess $\boldsymbol{g}_k$, $\boldsymbol{H}_v$, $\boldsymbol{h}_{kv}$ and $\boldsymbol{f}_{jv}$ to obtain the initial state $s$;
    for each step t do
        Obtain action $a$ from the actor-network;
        Observe instant reward $r$ and new state $s^{t+1}$ at given action $a$;
        Store the experience $\left(s,a,s^{t+1},r\right)$ in the replay memory;
        Sample mini-batches $M$ of experiences from replay buffer $\mathcal{R}$;
        Compute target actions from (15);
        Update the critic network by performing gradient descent ;
        Update the policy network;
        Every $t$ steps, update the target network by equation (12),(14);
    end for
end for
```

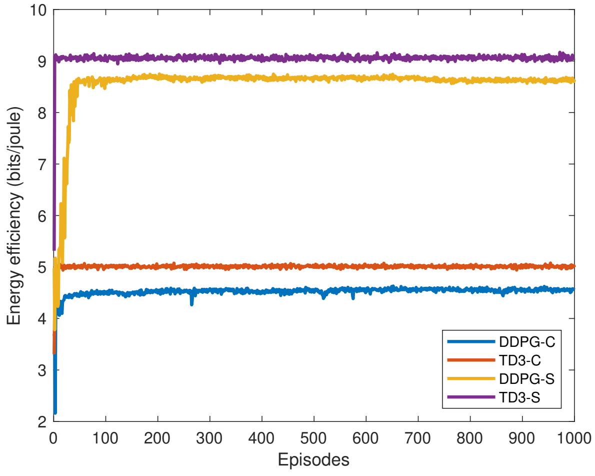

Fig. 1. Convergence of STAR-RIS

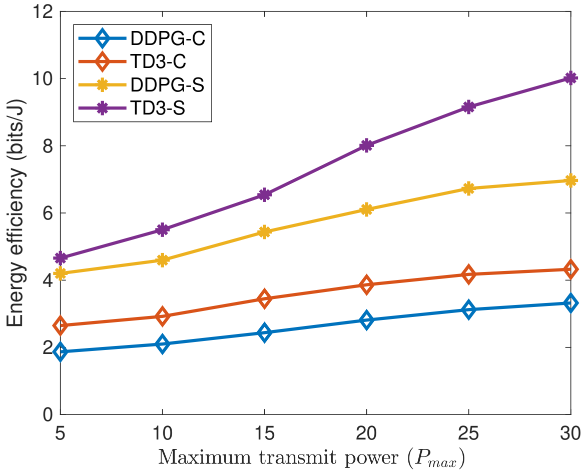

Fig. 2. EE versus max transmit power .

the respective large-scale fading coefficients given by $F = \frac{10^{-3.53}}{d^{3.76}}$, where d denotes the distance between the respective pair of devices. Similarly, $\mathbf{f}_{jv}$, $\mathbf{g}_{kv}$ can also be modelled as a Rician distribution channel (Details have been excluded due to limited space). Moreover, it is postulated that the direct link channel gains conform to a Rayleigh distribution.

Without loss of generality and for the sake of simulations, $M = 4$, $V = 4$, $N = 8$, $P_k = 10$ dBm, $P_j = 10$ dBm, $P_{max} = 50$ dBm, $P_R = 10$ dBm, $P_B = 39$ dBm, $\nu_0 = 0.8$, $\sigma^2 = -104$ dBm, and $B = 1$ MHz. The Rician factor is denoted by $K_v = 5$, and the remaining channels can be obtained similarly. The v^{th} RIS is located at $\left(\cos\left(\frac{2v\pi}{V}\right), \sin\left(\frac{2v\pi}{V}\right)\right) \times 100m$. The users are evenly deployed within a square region of size, $(200m \times 200m)$ with the BS located at its centre $(0, 0)$. The proposed algorithm is implemented using Tensorflow 2.7.0. with $\alpha = 0.99$, $\tau = 0.0001$, learning rate = 0.001, $R = 100000$, $E = 10000$ and number of hidden layer = 512.

To showcase the usefulness of the presented study and effectiveness of the proposed algorithm, which is abbreviated as TD3-S, we compare its performance with several benchmark schemes: 1) the DDPG-based solution, which is abbreviated as DDPG-S, 2) using the passive RIS instead of STAR-RIS and solving them using TD3 and DDPG, which is abbreviated as TD3-P and DDPG-P.

In Fig. 1, we analyse the comparison of the convergence among different methods by plotting it against the number of episodes. It can be observed that the proposed TD3- based algorithm outperforms the DDPG-based solution for both passive and STAR-RISs-aided networks. The main reason for this dominance is the fact that, unlike DDPG, it improves approximation and stability by utilising two distinct critic networks and choosing the lesser value from them to create its targets. By carefully selecting the appropriate learning rate and discount factor in terms of EE performance, the TD3-S framework can disregard irrelevant training and achieve faster optimization. It can also be seen that the use of STAR-RIS provides a significant performance improvement over conventional passive RIS. This is due to the fact that, unlike passive RIS which can only reflect signals towards front users, the STAR-RIS assists both front and back users using its simultaneous reflection and transmission properties.

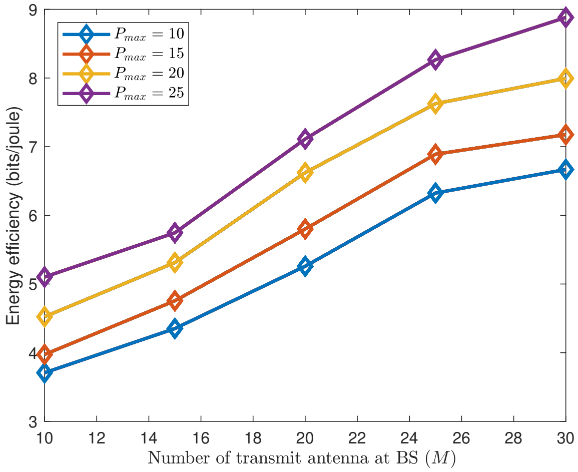

Fig. 3. EE versus M.

Fig. 2 demonstrates the impact of P_{max} on the average EE of the considered system for all the cases. It can be seen that the average EE increases with an increase in P_{max} for all the cases, which is because of the freedom to use a bit more transmit power by the BS leading to an increase in the signal strength at each user. However, it is clearly evident that TD3-S provides better performance compared to the other benchmark cases due the reasons explained earlier. It can also be noted that TD3-S can achieve a similar performance at lower P_{max} which DDPG-S achieves at higher P_{max}. For example, an average EE of around 6.3 bits/j can be achieved by using TD3-S with $P_{max} = 15$ dB or DDPG-S with $P_{max} = 25$ dB. This validates the effectiveness of TD3-S in achieving desired QOS in low power applications.

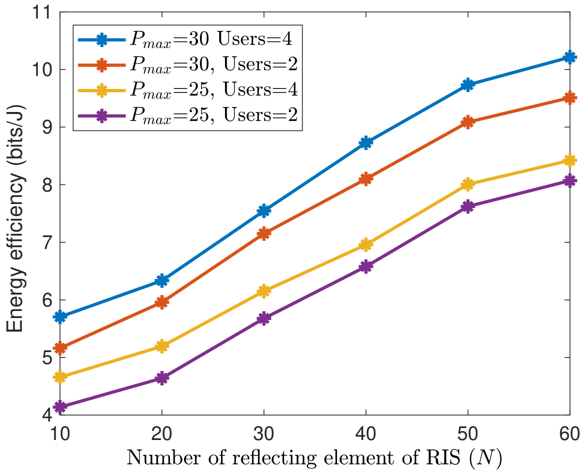

Fig. 4. EE versus N

Fig. 3 highlights the impact of the number of transmit antennas at the BS (M) on the performance of the considered system. It can be seen that due to the enhanced diversity, the average EE increases with an increase in M. However, the rate of increase in EE slows down as M increases with P_{max} = 30. It also demonstrates that the TD3-S is a strong fit for the proposed framework as it effectively handles the shortcomings of DDPG-S and DDPG-C, such as unstable action-value function overestimation. This makes TD3 a preferable option over DDPG.

Fig. 4 shows the impact of variation in the number of reflecting elements at the RIS (N) on the performance of the considered system. It is obtained by plotting average EE w.r.t. N by varying the transmit power and number of users. It can be clearly seen that the EE increases with the use of larger N at the RIS. This performance enhancement is due to increase in beamforming gain leading to a significant increase in the SINR at each user.

VI. Conclusions

We investigated the impact of DRL-based transmission design for a distributed STAR-RIS-assisted communication system. We discussed the effectiveness of two different approaches, DDPG and TD3, towards providing energy efficient resource allocation for the considered system. We demonstrated the effect of key system parameters, such as the maximum transmit power available at the BS, reflecting elements in each RIS and antenna at the BS. The simulation results demonstrated the effectiveness of the proposed algorithms, with the STAR-RIS approach outperforming the CR-RIS.

References

[1] S. Zhou, W. Xu, K. Wang, M. Di Renzo, and M.-S. Alouini, "Spectral and energy efficiency of IRS-assisted MISO communication with hardware impairments," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1366–1369, Sep. 2020.

[2] Y. Liu, X. Mu, X. Liu, M. D. Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface-aided multi-user networks: Interplay between NOMA and RIS," vol. 29, no. 2, pp. 169–176, Apr. 2022.

[3] K. K. Nguyen, S. R. Khosravirad, D. B. da Costa, L. D. Nguyen, and T. Q. Duong, "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 358–368, Dec. 2022.

[4] G. Li, M. Zeng, D. Mishra, L. Hao, Z. Ma, and O. A. Dobre, "Energy-efficient joint beamforming design for IRS-assisted MISO system," in *" in Proc. IEEE Int. Conf. Commun. Workshops,*, Jul. 2021, pp. 1–6.

[5] P. S. Aung, Y. Kyaw Tun, Z. Han, and C. S. Hong, "Energy-efficiency maximization of multiple RISs-enabled communication networks by deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. (ICC),*, May 2022, pp. 2181–2186.

[6] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.,*, vol. 21, no. 5, pp. 3083–3098, May 2022.

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

[8] H. R. Hashempour, H. Bastami, M. Moradikia, S. A. Zekavat, H. Behroozi, and A. L. Swindlehurst, "Secure SWIPT in STAR-RIS aided downlink MISO rate-splitting multiple access networks," *arXiv preprint arXiv:2211.09081*, 2022.

[9] P. P. Perera, V. G. Warnasooriya, D. Kudathanthirige, and H. A. Suraweera, "Sum rate maximization in STAR-RIS assisted full-duplex communication systems," in *Proc. IEEE Int. Conf. Commun. (ICC),*, May 2022, pp. 3281–3286.

[10] F. Fang, B. Wu, S. Fu, Z. Ding, and X. Wang, "Energy-efficient design of STAR-RIS aided MIMO-NOMA networks," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 498–511, Jan. 2023.

[11] Y. Guo, F. Fang, D. Cai, and Z. Ding, "Energy-efficient design for a NOMA assisted STAR-RIS network with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5424–5428, Apr. 2023.

[12] S. Kurma *et al.*, "DRL approach for spectral-energy trade-off in RIS-assisted full-duplex multi-user MIMO systems," in *Proc. IEEE Wireless Commun. Netw. (WCNC),*, May 2023, pp. 1–6.

[13] S. Kurma *et al.*, "Spectral-energy efficient resource allocation in RIS-aided FD-MIMO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, Oct. 2023.

[14] S. Kurma, K. Singh, M. Katwe, S. Mumtaz, and C.-P. Li, "RIS-empowered MEC for URLLC systems with digital-twin-driven architecture," in *Proc. IEEE Int. Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2023, pp. 1–6.

[15] R. H. Yoga Perdana, T.-V. Nguyen, Y. Pramitarini, K. Shim, and B. An, "Deep learning-based spectral efficiency maximization in massive MIMO-NOMA systems with STAR-RIS," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Mar. 2023, pp. 644–649.

[16] Z. Yang, M. Chen, W. Saad, W. Xu, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun*, vol. 21, no. 1, pp. 665–679, Jan. 2022.

[17] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Performance analysis of large intelligent surfaces (LISs): Asymptotic data rate and channel hardening effects," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2052–2065, Mar. 2020.