Deep Reinforcement Learning Based Energy-Efficient Design for STAR-IRS Assisted V2V Users

Shalini Yadav[1] and Rahul Rishi[1]

Dept. of CSE, UIET, MDU Rohtak, 124001, Haryana, India.
shalini.rs.uiet@mdurohtak.ac.in, rahulrishi@mdurohtak.ac.in

Abstract. Vehicle-to-Vehicle communication (V2V-C) is a cutting-edge approach in the world of 6G networks, thanks to the benefits it provides in terms of increased spectrum efficiency and energy efficiency (EE). Despite the potential benefits, there are still significant concerns with V2V-C, such as co-channel interference, cross-channel interference, and the requirement for huge connectivity. To overcome these issues, researchers have turned to simultaneous transmitting and reflecting reconfigurable intelligent surfaces (STAR-IRSs) as auxiliary devices to increase wireless network performance. These surfaces allow users on opposite sides to be served at the same time by sending or reflecting signals. We will investigate the downlink STAR-IRS-assisted communication system in the presence of vehicle-to-vehicle pairs (V2VPs) in this study. In this article, we will look at how to optimise energy efficiency (EE) for the STAR-IRS downlink network. Our research focuses on the hybrid space for optimising EE in the STAR-IRS downlink wireless network. Previous research has been limited to either continuous or discrete spaces, limiting optimisation parameters to either continuous or discrete nature. To get around this limitation, we use a hybrid space to optimise the network's EE, allowing one parameter to be continuous and the other to be discrete. The proposed scheme uses the parametrized deep Q-network (P- DQN) framework for estimating the beamforming vector and phase shift for EE optimisation. The simulation results demonstrate the efficacy of the proposed method in maximising the system's energy efficiency.

Keywords: STAR-IRS · P-DQN · V2V-C · EE

1 INTRODUCTION

Future wireless sixth-generation (6G) communication networks' performance could be improved by reconfigurable intelligent surfaces (IRSs) [1]. IRSs can be characterized as two-dimensional arrays made up of a lot of inexpensively reconfigurable passive elements.These elements are controlled by a FPGA controller.The propagation of wireless signals that are received can be changed by appropriately altering the amplitude and phase responsiveness of these components.The majority of the current research on IRSs is predicated on the idea that they are capable of only reflecting incident signals, necessitating the the placement of both the transmitter and receiver on a singular side of the Intelligent Reflecting surface (IRS).A novel IRS architecture named STAR-IRSs has been proposed to overcome this limitation, which employs simultaneous transmitting and reflecting. STAR-IRSs possess the capability to conduct simultaneous transmission and reflection of incident signals, thereby producing an all-encompassing coverage of space. This is a notable deviation from traditional IRSs that focus on reflection [2]. Buildings in metropolitan areas or hills and plants in rural areas may have substantial shadowing effects on V2X communications, lowering energy efficiency and spectrum efficiency.The utilization of STAR-IRS's reflection and transmission presents an opportunity to establish an alternative propagation channel that bypasses obstructions hindering direct Line-of-Sight (LoS) connection between the transmitter and receiver. As a result, the signal can maintain

a desired transmission rate and prevent penetration loss.The necessity for an intelligent and programmable wireless environment has been met, and IRS has proven to be an effective paradigm for addressing the needs of 5G and even beyond networks [3]. The optimization of IRS performance remains a challenging undertaking, owing to the vast quantity of programmable components and the rectifying capacities of the controller. Consequently, the development of novel techniques is imperative to enhance IRS performance. Reinforcement learning (RL) is a rigorous mathematical framework that enables an autonomous agent to engage with an uncertain and constantly evolving environment, and acquire an optimal policy by increase rapidly the aggregate reward [4]. While Q-learning and similar traditional RL algorithms have shown efficacy in diverse scenarios, their application has been restricted to networks of a small scale. The research community has responded to the challenge of managing large and complex networks by introducing a novel approach termed deep reinforcement learning, which amalgamates the principles of RL and DL.Deep Reinforcement Learning (DRL) effectively addresses the limitations of Reinforcement Learning (RL) in working with small data sets and complex problems through the application of Deep Neural Network's (DNN) function approximation attribute. The DRL technique allows an agent to learn through both online and offline approaches. Consequently, this approach yields optimal outcomes for every state-space-action pair and maximizes the total reward [5].

1.1 Related Work

Various schemes have been proposed by several researchers for increasing the Energy efficiency (EE) of the STAR-IRS network.By concurrently optimising active beamforming, transmission, and reflection coefficients with power constraints, the author maximises the lowest user energy efficiency (EE) [6]. To optimize the spectral efficiency, it is imperative to simultaneously regulate the beamforming intensity for individual users and the phase shift parameters of the STAR-IRS. Proximal policy optimization (PPO), a policy gradient-based reinforcement learning approach, is the method used to successfully address the current issue [7]. In an effort to optimize the long-term energy efficiency (EE) of a system that is subject to time-varying channels and user demands, the author presents a framework that utilizes both active and passive beamforming techniques. To facilitate the optimization of all passive beamforming for STAR-IRSs in an online mode, the author proposes a parallel deep reinforcement learning (DRL) algorithm [8]. In order to attain the highest level of energy efficiency, a number of factors are taken into consideration. These include the optimization of active-beamforming vectors at the transmitter, as well as the determination of appropriate transmission and reflection coefficients at the STAR-IRS. Furthermore, the decoding order of successive-interference cancellation (SIC), time-allocation, and constraints on quality-of-service (QoS), conservation of energy, time-allocation, phase-shifts, and SICdecoding are all carefully evaluated and optimized [9]. The author conducts an investigation into a groundbreaking mobile edge computing (MEC) system that is bolstered by a reconfigurable intelligent surface (STAR-IRS) that simultaneously transmits and reflects data. The author aims to minimize the overall energy consumption of all users by optimizing the transmission and reflection time and coefficients of the STAR-IRS in combination with the transmit power and the amount of data offloaded for each user [10].

1.2 Motivation and Contribution

Optimization of EE (Energy Efficiency) in STAR-IRS downlink wireless network in hybrid space has not yet received adequate attention. Available research either includes continous or discrete space, constrainting our optimization parameter to be either discrete or continous in nature. Considering this , we got motivated to use hybrid space to optimize the EE of the network. Using hybrid space we can have one parameter to be continous or other to be discrete.Contributions to this article are as follows:

- Our proposed wireless communication system network, known as the STAR-IRS network, is designed for the downlink scenario, signal sent from UAV (which is acting as base station) is blocked by building. STAR-IRS is placed on the building so that the user in reflection zone and transmission zone get the signal with high energy efficiency.
- To achieve the objective , we jointly optimize beamforming vector at UAV and phase shift at STAR-IRS by making the problem as an MDP. P-DQN algorithm is suggeated for optimization of parameters.

1.3 Organization

This scholarly article is structured as follows: Section II outlines the system model. The research problem and its corresponding solution are expounded upon in Section III. Section IV presents the numerical results, whereas Section V contains the concluding remarks of the proposed article.

2 System Model and Problem Formulation

2.1 System Model

A STAR-IRS, UAV added with antennas, downlink single cell wireless communication is considered as shown in figure. The proposed system model has a single UAV,single STAR-IRS, miltiple users located in transmission and reflection zone. All users are located on ground. Due to the prescence of obstacles and objects in real world problems, a direct connection between UAV positioned at certain distance and users at some other location is not feasible. In order to make establish a connection betweenUAV and users. A STAR-IRS is positioned at the building such that, when signal comes from the UAV it is incident on STAR-IRS and the service is accessible to people located in both the transmission and reflection regions.The group of all users are denoted by $\mathcal{N} = \{1, 2, \ldots, n\}$ and the group of all V2VPs are denoted by $\mathcal{M} = \{1, 2, \ldots, m\}$. A STAR-IRS is a homogeneous linear arrary of $l \times l$ reflecting elements.Each user sent its signal towards the UAV over each orthogonal subcell through OMA protocols. In each V2VP, the V2VT communicates with its respective V2VR through OMA protocols. The V2VPs and users both use the same subcell. It is assumed that each subcell is occupied by one user. Let J be the total network capacity, divided into $\mathcal{B}$ sub-cells. The set of sub-cells is specified as $\mathcal{B} = \{1, 2, \ldots, b, \ldots, B\}$, while the set of time slots is defined as $\mathcal{T} = \{1, 2, \ldots, t \ldots, T\}$ The UAV is positioned at $(x_{uav}, 0, G)$ where G is the UAV height. Let users are located in X-Y plane with their 3D Coordinates $(x_n, y_n, 0)$. Let us define Coordinates of STAR-IRS $X_{STAR-IRS} = (x_l, y_l, z_l)$. The distance between the UAV and l^{th}element of STAR-IRS is $d_{uav,l(t)} = \sqrt{(x_{uav} - x_l)^2 + (y_l)^2 + (z_l - G)^2}$. The distance. The distance between (STAR-IRS) l^{th} and n^{th} user $d_{l,n}(t) = \sqrt{(x_n - x_l)^2 + (y_n - y_l)^2) + (z_l)^2}$

2.2 Channel Model

As seen in figure 1, UAV equipped with D antennas transmits signals to multiple single-antenna users through a STAR-IRS consisting of L elements.Users O_1,O_2,..., O_N are situated at the back of the transmission zone.While the users M_1,M_2,...,M_N are situated in the front of the STAR-IRS in the reflection zone. We assume that walls or buildings are obstructing all direct connections between the UAV and all users. Individuals utilizing the system undergo clustering and are subsequently distributed via the downlink transmission mechanism [2]. Let $f^b_{l,n} \in \mathbb{C}^{1\times L}$ and $\widehat{L}^b_{uav,l} \in \mathbb{C}^{1\times M}$ defines the channel between STAR-IRS l^{th} component to the n^{th}user and from UAV to the l^{th} component of STAR-IRS respectively. Let $\widetilde{L}^b_{uav,n} \in \mathbb{C}^L$ define

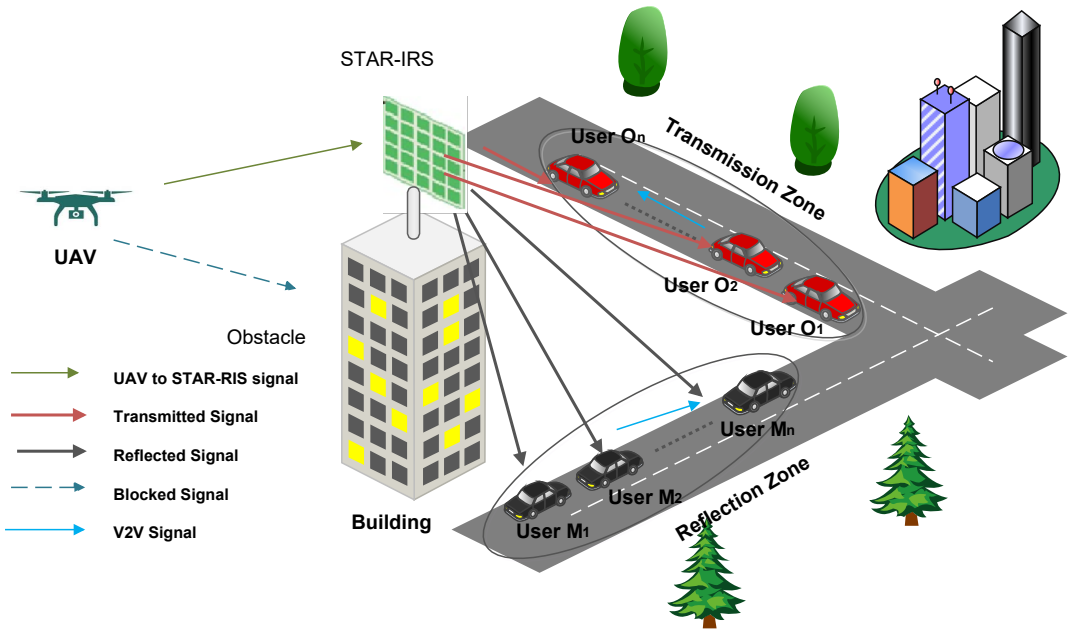

Fig. 1: Network Architecture.

channel from UAV to the n^{th} user. $g^b_{m,i} \in \mathbb{C}^L$ define the channel from m^{th} V2VT to i^{th} V2VR in b^{th} cell. Energy splitting protocol is considered to be followed by STAR-IRS [11], [12]. It implies that by implementing coefficient matrices concurrently, every component at STAR-IRS is capable of transmitting and reflecting incident signals simultaneously.The ES protocol's rule of amplitude coefficients is defined by the ideal situation in which the aggregate energy of the transmitted and reflected signals equates to the energy of the incident signals for every element $\lambda^O_n + \lambda^M_n = 1, \forall n \in \{1, 2, 3..., N\}$, λ^O_n and λ^M_n denote the transmission and reflection amplitudes coefficients for the n element at the STAR-IRS respectively.The coefficients matrices at the STAR-IRS are given by $\Theta^\nu = diag(\sqrt{\lambda^\nu_1}\exp^{j\phi^\nu_1}, \sqrt{\lambda^\nu_2}\exp^{j\phi^\nu_2}, ...\sqrt{\lambda^\nu_N}\exp^{j\phi^\nu_N}), \nu \in \{O, M\}$,where $\phi^\nu_n \in [0, 2\pi), \forall n \in \{1, 2, 3, \ldots, N\}$ denotes the phase shift for the n-th element.

UAV to User Channel Model The User gets the two signal. There is a direct signal UAV to the user whereas the second signal is transmiteed or reflected signal (UAV to User) through STAR-IRS.We assume that our system follow Rayleigh fadding.$f^b_{l,n}$,$\widehat{L}^b_{uav,l}$,$\widetilde{L}^b_{uav,n}$ are expressed as :

$$f^b_{l,n} = \sqrt{\delta_0 (a^b_{l,n})^{-\varphi_2}}$$

$$\left(\sqrt{\frac{E^b_{l,n}}{1 + E^b_{l,n}}} f^{LoS,b}_{l,n} + \sqrt{\frac{1}{1 + E^b_{l,n}}} f^{NLoS,b}_{l,n} \right), \tag{1a}$$

$$\widehat{L}^b_{uav,l} = \sqrt{\delta_0 (a^b_{uav,l})^{-\varphi_1}}$$

$$\left(\sqrt{\frac{E^b_{uav,l}}{1+E^b_{uav,l}}}\widehat{L}^{Los,b}_{uav,l}+\sqrt{\frac{1}{1+E^b_{uav,l}}}\widehat{L}^{NLoS,b}_{uav,l}\right), \tag{1b}$$

$$\widetilde{L}^b_{uav,n}=\sqrt{\delta_0(a^b_{uav,n})^{-\varphi_3}}\widetilde{L}^{NLoS,b}_{uav,n}, \tag{1c}$$

where path loss exponent and propagation loss at the reference distance is shown by δ_0 and φ respectively. $E^b_{l,n}$ and $Eb_{uav,l}$ are the Rician factors for STAR-IRS to nth user and uav to the STAR-IRS links.NLOS repsesents the Non Line of Sight and their elements are denoted by $f^{NLOS,b}_{l,n}$, $\widehat{L}^{NLOS,b}_{uav,l}$ and $\widehat{L}^{NLOS,b}_{uav,n}$ are described by Rayleigh fading.All elements belongs to $\mathcal{C}N(0,1)$. Hence, $f^{NLOS,b}_{l,n}$ and $\widehat{L}^{NLOS,b}_{uav,n}$ are defined as follows :

$$f^{LoS,b}_{l,n}=\left[1,\ldots,e^{j(n-1)\pi\ \sin(A_{l,n})},\ldots,e^{j(n-1)\pi\ \sin(A_{l,n})}\right]^{\mathbb{T}} \tag{2a}$$

$$\widehat{L}^{LoS,b}_{uav,n}=\left[1,\ldots,e^{j(uav-1)\pi\ \sin(A_{uav,n})},\ldots,e^{j(uav-1)\pi\ \sin(A_{uav,n})}\right]^{\mathbb{T}}, \tag{2b}$$

where $\sin(A_{l,n})=\frac{y_n-y_l}{\sqrt{(x_n-x_l)^2-(y_n-y_l)^2}}$ and $\sin(A_{uav,n})=\frac{y_uav}{\sqrt{(x_n-x_{uav})^2-(y_uav)^2}}$.

The signal transmitted by the uav to the n^{th} user can be defined as follows:

$$Z^n_{uav,n}=\underbrace{\left(f^b_{l,n}\Theta\widehat{L}^b_{uav,n}+\widetilde{L}^b_{uav,n}\right)\sum_{d\in\Psi}t_dr_d}_{\text{Reflected or Transmitted + Direct Signal}} \tag{3}$$

$$\underbrace{+\sum_{m\in\mathcal{M}}\epsilon^b_m(t)\sqrt{h^b_m}(t)g^b_{uav,m}r^b_{uav,m}}_{\text{UAV to V2VT Interference}}+\underbrace{\sigma^2}_{\text{AWGN}},$$

where $t_d\in\mathbb{C}^{M\times1}$ denotes the beamforming vector r_d is the signal symbol of the user and we assume $\mathbb{E}\{r_d\}^2=1$, $\sqrt{h^b_n}$ and $\sqrt{h^b_m}$ are the n^{th} and m^{th} V2VT transmitted power respectively. Transmitted symbols for the n^{th} user and m^{th} V2VT are denoted by $r^b_{n,uav}$ and $r^b_{m,uav}$ respectively. Cell allocation coefficient for n^{th} user and m^{th} V2VT link is given by $\epsilon^b_{n,m}$

$$\epsilon^b_{n,m}=\begin{cases}1 & \text{if } m^{th} \text{ V2VT occupies } n^{th} \text{ user,}\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

The SINR over the b^{th} subcell for the n^{th} user is given as:

$$\Gamma^b_{n,uav}=\frac{h_n\left|f^b_{n,l}\Theta\widehat{L}^b_{n,uav}+\widetilde{L}^b_{uav,n}\right|^2}{\sum_{m\in\mathcal{M}}h_m\left|g^b_{m,uav}\right|^2+\sigma^2} \tag{5}$$

V2VT-V2VR Channel Model The signal received by i^i V2VR from the m^{th} V2VT over the b^{th} subcell is given by

$$Z^b_{m,i}=\underbrace{\sqrt{h^b_m}g^b_{m,i}r^b_{m,i}}_{\text{Direct Signal}}+\underbrace{\sum_{i'\neq i,i'\in\mathcal{N}}\epsilon^b_m\sqrt{h^b_m}p^b_{m,i'}r^b_{m,i'}}_{\text{V2VT to V2VR interference}}$$

$$+ \underbrace{\sum_{n\in\mathcal{N}} \sqrt{h_n^b} f_{n,i}^b r_{n,i}^b}_{\text{User to V2VR interference}} + \underbrace{\sigma^2}_{\text{AWGN}}, \tag{6}$$

SINR for the ith V2VR over b^{th} is defined as

$$\Gamma_{m,i}^b = \frac{h_m \left| g_{m,i}^b \right|^2}{\sum\limits_{i'\neq i, i'\in\mathcal{M}} \epsilon_{n,i}^b h_m^b \left| g_{m,i'}^b \right|^2 + \sum\limits_{n\in\mathcal{N}} h_n^b \left| f_{n,i}^b \right|^2} \tag{7}$$

2.3 Energy Efficiency Evaluation

The transmission rate for the n^{th} users is given by as:

$$\mathbb{S}_{uav,n}^b = BW \log_2 \left[1 + \Gamma_{uav,n}^b\right]. \tag{8}$$

The transmission rate for the m^{th} V2VR is given as:

$$\mathbb{S}_{i,m}^b = BW \log_2 \left[1 + \Gamma_{i,m}^b\right] \tag{9}$$

The combined transmission rate of network is given as:

$$\mathbb{S}_{n,m}^b = \sum_{b\in\mathcal{B}} \left[\sum_{n\in\mathcal{N}} \mathbb{S}_{uav,n}^b + \sum_{m\in\mathcal{M}} \mathbb{S}_{i,n}^b \right]. \tag{10}$$

The following expression shows the network"s total power consumption.

$$\mathbb{P}_T^b = h_{UAV}^b + h_{STAR-IRS}^b + \sum_{n=1}^{N} h_n^k + \sum_{m=1}^{M} h_m^b, \tag{11}$$

In equation 11 $h_{STAR-IRS}^b$ represent the power usage of STAR-IRS,h_{UAV}^b term represent the power usage of UAV,$\sum_{n=1}^{N} h_n^b$ represent the power usage of user's,$\sum_{m=1}^{M} h_m^b$ represents the power usage ofV2VTs.

2.4 Problem Formulation

In this article, our principal objective is to optimize the energy efficiency (EE) of the network which we have proposed. The of energy efficiency (EE) can be defined as::

$$EE = \frac{\mathbb{S}_{n,m}^b}{\mathbb{P}_T^b}$$

$$= \frac{\sum\limits_{b\in\mathcal{B}} \left[\sum\limits_{n\in\mathcal{N}} \mathbb{S}_{uav,n}^b + \sum\limits_{m\in\mathcal{M}} \mathbb{S}_{i,m}^b \right]}{h_{UAV}^b + h_{STAR-IRS}^b + \sum_{n=1}^{N} h_n^k + \sum_{m=1}^{M} h_m^b} \tag{12}$$

This paper's primary goal is to amximize energy efficiency while simultaneously preserving user's and V2VTs SINR by jointly optimizing the STAR-IRS's phase shift and the beamforming vector . Accordingly, The problem of optimization can be expressed in the following manner:

$$\begin{aligned} \mathcal{P}.\mathcal{F}.: & \min_{(t_d,\phi)} EE(t), \\ s.t. \quad & \mathbb{D}_1 : \sum_{b=1}^{B} \epsilon_{n,m}^{b} \leq 1, \qquad \forall \mathcal{M}, \mathcal{N}, \\ & \mathbb{D}_2 : \Gamma_{uav,n}^{b} \geq \Gamma_{uav,n}^{b,\min}, \qquad \forall \mathcal{B}, \\ & \mathbb{D}_3 : \phi \in [o, 2\pi], \qquad \forall \mathcal{L}, \\ & \mathbb{D}_4 : X_{STAR-IRS} \in D, \end{aligned} \tag{13}$$

$\mathbb{D}_1$ ensures that each user is associated with a single V2VT. The minimal data rate requirement for user is represented by $\mathbb{D}_2$.$\mathbb{D}_3$ specifies the restricted range of IRS phase shift. $\mathbb{D}_4$ specifies that the UAV's position should be in the restricted area (D).The non-convexity of the optimization problem (13) is intricately linked to the fractional representation of the EE, in conjunction with the manifold constraints. This non-convexity poses a challenge in obtaining an optimal solution using convex optimization methods such as sequential convex approximation (SCA). We develop a DRL-based approach that jointly maximise EE by optimising beamforming vectors at the UAV and phase shift at the STAR-IRS. Deep Reinforcement Learning (DRL) is a paradigm of Artificial Intelligence (AI) technology that facilitates the development of fully autonomous agents by means of interacting with the environment and utilizing specific optimal strategies. Over time, the agents improve their performance through a process of trial and error.[10]. Our optimization quandary can be effectively resolved through utilization of P-DQN, a type of DRL that has been designed to address optimization problems that exist in hybrid spaces.

3 Proposed Scheme

We are jointly optomizing beamforming vector at UAV and phase shift at STAR-IRS by formulating the above described problem (13) as an markov decision process(MDP). We have put forth the P-DQN algorithm as a means of achieving joint optimization of both beamforming vector and phase shift.

3.1 Markov Decision Process Model

The optimization problem is transformed into MDP problem. An MDP problem has environment, Agent ,State space, Action space and reward.

3.2 Environment

The proposed communication systems make up our environment. An agent interacts with this environment to determine the best courses of action to implement in order to maximize cumulative rewards. The environment contains all network-related data, including UAV, vehicle, and STAR- IRS. The environment includes the characteIRStics of STAR-IRS components, the condition of the vehicles, and channel information. An agent observes a state $s^b(t)$ from the state space $\widehat{\mathcal{S}}$ at each time step t and then acts on that observation by selecting an action $a^b(t)$ from the action space $\widehat{\mathcal{A}}$. The environment's current state s[n] changes to the following state $s^b(t+1)$ when the action is completed. The agent also receives their current reward, $r^b(t)$.

3.3 Agent

In the proposed system STAR-IRS controller acts as an agent. STAR-IRS phase shift is controlled by STAR-IRS controller.

3.4 State Space

Agent checks the state and environment which consists of beamforming vector, phase shift, channel information.

$$s_t^b = [\phi, t_d, f, L] \tag{14}$$

3.5 Action Space

In the present study, the optimization of the beamforming vector of the unmanned aerial vehicle (UAV) and the phase shift of the (STAR-IRS) are concurrently considered to enhance the energy efficiency of the overall system. Therefore, the system's action space is

$$a_t^b = [\phi, t_d] \tag{15}$$

Note that beamforming vector has continous values and phase shift has discrete value so accordingly action space is hybrid.

3.6 Reward

The objective of the optimization problem at hand is to effectively maximize the energy efficiency of the system. As such, energy efficiency has been designated as the reward for the t-th training step.

$$r_t^b = EE(t) \tag{16}$$

We consider an MDP model having paramet IRSed action space $\widehat{\mathcal{A}}$. For $a^b \in \widehat{\mathcal{A}}$, Action valur function defined by $\mathcal{Q}(s^b, a^b)$ =$\mathcal{Q}(s^b, k, x_k^b)$ where $s^b \in \widehat{\mathcal{S}}$ k$\in [K]$.Let k_t is the discrete action chosen at time t and x_k^b be the associated continuous parameter.
So the Bellmen equation is

$$\mathcal{Q}(s_t^b, k_t, x_{kt}^b) = \mathbb{E}_{r_t^b, s_{t+1}^b}\left[r_t^b + \Upsilon \max_{k^b \in [K]} \mathcal{Q}(s_{t+1}^b, k^b, {x^b}_k^Q s_{t+1}^b) | s_t^b = s^b\right]. \tag{17}$$

Similar to deep Q-networks, Deep neural network $\mathcal{Q}(s^b, k^b, x_k^b; w^b)$ is used to approximate $\mathcal{Q}(s^b, k^b, x_k^b)$ where w^b represent the network weights.${x^b}_{k^b}^Q(s^b)$ is approximated with dterministic policy network $x_k^b(\theta^b)$: $\widehat{\mathcal{S}} \rightarrow \mathcal{X}_k^b$, where θ^b represent the network weights of the policy network. We want to determine θ^b such that when w^b is fixed such that

$$\mathcal{Q}(s_t^b, k, x_k^b(s^b; \theta^b); w^b) = sup_{x_k^b \in \mathcal{X}_k} \mathcal{Q}(s^b, k^b, x_k^b; w^b) \tag{18}$$

Then, using gradient descent to minimize the mean-squared Bellman error, we could estimate similarly to DQN. Let θ_t^b and w_t^b be the weights of the deterministic policy network and the value network, respectively, in the t-th step.We define the n-step target y_t^b by for a fixed $n \geq 1$ to accommodate multi-step techniques.

$$y_t^b = \sum_{i=0}^{n-1} \gamma^i r_{t+i}^b + \gamma^n \max_{k^b \in [K]} \mathcal{Q}(s_{t+n}^b, k^b, x_k^b(s_{t+n}^b, \theta_t^b); w_t^b) \tag{19}$$

The least squares loss function is used for w^b as in DQN. In addition, given our goal is to identify θ^b that maximize Q while holding w^b constant, we employ the following loss function for θ^b.

$$l_t^{\mathcal{Q}}(w^b) = 1/2[\mathcal{Q}(s_t^b, k_t^b, x_{k_t}^b; w) - y_t^b]^2 \tag{20}$$

$$l_t^{\mathcal{Q}}(\theta^b) = -\sum_{k=1}^{K} \mathcal{Q}(s_t^b, k^b, x_k^b(s_t; \theta^b); w_t^b) \tag{21}$$

By (20,21), we use stochastic gradient methods to update the weights.

We update w^b specifically with a stepsize α_t^b that is asymptotically negligible in compaIRSon to the stepsize β_t^b for θ^b. Additionally, we need α_t^b, β_t^b to meet the Robbins-Monro criterion for the stochastic approximation to be valid [Robbins and Monro, 1951]. In algorithm 1, we propose the P-DQN algorithm with experienced replay.Furthermore, we point out that asynchronous gradient descent can be simply added to our P-DQN technique to hasten training. We think about a distributed training system that is centrally managed, where each process computes its own local gradient and interacts with a global "parameter server" to share information.In particular, each local process uses its own transitions to compute gradients with respect to omega and theta and operates an independent environment to produce transition trajectories. Algorithm 2 contains the asynchronous n-step P-DQN (AP-DQN) algorithm. Here, we will just discuss the technique for each local process, that fetches ω^b and θ^b which retrieves and computes the gradient using the parameter server. The gradients sent from the local processes are used by the parameter server to update the global parameters.

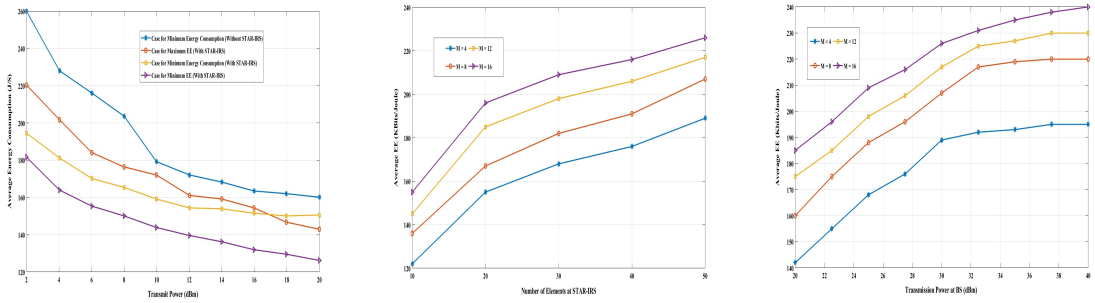

Fig. 2: Comparative Analysis (a) Average energy consumption over transmit power. (b) EE versus the number of elements at the STAR-IRS (c) EE versus the transmission power at the UAV(BS) .

4 Performance Evaluation

4.1 Simulation Parameters

The phase shift parameter of the STAR-IRS is arbitrarily assigned at the initial time-slot, and the UAV(BS) is positioned at a fixed location within the simulation. A cell surrounding the UAV contains distribution of N Users and M V2VPs. To construct the P-DQN training model, we utilized a structure compIRSed of three fully connected layers (input layer, hidden layers, output layer) with 500, 500, and 250 neurons respectively. Table I provides a comprehensive list of the remaining simulation parameters.

Algorithm 1 P-DQN with Experience Replay for Phase Shift and Beamforming Vector Coefficients

Input

- Step Sizes = $\{\alpha_t^b, \beta_t^b\} \forall t \geq 0$
- Exploration Parameter = ϵ
- Minibatch Size = E
- Probability Distribution = ξ

Initialization:

- Network Weights = ω^b & θ^b

for $(t = 1, t \leq T, t++)$ **do**
  Compute action parameters $x_k^b \leftarrow x_k^b(s_t^b, \theta_t^b)$
  Select action $a_t^b = (k_t^b, x_{k_t}^b)$
  According to ϵ-greedy policy, the value of a_t^b is given as follows:

$$a_t^b = \begin{cases} \text{a sample from distribution,} & \epsilon \\ (k_t^b, x_{k_t}^b), where\ k_t^b = \max_{k \epsilon [K]} Q(s_t^b, w_t^b) & 1 - \epsilon. \end{cases}$$

  Choose action a_t^b
  Renew the reward function as in (24)
  Evaluate the subsequent state, $s_{(t+1)}^b$
  Retain transition $\left(s_t^b, a_t^b, r_t^b, s_{(t+1)}^b\right)$ in experience replay buffer of capacity $\mathbb{G}$
  A small batch of transitions should be randomly selected $(s_i^b, a_i^b, r_i^b, s_{i+1}^b)$ from the replay buffer of capacity $\mathbb{G}$
  Define the target y_i^b as:

$$y_i^b = \begin{cases} r_i^b, & s_{i+1}^b \text{ is terminal state} \\ r_i^b + \max_{k^b \in [K]} \gamma Q(s_{i+1}^b, k^b, x_k(s_{i+1}^b, \theta_t^b); w_t^b) & \text{otherwise.} \end{cases}$$

  Computer stochastic gradient $\nabla_w l_t^Q(w^b)$ and $\nabla_\theta l_t^\Theta(\theta^b)$ using data (y_i^b, s_i^b, a_i^b)
  Update the Weights as follows:

$$w_{t+1}^b = w_t^b - \alpha_t^b \nabla_w l_t^Q(w_t^b)$$
$$\theta_{t+1}^b = \theta_t^b - \beta_t^b \nabla_\theta l_t^\Theta(\theta_t^b)$$

end for
Output: θ^b and P_v^r

Algorithm 2 AP-DQN with fast training process for Phase Shift and Beamforming Vector Coefficients

Input

- Exploration = $\mathcal{A}$
- Exploration Parameter = ϵ
- Probability Distribution = ξ
- Maximum length of multi step return = $t_{\max}$
- Maximum number of iterations = $\mathcal{N}_{step}$

Initialization:

- Network Weights = ω^b & θ^b
- Global Shared Counter = $\mathcal{N}_{step} = 0$
- Local step Counter $t \leftarrow 1$

repeat
  Clear local gradients $dw \leftarrow 0$, $d\theta \leftarrow 0$
  $t_{start} \leftarrow 0$
  Synchronize local parameters $w_1 \leftarrow w$, $\theta_1 \leftarrow \theta$ from the parameter server
  repeat
    Compute state s_t^b
    Assume $x_k^b \leftarrow x_k^b(s_t^b, \theta_1)$
    According to ϵ-greedy policy, the value of a_t^b is given as follows:

$$a_t^b = \begin{cases} \text{a sample from distribution,} & \epsilon \\ (k_t^b, x_{k_t}^b), k_t^b = \max_{k^b \epsilon [K]} Q(s_t^b, w_t^b) & 1-\epsilon. \end{cases}$$

    Select action $a_t^b = (k_t^b, x_{k_t}^b)$
    Renew the reward function r_t^b using (24)
    Determine the next state, $s_{(t+1)}^b$
    $t \leftarrow t+1$
    $\mathcal{N}_{step} \leftarrow \mathcal{N}_{step} + 1$
  until s_t^b is the terminal state or $t_{\max} = t - t_{start}$
  Define the target y_i^b as:

$$y_i^b = \begin{cases} 0, & \text{terminal state } s_t^b \\ \max_{k^b \in [K]} Q(s_t^b, k^b, x_k(s_t^b, \theta_1); w_1) & \text{non-terminal.} \end{cases}$$

  for $(t = t-1, t \leq t_{\max}, t_{\max}$++**)** **do**

$$y^b = r_i^b + \gamma y^b$$

    Update the Gradients as follows:

$$dw^b = dw^b + \nabla_{w^b} l_t^Q(w_1)$$

$$d\theta^b = d\theta^b + \nabla_\theta^b l_t^\Theta(\theta_1)$$

  end for
  Update global θ^b and w^b using $d\theta^b and dw^b$ using RMSProp
until $\mathcal{N}_{step} > \mathcal{N}_{\max}$ **Output**: θ and P_v^r

Table 1: Simulation Parameters

Parameters	**Values**
Cellular cell's Radius	400m
Carrier Frequency	1.5GHz
Distance between STAR-IRS and BS	50m
User transmission power	4W
Channel Power Gain	-30dB
Noise Power spectrum density	-172 dBm/Hz
Path loss exponent	5
Distance between STAR-IRS and User	10m
Pathloss MU-IRS links	$150 + 40\log d$
V2VPs number	5
Users number	7
Rician factor	6 dB
Factor of Discount	0.8
Starting learning rate	0.3
Declining learning rate	0.002
Replay storage capacity	1000
Small-batch Size	64
Number of Steps in Each Epoch	25
Episodes	100
Optimizer	Adam
Activation function	ReLu

4.2 Results and Discussion

The figure depicted in 2(a) displays the comprehensive energy consumption of the STAR-IRS system inclusive of the transmitter and receiver. It is noteworthy that an increase in transmit power results in a decrease in energy consumption for the STAR-IRS system. The base station (BS) expends the greatest amount of energy when attempting to reduce energy consumption without the aid of STAR-IRS. Figure 2a depicts the performance of the Energy Efficiency (EE) as a function of the number of elements at the STAR-IRS. Evidently, the system's EE demonstrates an upward trend with an increase in the number of elements at the STAR-IRS.Figure 2(c) presents the Energy Efficiency (EE) as a function of Transmission Power (TP) at the Base Station (BS) with varying numbers of antennas. The increment in Transmission Power (TP) at the Base Station (BS) leads to an increasing trend in the Energy Efficiency (EE), which attains a maximum value before stabilizing. This trend indicates that the Energy Efficiency (EE) cannot exhibit perpetual growth despite the continuous increase in power at the Base Station (BS).

5 Conclusion

In this paper, we designed an optimisation problem of energy-efficiency relevant to a downlink network employing a STAR-IRS. The proposed solution entails the creation of a P-QDN-based algorithm that allows for the joint optimisation of beamforming vectors at the Base Station (BS) and phase shift at the STAR-IRS, resulting in the maximisation of Energy Efficiency (EE). Numerical results have shown that the suggested approach is effective and convergent. Furthermore, we examined the EE trends obtained for varied levels of transmission power at the BS and the varying number of elements at the STAR-IRS.

Bibliography

[1] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE communications magazine*, vol. 58, no. 1, pp. 106–112, 2019.

[2] 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, 2021.

[3] I. Budhiraja, V. Vishnoi, N. Kumar, D. Garg, and S. Tyagi, "Energy-efficient optimization scheme for ris-assisted communication underlaying uav with noma," in *ICC 2022-IEEE International Conference on Communications*, pp. 1–6, IEEE, 2022.

[4] H. Sharma, I. Budhiraja, P. Consul, N. Kumar, D. Garg, L. Zhao, and L. Liu, "Federated learning based energy efficient scheme for mec with noma underlaying uav," in *Proceedings of the 5th international ACM mobicom workshop on drone assisted wireless communications for 5G and beyond*, pp. 73–78, 2022.

[5] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[6] K. Wang, L. Xue, Z. Yang, and M. Peng, "Max-min energy-efficiency fair optimization in star-ris assisted communication system," *IEEE Access*, 2023.

[7] P. S. Aung, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Deep reinforcement learning based spectral efficiency maximization in star-ris-assisted indoor outdoor communication," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6, IEEE, 2023.

[8] J. Chen, Z. Ma, Y. Zou, J. Jia, and X. Wang, "Drl-based energy efficient resource allocation for star-ris assisted coordinated multi-cell networks," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 4232–4237, IEEE, 2022.

[9] M. Asif, A. Ihsan, W. U. Khan, Z. Ali, S. Zhang, and S. X. Wu, "Energy-efficient beamforming and resource optimization for star-irs enabled hybrid-noma 6g communications," *IEEE Transactions on Green Communications and Networking*, 2023.

[10] Q. Zhang, Y. Wang, H. Li, S. Hou, and Z. Song, "Resource allocation for energy efficient star-ris aided mec systems," *IEEE Wireless Communications Letters*, vol. 12, no. 4, pp. 610–614, 2023.

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

[12] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (star) ris aided wireless communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3083–3098, 2021.