Article

Joint Beamforming Design for RIS-assisted Integrated Satellite-HAP-Terrestrial Networks Using Deep Reinforcement Learning

Min Wu [1], Shibing Zhu [1], Changqing Li [1], Yudi Chen [1,*], Feng Zhou [2] and Xiang Su [3]

[1] Space Engineering University; 1800022837@pku.edu.cn (M. W.); sbz_zhu@sohu.com (S. Z.);lcqqcl5577@sohu.com (C. L.); chenyudi9438@163.com (Y. C.);
[2] Yancheng Institute of Technology, Yancheng, China; zfycit@ycit.edu.cn (F. Z.);
[3] 714 Research Institute of China State, Beijing, China; sx_fly@0603@163.com (X. S.);
* Correspondence: sbz_zhu@sohu.com;

Abstract: In this paper, we consider a reconfigurable intelligent surface (RIS)-assisted integrated satellite-high altitude platform-terrestrial networks (IS-HAP-TNs) that can improve network performance by exploiting HAP's stability and RIS's reflection. Specifically, the reflector RIS is installed on the side of HAP to reflect signals from the multiple ground user equipments (UEs) to the satellite. To aim at maximising system sum rate, we jointly optimize the transmit beamforming matrix at the ground UEs and RIS phase shift matrix. Due to the limitation of the unit modulus of the RIS reflective elements constraint, the combinatorial optimization problem is difficult to tackle it effectively by traditional solving methods. Based on this, this paper studies deep reinforcement learning (DRL) algorithm to achieve online decision making for this joint optimization problem. In addition, it is verified through simulation experiments that the proposed DRL algorithm outperforms the standard scheme in terms of system performance and execution time, and higher computing speed, making real-time decision making truly feasible.

Keywords: Reconfigurable intelligent surface (RIS); integrated satellite-HAP-terrestrial networks (IS-HAP-TNs); deep reinforcement learning (DRL); optimization performance;

1. Introduction

As fifth-generation mobile communication systems enter commercial operation worldwide, terrestrial wired and wireless networks are beginning to provide instant, high-speed data transmission services to users in high-density population areas, but due to geographical conditions and business models, networks in remote areas are still unable to meet multiple users' needs for full-area coverage and ubiquitous access. Compared with traditional terrestrial wireless communication systems, the integration of satellite, aerial platform and terrestrial communications into the integrated satellite-high altitude platform-terrestrial networks (IS-HAP-TNs) have emerged as a very potential infrastructure in the future wireless communication networks, which can establish seamless coverage and massive connectivity for the explosive growth of terrestrial users [1,2]. Nevertheless, IS-HAP-TNs also raise serious concern about the rapidly growing energy consumption and wireless security in the transmission process, which are of great significance for maintaining green and reliable communication schemes [3].

Among the various candidates, a novel energy-efficient mode, known as reconfigurable intelligent surface (RIS), has been widely applied to improve communication security and network performance[4,5]. Each of the RIS reflective element is a varactor diode that allows the amplitude and/or phase shift of the incident signal to be independently controlled by an embedded RIS central controller [6]. An extensive study in [7] shows that RIS has already been applied in many different communication network scenarios, such as ambient reflectors, signal transmitters, and even signal receivers.

Meanwhile, RIS is also used in ambient forward scatter/backscatter communication systems, which is a seminal contribution, as in [8].

Recently, deep reinforcement learning (DRL) has made a splash in non-convex optimization problems, including hybrid beamforming design[14], spectrum intelligence sensing [15], channel state estimation [16], and power allocation strategy optimization [17]. Compared with deep learning (DL), the DRL algorithm does not require a large amount of training labeled data as inputs and is therefore very friendly for optimization of wireless communication systems where obtaining data is more tedious. By interacting with the environment to obtain rewards from the network, DRL can learn and construct wireless channel knowledge without knowing the complete channel model information and the precise movement pattern, while implementing efficient algorithm design through embedded neural networks to sequentially find optimal solutions to complex multi-objective optimization problems. In [18], a deep Q-network (DQN) with greedy characteristics is proposed for the jointly optimizing of beamforming design, power allocation strategy and interference coordination for maximizing the signal to interference plus noise ratio (SINR). In [19], by using the DRL framework, the user distribution model is tracked and predicted to autonomously and dynamically optimize the MIMO broadcast beam and propose the optimal broadcast beam for each served cell. The results confirm that optimal coverage can be achieved using the DRL framework in both single-sector and multi-sector environments, and in both periodic and Markovian mobility modes.

Currently, the joint beamforming design technique is also a hot issue, which can greatly improve the communication efficiency, system capacity and transmission rate of wireless communication systems to some extent. Motivated by the above analysis, our objective is to design a novel DRL framework that considered as the DDPG algorithm for the jointly optimization problem under the proposed RIS-assisted IS-HAP-TNs [20]. In particular, considering the high dynamics of RIS-assisted IS-HAP-TNs transmission process, the main work and contributions of this paper can be summarized as follows

- Firstly, considering the time-varying characteristics of IS-HAP-TNs fading channel model and signal transmission model, the system sum rate formulations are given under this system model constraints using the active transmit beamforming at the ground user equipments and the phase shift matrix at the RIS, and the maximization expressions under the proposed constraints.
- Secondly, a soft-update strategy framework based on DDPG framework is designed to optimize the above target problems. The framework does not need to know the explicit model and specific mobile model of wireless environment, and can well deal with the continuous state space, action space and reward function, and solve the formal problem of the system.
- Finally, the simulation experiments on the number of RIS elements as well as the average reward show that the designed DRL algorithm framework outperforms other algorithms, which is a guideline for real-time decision making in dynamic IS-HAP-TNs communication environments.

The remainder of this paper is arranged as follows. Section II describes the considered system model and identifies the optimization objective problem under the constraints. Section III gives the basic framework of the soft update parameter strategy and gives the design flow for the optimization of the active transmit beamforming matrix and the RIS phase shift matrix under this framework. Section IV plots the network performance simulation results under this framework and provides a detailed theoretical analysis. Finally, Section V concludes the whole work.

2. System Model Description

In this illustration, we envision an uplink transmission communication system includes geosynchronous earth orbit (GEO) satellite, backward high altitude platforms (HAPs) deployed with RIS, the K ground user equipments (UEs) employs a single antenna as shown in Fig. 1. In our proposed system model, the UEs transmission communication information through RF links to the RIS which installed on the HAP with M reflective elements, which acts as a reflecting relay with changeable transmission links, and sends the received signal to the satellite.

It is noted that the satellite are linked to the cloud data computing processing center by free-space optical (FSO), which can collect global communication information such as the user's requirements as system control link. Instead of coding satellites and HAPs separately, it centralizes the baseband processing of the entire network in the cloud, with the cloud as the core, taking into account resource management and environmental feedback [21].

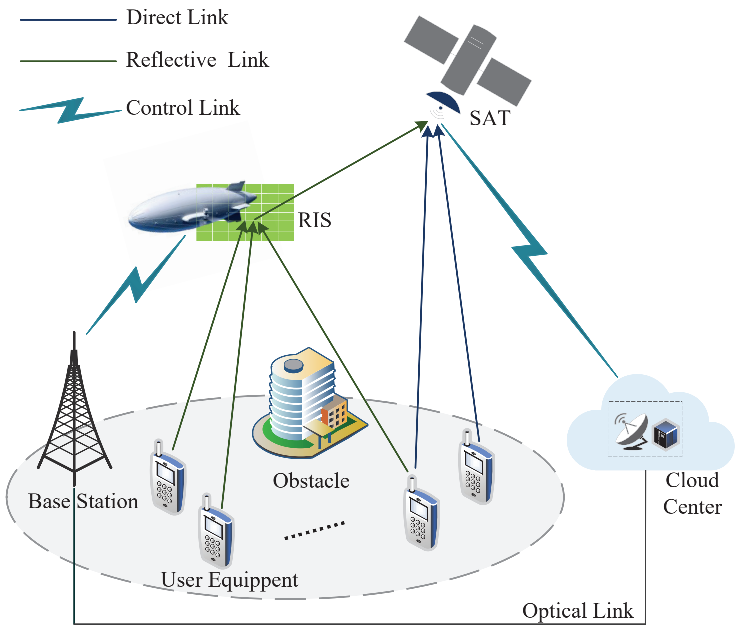

Figure 1. Illustration of a RIS-assisted IS-HAP-TNs system

In order to realistically simulate the UEs-RIS link, where the RIS is mounted on the HAP in the aerial. Here, we consider the small-scale path loss model [22], then the channel vector of the UEs-RIS can be expressed as

$$\mathbf{h}_{UR} = \sqrt{\frac{MK}{L_{total}}} \sum_{l=1}^{L_{total}} \alpha_l \mathbf{g}(m, \varphi_{AR}) \mathbf{g}^T(k, \varphi_{DU}) \tag{1}$$

where L_{total} denotes the number of the total transmission path, α_l represents the Nakagami-m channel model random variable, $\varphi_{AR,l}$ and $\varphi_{DU,l}$ denote the the angle of arrival (AoA) of RIS and the angle of departure (AoD) of the UEs in the l-th transmission path. The channel model vector $\mathbf{g}(L, \varphi)$ as a function of the transmission path L and the AOA or AOD φ can be expressed as

$$\mathbf{g}(L, \varphi) \triangleq \frac{1}{\sqrt{L}} \left[1, e^{j\pi \cos\varphi}, e^{2j\pi \cos\varphi}, \ldots e^{(L-1)j\pi \cos\varphi}\right]^T \tag{2}$$

The RIS-satellite uplink channel vector is denoted by $\mathbf{H}_{RS}$, which can expressed as

$$\mathbf{H}_{RS} = \sqrt{MN_s P_r}[\mathbf{g}(N_s, \varphi_{AS})]\mathbf{g}\left(M^T, \varphi_{DR}\right) \tag{3}$$

where the N_s denotes the antenna numbers of the uniform linear array (ULA) in the satellite, φ_{AS} and φ_{DR} are the AOA of the satellite and the AOD of the RIS, respectively. Meanwhile, the P_r is the free space path loss between the RIS and the satellite [23]. Note that in the RIS-satellite uplink channel model, considering that the HAP flies at a higher altitude than most ground buildings and the RIS is mounted on the HAP, thus we only assume the line-of-sight (LoS) transmission path between the RIS and the satellite, the P_r can be expressed as by the following formula

$$P_r = \frac{\lambda^2 G_{sr} G_{st}}{(4\pi)^2 {d_{sr}}^2 \kappa_a T_a B_W} \tag{4}$$

where λ, G_{sr}, G_{st}, d_{sr}, κ_a, T_a, B_W denote the carrier wavelength of signal, the gains of every RIS reflection unit, antenna gain of each satellite, the transmission distance between RIS to center of satellite coverage area, the Boltzmann constant, the temperature of the propagating noise and the frequency band of signal, respectively. The direct uplink channel $\mathbf{H}_{US}$ from the UE and the satellite is basically a standard MIMO channel and can be characterized by existing methods to express its channel characteristics [24].

We defined that Φ means diagonal matrix for input by $\Phi(m,m) = \phi_m = \chi_m e^{j\varphi_m}, m = 1, 2, ..., M, \chi_m \in [0,1]$, representing the magnitude and phase shift of each RIS element, respectively. It is supposed that $\mathbf{H}_{RS}$ and $\mathbf{H}_{US}$ remain unchanged in consecutive K consecutive time slots under the assumption of block attenuation of channel model, its purpose is to convert the signal $E\left[|s_k(t)|^2\right] = 1$. Next, the signal received by the satellite can then be expressed in the following

$$y_k(t) = \sum_{k=1}^{K} (\mathbf{H}_{US} + \mathbf{H}_{RS}\Phi\mathbf{h}_{UR})\mathbf{w}_k s_k + n_0 \tag{5}$$

where $\mathbf{w}_k$ means the transmit power matrix coefficient vector at the k-th UEs of the total transmit beamforming matrix $\mathbf{W}$, n_0 denotes the system noise followed by $n_0 \sim \mathcal{CN}(0, \sigma^2)$, respectively.

2.1. Problem Formulation

It can be seen from Eq.(5) that the RIS-assisted IS-HAP-TNs system dose not introduce extra noise compared with the conventional relay-assisted system. This is because the RIS does not need to decode and encode the signal, but only acts as a simple reflective device and reflects the signal incident on it. The overall system sum rate is given by the following formula

$$R_k = \log_2(1 + \gamma_k) \tag{6}$$

where γ_k is the received signal-to-interference-plus-noise ratio (SINR), which can be shown as

$$\gamma_k = \frac{|(\mathbf{H}_{US} + \mathbf{H}_{RS}\Phi\mathbf{h}_{UR})\mathbf{w}_k|^2}{\sum_{j \neq k}^{K} |(\mathbf{H}_{US} + \mathbf{H}_{RS}\Phi\mathbf{h}_{UR})\mathbf{w}_j|^2 + \sigma^2} \tag{7}$$

The system can be described that the K UEs transmits signals to satellite, so its network performance should be the sum rate of the total K UEs, which can be modeled as

$$C(\mathbf{H}_{US}, \mathbf{H}_{RS}, \Phi, \mathbf{w}, \mathbf{h}_{UR}) = \sum_{k=1}^{K} R_k \tag{8}$$

Unlike traditional deep neural networks (DNNs), which require two phases, online learning phase and offline training phase, each CSI is used to set up the states by our proposed DRL approach, and the algorithm is used to obtain a continuous two matrices through calculation. Mathematically speaking, the problem of RIS-assisted IS-HAP-TNs performance optimization design can be represented as

$$\begin{aligned} &\max_{\{\mathbf{W},\mathbf{\Phi}\}} \ C(\mathbf{H}_{US}, \mathbf{H}_{RS}, \Phi, \mathbf{w}, \mathbf{h}_{UR}) \\ &\text{s.t.}\, C_1 : tr\{\mathbf{W}\mathbf{W}^{\mathcal{H}}\} \leq P_{\max} \\ &\quad C_2 : |\phi_m| \leq 1, \forall m = 1, 2, \ldots, M. \end{aligned} \tag{9}$$

where $P_{\max}$ represents the maximum link transmission power. The constraint C_1 regulates the UEs transmission maximum power. The constraint C_2 represents the constraints on RIS reflective elements. Obviously, the above optimization problem are all based on non-convex constraints and can hardly be settled by conventional improvement approaches. If the classical mathematical tools are used, you must use exhaustive exhaustive search to get a locally optimal or sub-optimal solution, which requires a lot of computing resources or even impossible, especially for the large-scale network communication scenarios proposed in this paper.

In each iteration of traditional alternating optimization algorithms, the globally sub-optimal $\mathbf{W}$ is solved by first fixing Φ and the sub-optimal Φ is solved by fixing the matrix $\mathbf{W}$ until the algorithm converges. For the design of high-dimensional continuous variables, including the transmit power matrix, the phase shift matrix, etc., traditional DRL methods such as DQN and DDPG cannot effectively solve these problems, and often generate local optimal deviations.

3. Soft-DDPG-Based Joint Active and Passive Beamforming Design

In this section, the method of DRL is used to jointly optimize the transmit beamforming shape and phase shift array, and utilizing DDPG structure shown in Fig. 2. First, we briefly discuss the soft-DDPG principle and operation process. Then, we will introduce the proposed DNN architecture and provide a detailed description of the *state*, *action*, *reward*, and the algorithm framework.

3.1. Overview of soft-DDPG

It is supposed that there exists a central controller or a learning agent in this network that can collect channel information or communication date immediately, such as the RIS to satellite channel $\mathbf{H}_{RS}$ and $\mathbf{h}_{UR}$ and the UE to the RIS channel $\mathbf{h}_{UR}$. Fig. 2 displays the soft-DDPG architecture suitable for the earning agents to interact with high dynamic communication environments to get pre-defined rewards or punishments. The core concept of the soft-DDPG framework proposed in this letter is to perform effective beamforming design and phase shift convert under unforeseen circumstances such as local state observations such as RIS. The algorithm mainly includes two kinds of deep neural networks (DNN), namely the training network and the target network. To avoid or mitigate the issue of updating state participant values in a single case, we assume that the target and training networks have the same neural network architecture.

Based on the above extensions, we can more clearly portray the framework covered in this article, with four DNNs are drawn in detail, which are the training critic network, the training actor network, the target critic network and the target actor network. The functions of these four neural networks described above are described below. The training critic network need to input the current state $s^{(t)}$ into the action network and output the current action $a^{(t)}$, and the training actor network need to input the state $s^{(t)}$ and action $a^{(t)}$ into the training critic network and output the Q value $Q_\pi(s^{(t)}, a^{(t)})$. The target critic network need to input the updated state $s^{(t+1)}$ to the target actor network and output the $a^{(t+1)}$. The target actor network need to input the updated $s^{(t+1)}$ and $a^{(t+1)}$ to the target critic network and output the target Q value $Q_\pi(s^{(t+1)}, a^{(t+1)})$.

Considering the existence of plural inputs in the neural network input, this proposed model uses the *tanh* as the activation function of the hidden layer to limit the action space in the interval $(0, 2\pi)$, and to eliminate the effect of the change in the distribution of the hidden layer data brought by the parameter update. this proposed DRL framework introduces a batch normalization layer after each hidden layer to process its output. The batch normalization layer can effectively combat the gradient disappearance phenomenon, improve the training efficiency, and make the training process of the deep layer network more stable. In addition, according to the constraints of transmitting power and phase shift coefficients, the proposed model adds *tanh* activation function to the output layer of the actor network to restrict the output to the interval [-1,1], and subsequently transforms the action into the data format required by the optimization problem by taking absolute value normalization and range mapping methods to meet the constraints of power allocation and phase shift, so as to calculate the system sum rate as the Eq. (8).

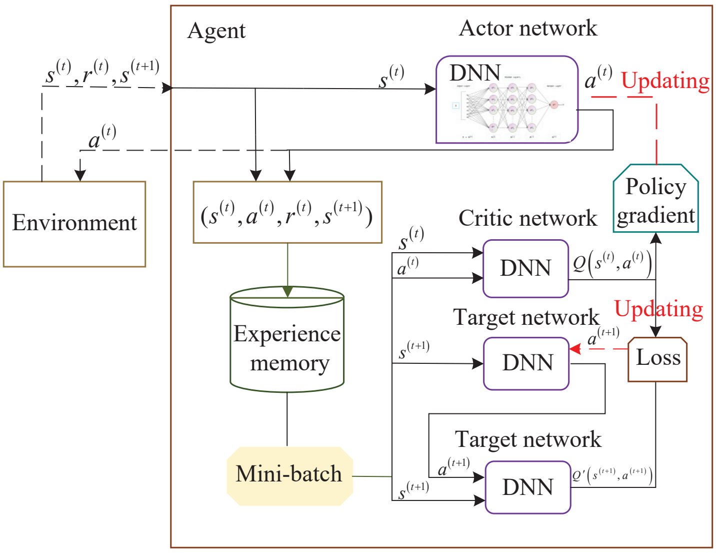

Figure 2. The DRL-based active transmit matrix and phase shift design framework using DDPG.

We generate different transmission link channel information by following the channel model features described earlier when channel state information (CSI) and the previous action $\mathbf{W}^{(t-1)}$ and $\Phi^{(t-1)}$ are known at t-th time step, and the leaning agent can establish the knowledge about the current state space $s^{(t)}$ in the t-th time step. It is considered that the difficulty of joint optimal design of active transmission beamforming and passive RIS phase shift matrix are discrete and presents a great challenge to continuous state space and action space settings. Next, the detail of DRL-based algorithm state space S, action space A and the instant reward function R are explained below.

State: State space is generally a description of the environmental observations at t-th time step. In this paper, the DRL algorithm state space includes three parts, i.e., the last time action space, the satellite-RIS channel $\mathbf{H}_{RS}$ and the RIS-UEs channel $\mathbf{h}_{UR}$. Next, the state of the t-th state space is defined as

$$s^{(t)} = \left[a^{(t-1)}, \mathbf{H}^{(t-1)}, \mathbf{h}_1^{(t-1)}, ..., \mathbf{h}_K^{(t-1)}\right] \tag{10}$$

Action: Action space is generally a series of choices for the next action. Once the agent performs the current action $a^{(t)}$ step by step during the learning process according to the transfer policy π at the t-th time slot, the state space of the environment will be shifted from $s^{(t)}$ to the next state $s^{(t+1)}$. The actions $a^{(t)}$ are mainly related to the two variables $\mathbf{W}$ and phase-shift matrix Φ to be optimized. Since Φ is a complex vector, to simplify the action dimension, this paper takes its phase part. Thus, the action space is modelled as

$$a^{(t)} = \left[\mathbf{w}_1{}^{(\mathrm{t})}, \ldots, \mathbf{w}_K^{(t)}, \phi_1^{(t)}, \ldots, \phi_M^{(t)}\right] \tag{11}$$

And, in the action space, to ensure that the neural network inputs are real numbers and match the neural network input formats, the variables to be optimized need to divide in real part and imaginary part, thus we define $\mathbf{W} = \mathrm{Re}\{\mathbf{W}\} + \mathrm{Im}\{\mathbf{W}\}$ and $\mathbf{\Phi} = \mathrm{Re}\{\mathbf{\Phi}\} + \mathrm{Im}\{\mathbf{\Phi}\}$.

Reward: The purpose of this paper is to maximize the system sum rate and Eq. (10) is adopted as the reward function:

$$r^{(t)} = C(\mathbf{H}_{US}, \mathbf{H}_{RS}, \Phi, \mathbf{w}, \mathbf{h}_{UR}) \tag{12}$$

Algorithm 1: Soft-DDPG-based Algorithm

```
Initialize experience memory $D$ to empty;
Randomly initialization generate actor target/train network $\psi'(\cdot)$ and critic target/train network $Q'(\cdot)$ with parameters $\xi_a'$ and $\xi_c'$, separately;
**Input**: $\mathbf{w}$, $\phi$, $\mathbf{H}_{RS}$ and $\mathbf{h}_{UR}$;
**Output**: Optimal action $a_{opt}^{(t)}$;
**for** each episode **do**:
    Initialize state $s^{(0)} \in S, S \leftarrow s^{(0)}$;
    **for** $t = 0, 1, 2, ...T-1$ **do**:
        Choose action $a^{(t)} = \pi\left(s^{(t)} \mid \theta^{\pi}\right) + \mathcal{N}$;
        Take action $a^{(t)}$, get reward $r^{(t)}$ and $s^{(t)}$ evolves into new state $s^{(t+1)}$;
        Save $(s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)})$ into $D$;
        Randomly sample $\zeta$ transitions form $D$;
        Traning framework via DNN;
        Compute target value for the critic's evaluation network by
        $y^{(i)} = r^{(i)} + \gamma Q'_{\pi'}\left(s^{(i+1)}, \pi'\left(s^{(i+1)}|\theta^{\pi'}\right)|\theta^{Q'}\right)$
        Update the parameters of the critic's evaluation network by
        $L(\theta^{Q}) = \frac{1}{\zeta}\sum_{i=1}^{\zeta}\left(y^{(i)} - Q_{\pi}\left(s^{(i)}, a^{(i)} \mid \theta^{Q}\right)\right)^{2}$;
        Update the parameters of actor network with sampled policy gradients by
        $\nabla_{\theta^{\pi}} J = \frac{1}{\zeta}\sum_{i=1}^{\zeta}\nabla_{a} Q_{\pi}(s, a|\theta^{Q})\Big|_{a=\pi\left(s^{(i)}|\theta^{\pi}\right)} \nabla_{\theta^{\pi}}\pi(s|\theta^{\pi})$;
        Soft-update the parameters of DDPG's target networks by
        $\theta_{c}^{(target)} \leftarrow \tau_{c}\theta_{c}^{(train)} + (1-\tau_{c})\theta_{c}^{(target)}$
        $\theta_{a}^{(target)} \leftarrow \tau_{a}\theta_{a}^{(train)} + (1-\tau_{a})\theta_{a}^{(target)}$;
        Update the state $s^{(t+1)}$;
    **end for**;
**end for**;
```

3.2. The Process of Algorithm Training

In order to break the coupling between experiences and adapt to a high dynamic environment, the experimencre replay approach allows agent access to previous historical experiences in subsequent training, the DDPG framework considered in this article.

For policy-class-based algorithms, the agent collects experience in episode. After run an episode, then lose your experience. Better with a multi-threaded parallel architecture. This not only solves the previous problems, but also makes efficient use of computing resources and improves training efficiency.

In the proposed DDPG framework, the entire agent consists of a global network and multiple parallel independent workers, each including a set of actor network and critic network. Each worker interacts independently with their own environment, gaining independent sampling experiences that are independent of each other, thus breaking the coupling between experiences to match the experience replay. Most of the underlying algorithms in DRL are single-threaded, that is, a learning agent that interacts with the environment to generate experience. Including the underlying version of actor network and critic network, because the environment is fixed and the action of the agent needs to be continuous, the experience gathered has strong timing associations and only part of the state and action space can be explored in a limited amount of time. To solve this problem, we adopt the DDPG scheme to optimize the design process and present the corresponding pseudo-code in Algorithm 1.

In the initial stage of the algorithm, the experience replay buffer D, the training actor network $\psi(\cdot)$ and training critic network $Q(\cdot)$ need to be initialized randomly(Line 1-2). They are copied to the target network $\psi'(\cdot)$ and $Q'(\cdot)$ (Line 3). After initializing and randomly generating the RIS-assisted IS-HAP-TNs communication channel environment state, the state is processed via DNN and the output Y_t (Line 5-6). The action is derived based on Y_t, where $\mathcal{N}$ is denoted as random noise, with the aim of seeking efficient exploration (Line 7-8). In this letter, we employ the mini batch to reduce the sample training amount of sampling and ensure the quality of gradient reduction. After the transformation sequence is saved in the memory replay buffer D (line 10), to achieve the optimal action that maximizes the output of the critic train network, the two train networks are updated using the minibatches of size ζ randomly sampled from replay buffer D (line 11). Update the critic target network parameters $Q(\cdot)$ by minimizing the variance loss (line 14). Make use of linking rules to update the actor networks parameters $\psi(\cdot)$. Finally, the target networks parameters of actor network and critic network are slowly soft updated using the control factor τ as the decaying rate(line 15).

During each iteration t each of the learning process, the actor train network will select the action from the continuous action space based on the current state $s^{(t)}$. During this training process, in order to effectively explore the optimal action, the stochastic noise $\mathcal{N}_a$ is also taken into account in the algorithm framework to obtain the deterministic strategy, i.e., $a^{(t)}=\pi\left(s^{(t)}\mid\theta^{\pi}\right)+\mathcal{N}_a$, where θ^{π} is the actor train network parameter, and π is the transfer policy. When the operation ends, the environment will transit the last action to the next state $s^{(t+1)}$ to obtain instant reward $r^{(t)}$, and then get an evaluation for the action to evaluate the optimal action $a^{(t)}$. Modeling a state-action value function by parameterized by θ^{Q} as

$$\begin{aligned} Q_{\pi}\left(s^{(t)},a^{(t)}\mid\theta^{Q}\right) &\leftarrow \alpha Q_{\pi}\left(s^{(t)},a^{(t)}\mid\theta^{Q}\right) \\ &+(1-\alpha)\left[r^{(t)}+\gamma\max_{a'}Q_{\pi}\left(s^{(t+1)},a'\mid\theta^{Q}\right)\Big|_{a'=\pi\left(s^{(t+1)}\mid\theta^{\pi}\right)}\right] \end{aligned} \tag{13}$$

where α denotes the algorithm learning rate in this algorithm framework. To ensure the stability, the target actor network parameterized by $\theta^{\pi'}$ and the target critic network characterized by $\theta^{Q'}$, which is parameterized at intervals according to online network parameters. Thus, considering the parameter update strategy is a soft update method, so the algorithm is called soft-DDPG. The soft update method of parameters ensures the slow update of parameters and alleviates the instability problem of the policy network during the learning process.

doi:10.20944/preprints202212.0038.v1

4. Numerical Simulation Results

In this section, we will evaluate the performance improvement of the proposed DRL-based algorithm framework for the proposed system model from different perspectives.. First, we will randomly generate channel model matrix $\mathbf{H}_{RS}$ and $\mathbf{h}_{UR}$ following shadowed-Rician fading distribution and Rayleigh distribution, respectively [25]. The system parameters and hyperparameters of the DDPG algorithms are listed in Table 1 [26]. To test whether our algorithm improves network performance, we will also consider three other standard solutions:

1) Hard-DDPG: The scheme indicates that the parameters in the DRL framework are updated in a hard-update strategy which that allows the network to copy all the parameters in the network at this time directly into the target network after every t_u training sessions by pre-setting the parameter update interval t_u .

2) Random RIS: The scheme denotes that the RIS phase shift matrix Φ is randomly generate.

3) Without RIS: This scheme denotes that the communication scenario without RIS and the UEs can send the signals directly to the satellite. Considering that the process is a continuous transmission, thus we assume that the successful transmission signal is $1/2$.

Fig. 3 plots the relationship between the number of RIS reflection elements and the system sum rate. As can be seen from the figure, the system sum rate is significantly higher for all algorithms as the number of RIS elements increases, due to the fact that more RIS reflection elements increase the reflection channel gain, but also sacrifices the complexity of the RIS deployment at the HAP. In addition, we can observe that the soft-update parameter strategy obtains a higher system sum rate than the hard update parameter strategy, alleviates the instability of the Q-value network in the learning process, and the soft-update strategy obtains a higher system ensemble rate by more flexible interacting with the environment to design the phase shift matrix more flexibly.

The setting of hyper-parameters will have a great impact on the performance, like the stability and convergence speed of neural networks. This paper also explores the effect of different learning rates on the performance and convergence speed of the model in our proposed DRL framework. The average reward is used to measure its performance, which can be shown as

$$\text{average\_reward}\left(T_i\right) = \frac{\sum\limits_{t=1}^{T} r^{(t)}}{T_i}, T_i = 1,2,\ldots,T \tag{14}$$

where T is the maximum step size of sample training. Fig. 4 shows the average reward versus time step under different learning rates, and it can be seen that the effect of different learning rate settings in the neural network on the performance of the DRL algorithm varies greatly. In particularly, the considered DRL framework with a learning rate of 0.001 performs the best, but converges more slowly than the others. As the RIS reflection element increases, the average system reward also increases gradually as expected with the addition of reflection channels, but this does not significantly increase the convergence time of the proposed DRL framework.

Fig. 5 shows a schematic comparison of the average reward performance and the outdated CSI coefficients, respectively. In the proposed DRL framework, we choose the last moment CSI as the state space input, and we can see that the average reward of all algorithms decreases gradually as the outdated CSI coefficient decreases. However, the proposed soft-DDPG framework remains at a favorable level compared to the existing scheme and the hard-DDPG scheme. Compared with the advanced DRL schemes, which do not require an exact channel model information, the existing alternating parameter optimization scheme relies on the knowledge of static exact channel model, but because of the high dynamic communication scenario, the system performance is not as good as soft DDPG and hard DDPG scheme.

Table 1: System and DNN Parameters

System parameters	**Value**
Frequency band	$f = 2$ GHz
Wavelength	$\lambda = 150$ mm
Noise power spectral density	-169 dBm/Hz
Link bandwidth	$W = 15$ MHz
Noise temperature	$T = 300K$
Height of HAP	20km
Number of the UEs	K=3
Transmission path	L=3
DNN hyperparameters in DDPG	**Value**
Reward discount rate	0.99
Numbers of experiences with the mini-batch	16
Learning rate	0.0001
Decaying rate	0.0001
Experience replay buffer size	100000
Numbers of steps in each training episode	10000

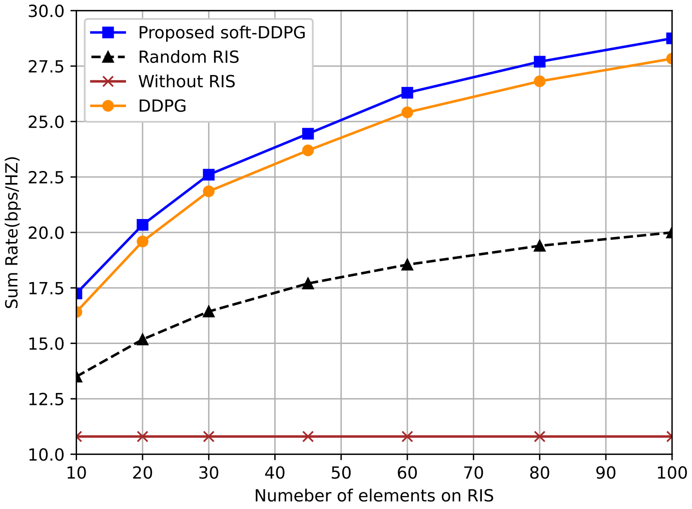

Figure 3. Sum rate performance relative to the increasing number of elements on RIS.

5. Conclusion

This paper discussed the joint optimal design scheme of transmitting active beamforming and passive beamforming for maximizing system sum rate. In the IS-HAP-TNs assisted by RIS, it is hard to sense the channel state information in the dynamic environment accurately and comprehensively. On this basis, a novel type of DRL architecture, namely soft-DDPG algorithm. With the help of the network parameter soft-update strategy, the coordination of the phase shift matrix can be obtained even when the increasing number of RIS reflective elements amplitude changes. Simulation results show that the proposed framework can achieve better network performance in a lower operation duration and can be applied to the real-time control of IS-HAP-TNs system.

Acknowledgments: This work was supported by the National Science Natural Foundation of China (No.61901502, No.61971474 and No.62001517), the National Postdoctoral Program for Innovative Talents (No.BX20200101), the Research Project of Science and Technology on Complex Electronic System Simulation Laboratory (No.DXZT-JC-ZZ-2019-005), and in part by Beijing Nova Program (No.Z201100006820121) and the Project of Space Engineering University (No.2020XXAQ01 and No.2019XXAQ05)

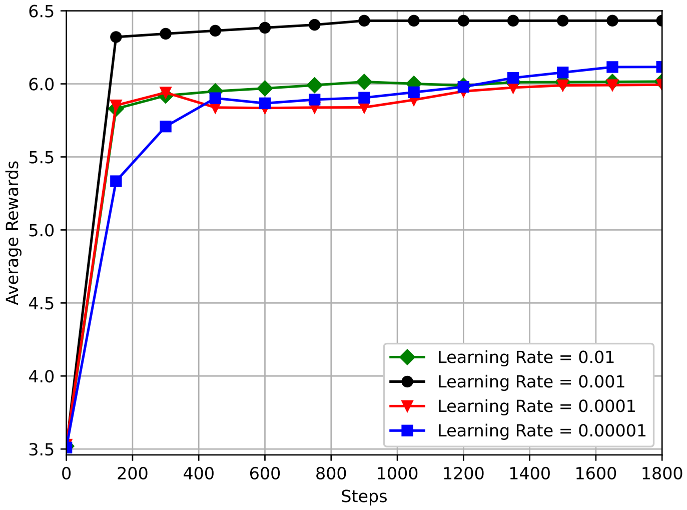

Figure 4. Variation of average reward under different learning rate.

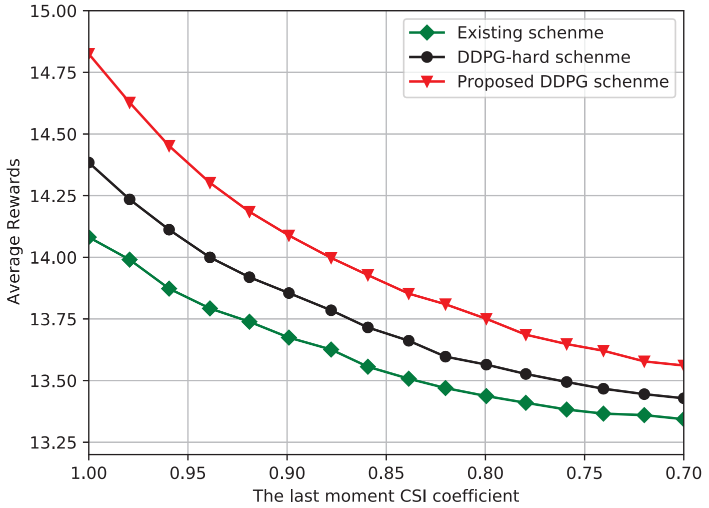

Figure 5. Average reward against the last moment CSI coefficient.

Author Contributions: Min Wu, Shibing Zhu, Changqing Li, Yudi. Chen, Feng Zhou and Xiang Su conceived and designed the experiments; Min Wu performed the experiments; Shibing Zhu, Changqing Li and Yudi. Chen analyzed the data; Feng Zhou and Xiang Su contributed analysis tools; Min Wu, Shibing Zhu and Feng Zhou wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. K. An, M. Lin, J. Ouyang, and W. P. Zhu, "Secure transmission in cognitive satellite terrestrial networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 3025-3037, Nov. 2016.
2. Lin, Z., Lin, M., Cola, T. de, Wang, J.-B., Zhu, W.-P., Cheng, J. "Supporting IoT with rate-splitting multiple access in satellite and aerial integrated networks," *IEEE Internet Things J.*, **2021**, *8*, 11123-11134.
3. Liu, R., Guo, K., An, K., Zhu, S., Shuai, H. "NOMA-based integrated satellite-terrestrial relay networks under spectrum sharing environment," *IEEE Wireless Commun. Lett.*, **2021**, *10*, 1266-1270.

4. L. Yang, P. Li, Y. Yang, S. Li, I. Trigui and R. Ma, "Performance Analysis of RIS-Aided Networks With Co-Channel Interference," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 49-53, Jan.
5. H. Luo, L. Lv, Q. Wu, Z. Ding, N. Al-Dhahir and J. Chen, "Beamforming Design for Active IOS Aided NOMA Networks," *IEEE Wireless Commun. Lett.*, 2022.
6. H. Niu, Z. Chu, F. Zhou, et al., "Robust design for intelligent reflecting surface-assisted secrecy SWIPT network," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4133-4149, June. 2022.
7. S. Gong, X. Lu, D. T. Hoang, et al., "Towards smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surv. Tut.*, pp. 1-33, Jun. 2020.
8. X. Li et al., "Hardware impaired ambient backscatter NOMA systems: Reliability and security," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2723-2736, April 2021.
9. K. Guo, K. An, et al., "Physical layer security for multiuser satellite communication systems with threshold-based scheduling scheme," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5129-5141, May. 2020.
10. J. Wang, Y .-C. Liang, et al., "Robust beamforming and phase shift design for IRS-enhanced multi-user MISO downlink communication," *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, pp. 1-6, Jun. 2020.
11. K. B. Letaief, W. Chen, et al., "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84-90, Aug. 2019.
12. X. Yue and Y. Liu, "Performance Analysis of Intelligent Reflecting Surface Assisted NOMA Networks," *IEEE Trans. Wireless Commun.* , vol. 21, no. 4, pp. 2623-2636, April 2022.
13. Q.-U.-U. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," 2019, *arXiv:1903.08127.* [Online]. Available: http://arxiv.org/abs/1903.08127.
14. M. Fozi, A. R. Sharafat and M. Bennis, "Fast MIMO beamforming via deep reinforcement learning for high mobility mmWave connectivity," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 127-142, Jan. 2022.
15. Y. Li, W. Zhang, C. -X. Wang, et al., "Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks," *IEEE Trans. Cognitive Commum. Net.*, vol. 6, no. 2, pp. 464-475, June. 2020.
16. H. Ren, C. Pan, L. Wang, et al., "Long-term CSI-based design for RIS-aided multiuser MISO systems exploiting deep reinforcement learning," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 567-571, March. 2022.
17. M. Zhang, S. Fu and Q. Fan, "Joint 3D deployment and power allocation for UAV-BS: A deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2309-2312, Oct. 2021.
18. F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581-1592, Mar. 2020.
19. R. Shafin, M. Jiang, S. Ma, L. Piazzi and L. Liu, "Joint Parametric Channel Estimation and Performance Characterization for 3D Massive MIMO OFDM Systems," *2018 IEEE International Conference Commun. (ICC)*, 2018, pp. 1-6.
20. C. Huang, R. Mo and C. Yuen, "Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839-1850, Aug. 2020.
21. Liang Yang, Qi Zhu, Sai Li, Imran Shafique Ansari, and Siyuan Yu, "On the Performance of Mixed FSO-UWOC Dual-Hop Transmission Systems," *IEEE Wireless Commun. Lett.*, vol.10, no.9, pp.2041-2045, Sep.2021.
22. Yang, F. Meng, Q. Wu, D. B. da Costa and M. -S. Alouini, "Accurate Closed-Form Approximations to Channel Distributions of RIS-Aided Wireless Systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1985-1989, Nov. 2020.
23. J. He, M. Leinonen, H. Wymeersch, et al., "Channel estimation for RIS-aided mmWave MIMO systems," *in Proc. 2020 IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, pp. 1–6, Dec. 2020.
24. Yan, X.; Xiao, H.; An, K.; Zhen, G.; Chatzainotas, S. "Ergodic capacity of NOMA-based uplink satellite networks with randomly deployed users," *IEEE Syst. J.*, 2020, *14*, 3343-3350.
25. Bletsas, A.; Shin, H.; Win, M.Z. "Cooperative communication with outage-optimal opportunistic relaying," *IEEE Trans. Wireless Commun.*, 2007, *6*, 3450-3460.

26. K. Guo, M. Lin, J.-B. Wang, et al., "On he performance of LMS communication with hardware impairments and interference," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1490-1505, Feb. 2019.