Deep Reinforcement Learning-Assisted NOMA Age-Optimal Power Allocation for S-IoT Network

Qingxi Liu[1], Jian Jiao*[,1,2], Shaohua Wu[1,2], Rongxing Lu[3], and Qinyu Zhang[1,2]
[1]Communication Engineering Research Centre, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
[2]Peng Cheng Laboratory, Shenzhen 518055, China
[3]University of New Brunswick, Fredericton, NB E3B 5A3, Canada
{20s152067@stu.hit.edu.cn, {jiaojian, hitwush, zqy}@hit.edu.cn, rlu1@unb.ca}

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9839197

***Abstract*—In this paper, we consider a satellite-based Internet of Things (S-IoT) network under shadowed-Rician fading channels, where a satellite transmits timely status updates to multiple user equipments (UEs) with non-orthogonal multiple access (NOMA). In each transmission, the satellite needs to allocate limited power to the status updates for UEs in an appropriate way to guarantee the freshness of updates, characterized by age of information (AoI). To minimize the average AoI of S-IoT network, we formulate a power-constrained optimization problem and then reformulate it as a Markov decision process (MDP). Considering the non-convexity of the optimization problem and the high dimensionality of the multiuser MDP with large state and action spaces, we propose a deep reinforcement learning-assisted age-optimal power allocation (DRAP) scheme to solve the problem and obtain an optimal power allocation policy. Furthermore, a double-network deep reinforcement learning structure is designed to enhance the training effectiveness for our optimization problem. Finally, simulation results show that our proposed DRAP scheme outperforms the benchmark schemes.**

***Index Terms*—Satellite-based Internet of Things, non-orthogonal multiple access, age of information, deep reinforcement learning, power allocation**

I. Introduction

Recently, the rapidly developing high-throughput satellite has been viewed as a crucial technology to meet the demands of high-throughput and global coverage applications in Internet of Things (IoT) [1]. As a result, satellite-based Internet of Things (S-IoT) will play an important role in beyond 5G (B5G) and 6G wireless networks. Furthermore, the information freshness is becoming an increasingly significant performance metric in various S-IoT scenarios, such as environment monitoring, smart city, self-driving cars, etc [2], [3]. In these applications, the latest status updates need to be transmitted to receivers within certain deadlines, since outdated information is of no use and may even lead to catastrophic results. To measure the information freshness, age of information (AoI) has been proposed, which is defined as the elapsed time from the generation of status to the present [4].

How to optimize the AoI and improve the timeliness of networks has received widespread attentions. The works in [5]–[7] investigate the average AoI from the scheduling perspective. In [5], Gong *et al.* design a dynamic programming (DP) algorithm and a low-complexity myopic policy to obtain the optimal scheduling policy for a multiuser uplink network. Hsu *et al.* [6] consider a wireless broadcast network, where a base-station (BS) timely transmits information to many network users under a transmission capacity constraint, and the BS can serve at most one user for each transmission. They design a structural scheduling algorithm and an index scheduling algorithm to minimize the average AoI. In [7], Bedewy *et al.* show that the maximum age first (MAF) scheduler and the zero-wait sampler are jointly optimized for the data freshness in multi-source system.

The previously mentioned works focus on the orthogonal multiple access (OMA) scheme. To enhance spectral efficiency and user fairness, non-orthogonal multiple access (NOMA) is investigated on AoI recently [8]. In [9], Maatouk *et al.* compare the average AoI performance between OMA and NOMA schemes, where NOMA scheme can achieve better spectral efficiency and result in a lower average AoI under some simulation setups. Moreover, AoI optimization for hybrid NOMA/OMA systems is investigated in [10], [11]. Guo *et al.* [10] propose an adaptive NOMA/OMA power allocation policy based on Lyapunov Optimization. Wang *et al.* [11] prove the existence of optimal stationary and deterministic policy in the hybrid NOMA/OMA system, and obtain an optimal policy that decides when the BS uses NOMA or OMA to minimize average AoI for a downlink network with two users. However, the applicable scenarios in most of these existing schemes are limited due to high computational complexity, especially when the number of users increases.

In this paper, we aim to minimize average AoI in multiuser NOMA S-IoT network subject to power constraints. It is worth noting that power allocation problems for AoI minimization in multiuser NOMA system are more challenging than scheduling problems, because of the larger state and action space in power allocation problems. To our best knowledge, this is the first work to design a deep reinforcement learning-

*Jian Jiao is the corresponding author. This work was supported in part by the National Natural Sciences Foundation of China (NSFC) under Grant 62071141, Grant 61871147, Grant 61831008, and Grant 62027802, in part by the Shenzhen Basic Research Program under Grant GXWD20201230155427003-20200822165138001, in part by the Natural Science Foundation of Guangdong Province under Grant 2020A1515010505, in part by the Guangdong Science and Technology Planning Project under Grant 2018B030322004, and in part by the Major Key Project of PCL under Grant PCL2021A03-1.

assisted age-optimal power allocation (DRAP) scheme in NOMA S-IoT network. Specifically, contributions of our work are summarized in the following.

- We formulate the age-optimal power allocation optimization problem for NOMA S-IoT network subject to the power constraints inherited from satellite. To minimize the average AoI, we reformulate the optimization problem as a Markov Decision Process (MDP) based on the instantaneous AoI of all user equipments (UEs).
- Due to the non-convexity of the optimization problem and the curse of dimensionality of MDP, we propose a DRAP algorithm to obtain the optimal power allocation policy. We delicately design a double-network deep reinforcement learning (DRL) architecture for our optimization problem and give training details.
- Simulation results show that our proposed DRAP scheme outperforms the benchmark schemes in terms of average AoI performance.

The rest of this paper is organized as follows. Section II describes the NOMA S-IoT network and presents the problem formulation. Section III reformulates the optimization problem as a MDP problem. In section IV, we develop a DRAP algorithm to obtain the power allocation policy. In section V, we train the DRAP algorithm and compare our scheme to other two existing schemes. Conclusion is drew in section VI.

II. System Model and Problem Formulation

A. System Model

We consider a NOMA downlink S-IoT network which contains a satellite and multiple UEs. The satellite communicates with UEs in a hybrid access mode within its multi-beam coverage, that is, OMA manner is adopted for different beams, and NOMA scheme is used to serve multiple UEs in the same beam. In this way, interference between adjacent beams can be prevented while ensuring spectrum efficiency, and we focus on the single-beam transmission process in the following. Moreover, we assume that there are K UEs in the coverage of a beam and the status updates for different UEs are distinguished by using different powers. We denote the set of UEs with $\mathcal{U} = \{\mathrm{UE}_1, \mathrm{UE}_2, ..., \mathrm{UE}_K\}$. Time is divided into slots of equal durations denoted by t, and the total number of time slots is T. Without loss of generality, we assume that the distances from satellite to K UEs in a spot beam are the same, and the duration of time slot equals to the propagation delay from satellite to UEs. Moreover, K UEs are stationary during each slot, which leads the Doppler shifts caused by the mobility of satellite for different UEs are identical [12]. Thus, we can relieve the influence of Doppler shifts by setting the guard bandwidth in satellite to twice as much as the Doppler shifts [13].

Unlike terrestrial wireless communication, channels between the satellite and ground UEs are not simply the line of sight loss (LoS) but also suffered masking effects of obstacles and obscurations. Hence, we assume that the considered satellite links follow the independent and identically distributed (i.i.d.) shadowed-Rician fading, and the probability density function (PDF) of channel gain $|h_i|^2$ can be described as

$$f_{|h_i|^2}(x) = \left(\frac{2b_0 m}{2b_0 m+\Omega}\right)^m \frac{1}{2b_0} \exp\left(-\frac{x}{2b_0}\right) \cdot {}_1F_1\left(m, 1, \frac{\Omega x}{2b_0(2b_0 m+\Omega)}\right), \tag{1}$$

where $2b_0$ is the average power of the scatter component, m is the Nakagami-m parameter, Ω is the average power of the LOS component, and ${}_1F_1(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function [14].

Let $s_i(t)$ denote the status update transmitted to UE_i in time slot t. The superimposed signal at the satellite can be written as

$$S(t) = \sum_{i=1}^{K} \sqrt{\alpha_i(t) P} \cdot s_i(t), \tag{2}$$

where $\alpha_i(t)$ is the power coefficient allocated by the satellite to UE_i, and P is total transmission power. We have $\sum_{i=1}^{K} \alpha_i = 1$ to achieve better performance as much as possible. The signal received by the UE_i can be given by

$$r_i(t) = g_i(t)\ S(t) + n_i(t), \tag{3}$$

where $g_i = h_i(t)\sqrt{G_S G_{UE} L_F}$ is joint channel gain. G_S and G_{UE} denote satellite beam gain and antenna gain respectively, L_F denotes the free space loss, and $n_i(t)$ is additive white Gaussian noise (AWGN) with variance σ^2.

B. SIC Decoding

Without loss of generality, we assume that the channel gain of each UE satisfies $|g_i|^2 > |g_{i+1}|^2$, for $i \in \{1, 2, 3, ..., K\}$. According to the principle of NOMA, we have $\alpha_i < \alpha_{i+1}$, $i \in \{1, 2, 3, ..., K\}$. The decoding process follows the descending order of the power coefficients of UEs. Specifically, UEs utilize the successive interference cancelation (SIC) to recover $s_k(t)$ from the received signals $r_k(t)$ by treating the unrecovered status updates as interferences. If the decoding is successful, the influence of $s_k(t)$ will be eliminated, and then decode the next highest power status update $s_{k-1}(t)$ until $s_1(t)$ is recovered at UE_1. In the decoding process, the signal-to-interference-and-noise-ratio (SINR) of the UE_i can be written as

$$\rho_i(\alpha) = \frac{\alpha_i P |g_i|^2}{\sum_{j=1}^{i-1} \alpha_j P |g_i|^2 + \beta \sum_{j=i+1}^{K} \alpha_j P |g_i|^2 + \sigma^2}, \tag{4}$$

where $\left(\sum_{j=1}^{i-1} \alpha_j P |g_i|^2\right)$ represents the interferences of the unrecovered status updates which have lower power than UE_i's signal. The second interference $\left(\beta \sum_{j=i+1}^{K} \alpha_j P |g_i|^2\right)$ comes from residual component of SIC decoded UEs, and β is called the error-propagation factor whose value is ranged from 0 to 1 [15]. Specifically, $\beta = 0$ means perfect SIC and $\beta = 1$ means SIC is failed.

Let B denote bandwidth of the channel, then the data rate of UE_i can be expressed by $R_i(\alpha) = B \log_2(1 + \rho_i(\alpha))$. The target rate of UE_i is denoted as R_i^{target}, and transmission fails when R_i^{target} is greater than $R_i(\alpha)$. Therefore, the outage probability of UE$_i$ is

$$P_i(\alpha) = 1 - P\left(R_i(\alpha) \geq R_i^{\text{target}}\right). \tag{5}$$

C. Problem Formulation

We adopt AoI to measure the timeliness of status updates received at UEs. Let $\Delta_i(t)$ denote the instantaneous AoI of UE$_i$ at time slot t. According to the definition of AoI, $\Delta_i(t)$ decreases to 1 when UE$_i$ receives a status update successfully at time slot t. $\Delta_i(t)$ is increased by 1 if an outage occurs at UE$_i$, since the status update at UE$_i$ is one slot older. Hence, we have

$$\Delta_i(t+1) = \begin{cases} 1, & u_i(t) = 1, \\ \Delta_i(t) + 1, & u_i(t) = 0, \end{cases} \tag{6}$$

where $u_i(t) = 1$ indicates the transmission is successful and the decoding is correct, and $u_i(t) = 0$ otherwise. Note that we assume the status updates always generate at the beginning of each time slot and are buffered at separate last-come-first-served (LCFS) queues for all the UEs. Hence, the satellite always has status updates to transmit.

We use the expected weighted sum AoI (EWSAoI) of all the UEs to measure the AoI performance of the S-IoT system, which is given by

$$\bar{\Delta} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{T} w_i \Delta_i(t)\right], \tag{7}$$

where w_i is a positive weight of UE denoting the relative importance of UEs' application. This setting has practical significance because different UEs have diverse requirements for timeliness of information. For instance, earthquake prediction system cares more about timeliness than parking system, and we should assign a higher weight for it. For simplicity, we assume that the weights of all UEs are the same, i.e., $w_1 = w_2 = ... = w_i = 1$, and it can be readily extended to the cases with different weights.

We aim to find an optimal power allocation policy π that minimizes the EWSAoI, which can be formulated as

$$\min_{\boldsymbol{\pi}} \bar{\Delta} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{T} w_i \Delta_i(t)\right], \tag{8a}$$

$$\text{s.t. } C1: \alpha_i(t) < \alpha_j(t), \text{for } j > i \text{ and } i, j \in \{1, 2, ..., K\}, \tag{8b}$$

$$C2: \sum_{i=1}^{K} \alpha_i(t) = 1, \tag{8c}$$

$$C3: \alpha_i(t) \geq 0, \forall i \in \{1, 2, ..., K\}, \tag{8d}$$

where $C1$ indicates that the UE with weaker channel gain should be allocated more power to meet the requirements of SIC decoding. $C2$ and $C3$ guarantee that the power allocation coefficients should be non-negative values, and the sum of them is equal to 1.

III. MDP Formulation

MDP is suitable for sequential decision-making problems, which is consistent with our age-optimal power allocation problem. We reformulate the optimization problem as a MDP to obtain the optimal solution for the NOMA downlink S-IoT system, which is described as follows:

State: We denote the state at time slot t by $s_t \triangleq (\Delta_1(t), \Delta_2(t), \ldots, \Delta_K(t))$, where $\Delta_i(t) \in \{1, 2, ..., T\}$ is the instantaneous AoI of UE$_i$. The set of all possible states is denoted as $\mathcal{S}$. In order to have knowledge of AoI at UEs, satellite receives a binary observation (i.e. ACK signal) to indicate whether the message is successfully transmitted or not.

Action: The action of MDP is defined as power allocation coefficients for UEs at time slot t and denoted by $a_t \triangleq (\alpha_1(t), \alpha_2(t), \ldots, \alpha_K(t))$. The coefficients of UEs are considered to only take value from a discrete set, that is, $\alpha_i(t) \in \{0, \frac{1}{M}, \frac{2}{M}, ..., 1\}$, where M is the number of power levels. All possible combinations of power allocation coefficients form an action space $\mathcal{A}$.

Transition probability: We use $\Pr(s_{t+1}|s_t, a_t)$ to denote the probability of transforming state s_t to s_{t+1} when taking action a_t. Taking two UEs as an example, the state transition probabilities can be expressed as follows:

$$\begin{aligned} &\Pr\left((1, \Delta_2(t)+1) \mid (\Delta_1(t), \Delta_2(t))\right) = (1 - P_1)\, P_2, \\ &\Pr\left((\Delta_1(t)+1, 1) \mid (\Delta_1(t), \Delta_2(t))\right) = P_1\, (1 - P_2), \\ &\Pr\left((1, 1) \mid (\Delta_1(t), \Delta_2(t))\right) = (1 - P_1)\, (1 - P_2), \\ &\Pr\left((\Delta_1(t)+1, \Delta_2(t)+1) \mid (\Delta_1(t), \Delta_2(t))\right) = P_1 P_2, \end{aligned} \tag{9}$$

where P_1 and P_2 are outage probabilities of UE$_1$ and UE$_2$, which can be obtained by Eq. (5). It is worth noting that we consider a multiuser system in this paper, and the actual state transition probability matrix will be very large. Nevertheless, we adopt a model-free DRL method to solve the problem, which can bypass this trouble, because DRL will learn strategies from iterative transmissions without knowing the exact state transition probability.

Reward: Taking action a_t in state s_t will get a reward r_{t+1}, which can be described by reward function $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Specifically, the reward function of our MDP model is defined as

$$r_{t+1}(s_t, a_t) = \frac{1}{K}\sum_{i=1}^{K}\left(\Delta_i(t) - \Delta_i(t+1)\right). \tag{10}$$

The above expression shows that the reward obtained by taking a_t in state s_t is positively correlated with AoI variation, that is, the more system average AoI decreases, the more reward will be obtained in this transmission.

Bellman equation: To reflect the long-term benefits of the system, we define return as the sum of discounted future rewards $G_t = \sum_{j=t}^{T} \gamma^{j-t} r_{j+1}$, where $\gamma \in [0, 1]$ is a discounted factor. Taking different actions in a certain state will result in different rewards. Hence, we further define the action-value function as the expected return when taking an action a_t based on policy π in state s_t, which can be described by

$$Q^{\pi}\left(s_t, a_t\right) = \mathbb{E}_{a_{l>t}\sim\pi}\left[G_t \mid s_t, a_t\right]. \tag{11}$$

According to the definition of G_t, Eq. (11) can be expressed in recursive form:

$$\begin{aligned} Q^{\pi}\left(s_t, a_t\right) &= \mathbb{E}_{a_{l>t}\sim\pi}\left[r_{t+1} + \gamma G_{t+1} \mid s_t, a_t\right] \\ &= r_{t+1} + \gamma \mathbb{E}_{a_{l>t+1}\sim\pi}\left[G_{t+1} \mid s_{t+1}, a_{t+1}\right] \\ &= r_{t+1} + \gamma Q^{\pi}\left(s_{t+1}, a_{t+1}\right). \end{aligned} \tag{12}$$

Eq. (12) is the Bellman equation, which shows that the action-value function is composed of the current reward and the action-value function at next state. Our goal is to find the optimal policy π to maximize action-value function.

Note that the above MDP optimization problem is non-convex due to the randomness of shadowed-Rician fading, and we omit the specific proof since the limited pages in this paper. Moreover, this is a complicated multiuser problem with large state space and action space, and the existing methods such as dynamic programming (DP), Lagrange multiplier and Karush Kuhn Tucker (KKT) conditions are difficult to deal with this problem due to high computation complexity and large storage space. Therefore, we utilize the DRL to solve our problem.

IV. The Proposed DRAP Algorithm

In this section, we first describe the idea of Deep Q-Network (DQN) and then propose our DRAP algorithm.

A. *Principle of DQN*

Q-learning is a commonly used RL algorithm whose basic idea is to estimate the Q-value, i.e. $Q(s_t, a_t)$, by using the Bellman equation as an iterative update [16]. The Q-value is updated by the reward obtained in current state and the maximum Q-value of the next state, which can be expressed as

$$\begin{aligned} Q(s_t,&a_t) \leftarrow Q(s_t, a_t) \\ &+ \lambda\left[r_{t+1} + \gamma \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}\right) - Q(s_t, a_t)\right], \end{aligned} \tag{13}$$

where $r_{t+1} + \gamma \max\limits_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}\right) - Q(s_t, a_t)$ represents the deviation between the current Q-value and the maximum Q-value of the next state, and λ is the learning rate. The Q-value function eventually converges to Q^* after certain iterations, and we can obtain the optimal policy expressed by

$$\pi^* = \arg\max_{a\in\mathcal{A}} Q^*(s, a). \tag{14}$$

Q-learning performs well when dealing with simple problems, but it becomes impractical when the state and action spaces are large due to high computational complexity of the iterative method.

To overcome this issue, researchers combined neural networks and Q-learning to develop an algorithm called DQN, which is a pioneering work in the field of DRL [16]. DQN uses deep neural networks (DNN) to approximate the action-value function. Given the state s, the network will output a vector of action-value $Q(s, \cdot; \theta)$, where θ are the parameters of the Q-network.

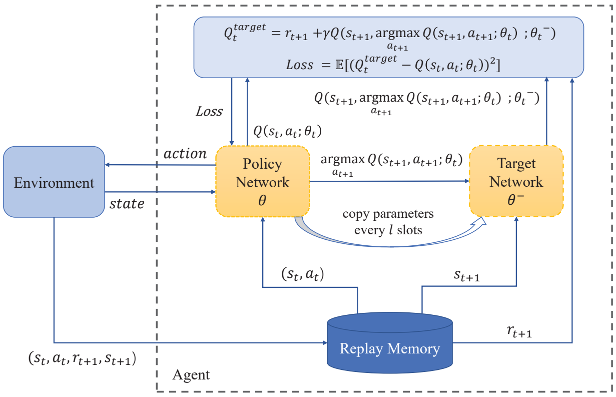
$Q_t^{target} = r_{t+1} + \gamma Q(s_{t+1}, \operatorname{argmax}_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_t)\ ; {\theta_t}^{-})$
$Loss\ = \mathbb{E}[(Q_t^{target} - Q(s_t, a_t; \theta_t))^2]$
Loss
$Q(s_t, a_t; \theta_t)$
$Q(s_{t+1}, \operatorname{argmax}_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_t)\ ; {\theta_t}^{-})$
action
state
Environment
Policy Network θ
$\operatorname{argmax}_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_t)$
Target Network θ^-
copy parameters every l slots
(s_t, a_t)
s_{t+1}
r_{t+1}
Replay Memory
$(s_t, a_t, r_{t+1}, s_{t+1})$
Agent

Fig. 1. The architecture of the proposed DRAP algorithm.

B. *Architecture of the proposed DRAP algorithm*

Inspired by the idea of DQN, we propose the DRAP algorithm to solve the optimization problem. The architecture of the algorithm is shown in Fig. 1. The agent interacts with the S-IoT environment through a sequence of states, actions and rewards to select the optimal actions to maximize the Q-value.

Agent transits to state s_{t+1} and gets reward r_{t+1} when the satellite takes an action a_t to communicate with UEs in state s_t. We store the generated data $(s_t, a_t, r_{t+1}, s_{t+1})$ at each time slot t in a replay memory with size D, and randomly use a mini-batch of samples during the training process. This technique is called experience replay, which can reduce the correlation between samples [16].

A DNN with parameters θ is used to approximate the Q-value function. The network learns the parameters by updating the current value $Q\left(s_t, a_t; \theta_t\right)$ toward a target value defined as

$$Y_t^{\mathrm{Q}} = r_{t+1} + \gamma \max_{a_{t+1}} Q\left(S_{t+1}, a_{t+1}; \theta_t\right). \tag{15}$$

The loss function that is minimized for training the network can be written by

$$L_t\left(\theta_t\right) = \mathbb{E}_{\left(s_t, a_t, r_{t+1}, s_{t+1}\right)}\left[\left(Y_t^{\mathrm{Q}} - Q\left(s_t, a_t; \theta_t\right)\right)^2\right]. \tag{16}$$

In the original DQN, the max operator in Eq. (15) uses the same values to select and evaluate the action. This makes it tend to overestimate the Q-value, and leading to an inaccurate training model. Hence, we use a double-DQN structure to overcome this problem [17]. Specifically, we use a policy network with parameters θ to predict action and evaluate the action by a target network with parameters θ^{-}. Hence, Eq. (15) can be rewritten as

$$Q_t^{\text{target}} = r_{t+1} + \gamma Q\left(s_{t+1}, \underset{a_{t+1}\in\mathcal{A}}{\operatorname{argmax}} Q\left(s_{t+1}, a_{t+1}; \theta_t\right); \theta_t^{-}\right). \tag{17}$$

Both networks use the same structure, and the parameters of policy network will be copied to target network every l slots.

Algorithm 1: DRAP Algorithm

```
Input: $N$, $T$, $l$, $K_s$, $\varepsilon$, $\eta$, $\gamma$;
Output: The optimal policy network $Q^*(s, a \mid \theta)$;
Initialize the network parameters $\theta$ and $\theta^-$ randomly;
Initialize the state $s_0$;
Initialize the replay memory with size $D$;
for $n = 1; n \leq N; n++$ do
    Reset the state $s_t = s_0$;
    for $t = 1; t \leq T; t++$ do
        Input $s_t$ into policy network and obtain Q-values $Q(s_t, \cdot; \theta_t)$ for all actions;
        Generate a random number $e \in [0, 1]$;
        if $e < \varepsilon$ then
            Randomly select an action $a_t \in \mathcal{A}$;
        else
            Select the action with the largest Q-value $a_t = \underset{a_t \in \mathcal{A}}{\arg\max} Q(s_t, a_t; \theta_t)$;
        end
        Update $\varepsilon = \eta\varepsilon$;
        Satellite transmits status updates to UEs by taking action $a_t$, and obtain the reward $r_{t+1}$ and $s_{t+1}$;
        Store data $(s_t, a_t, r_{t+1}, s_{t+1})$ in replay memory;
        Randomly select $K_s$ samples from replay memory;
        Update the target Q-value based on $Q_t^{\text{target}} = r_{t+1} + \gamma Q\left(s_{t+1}, \text{argmax}\, Q\left(s_{t+1}, a_{t+1}; \theta\right); \theta^-\right)$;
        Train the parameters $\theta$ with with gradient descent method;
    end
    if $t \% l = 0$ then
        Update the target network as $\theta^- = \theta$;
    end
end
Return policy network $Q^*$ with $\theta$;
```

C. Training the DRAP Algorithm

In this subsection, we describe the training process and detailed implementations in Algorithm 1. The training process is consists of N episodes, and each episode contains T time slots.

Before training, the satellite first performs D transmissions to obtain enough data to initialize the replay memory. In the training phase, the action of transmission in time slot t is obtained by ε-greedy policy, that is, select action a_t randomly with the probability of ε, and select the action according to the policy network with probability $(1-\epsilon)$. This is to allow more actions to be explored and evaluated. After taking the action a_t, the agent will move to state s_{t+1} and obtain a reward r_{t+1}, then the data $(s_t, a_t, r_{t+1}, s_{t+1})$ is stored in replay memory. A batch of K_s samples are selected to train the policy network, and the target network is updated with the period of l time slot.

TABLE I
SIMULATION PARAMETERS.

Parameters	Value
Bandwidth B	200 MHz
Target rate R^{target}	150 Mbit/s
Height of satellite	300 km
Duration of time slot	1 ms
The average power of scatter component b_0	0.063
The average power of LoS component Ω	0.00897
The Nakagami-m parameter m	1
error-propagation factor β	0.05
Discount factor γ	0.9
Initial ε	1.0
Replay memory size	10000
Batch size	32
Training episodes	5000
Time slots per episode	100

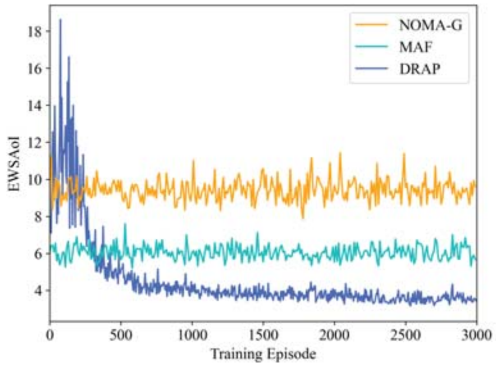

Fig. 2. EWSAoI of DRAP scheme compared to NOMA-G and MAF scheme versus training episode, where $K = 4$ and SNR = 17 dB.

V. Numerical Results

A. Simulation Setup

In this section, we present the simulation results of our proposed DRAP scheme and compare it with two existing schemes: 1) NOMA-G scheme [14], where satellite allocates power to UEs depending on the channel conditions. UE$_i$ with smaller channel gain will be allocated more power; 2) MAF scheme [18], where UE$_i$ with the highest value of $\Delta_i(t)$ is greedily allocated as much power as possible with constraint $C1 \sim C3$ in each slot t. Simulation parameters are summarized in Table I.

B. Simulation Results

The convergence of the network is shown in Fig. 2, and we set $K = 4$ and transmit SNR $= 17$ dB. The simulation result gives an observation that EWSAoI of our proposed DRAP scheme fluctuates at relatively high values in the first 250 episodes. This is because the agent randomly explores the value of actions with a large probability at the beginning of training, and the network is not convergent initially. After 400 episodes, our DRAP scheme outperforms the MAF scheme. After 2000 episodes, the EWSAoI of DRAP is approximately

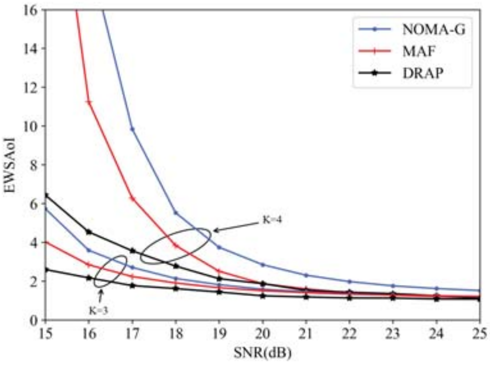

Fig. 3. EWSAoI versus transmit SNR by using the DRAP scheme, MAF scheme and NOMA-G scheme.

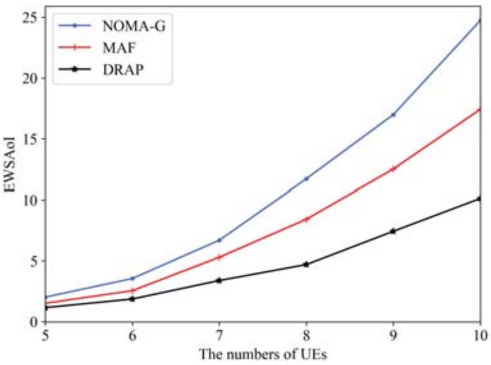

Fig. 4. EWSAoI versus the different numbers of the UEs in the network, where SNR=30dB.

62% and 41% lower than that of NOMA-G scheme and MAF scheme, respectively.

Fig. 3 illustrates the EWSAoI performance of 3 UEs and 4 UEs versus transmit SNR by using the DRAP scheme, MAF scheme and NOMA-G scheme. We can find that the AoI performance of our proposed DRAP scheme outperforms the other two existing schemes especially when SNR is low. The gap of EWSAoI between the DRAP scheme and the other two schemes narrows as the SNR increases. This is because the outage probability of each transmission is at a low level when the SNR is high. Moreover, in the case of 4 UEs, the performance of NOMA-G scheme and MAF scheme are worse in low SNR due to more interferences during SIC decoding, but our DRAP scheme can still keep EWSAoI at relatively low values. This shows the advantage of DRAP scheme in multiuser system when SNR is low.

In addition, we simulate the performance of these schemes versus the different numbers of UEs where SNR =30 dB in Fig. 4, which shows that the EWSAoI of all the schemes increase as the numbers of UEs increase. This is because the increasing number of UEs will lead to insufficient power resources in the network and more interferences. We can also observe that the EWSAoI of DRAP scheme considerably outperforms the other two schemes.

VI. Conclusions

In this paper, to minimize the average AoI of a NOMA downlink S-IoT network, we have formulated an optimization problem subject to power constraints and then reformulated it as a MDP problem. To overcome the curse of dimensionality of the MDP and the non-convexity of the optimization problem, we have developed a DRAP algorithm and trained a double-network DRL under different system settings to obtain the optimal solution, and the simulation results showed that our proposed DRAP scheme can significantly outperform the benchmark schemes. In future work, different arrival rates of status update in the satellite can be studied in our optimization framework.

References

[1] J. Jiao, et al., "Massive access in space-based Internet of Things: Challenges, opportunities, and future directions," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 118–125, 2021.

[2] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5712-5728, Sept. 2020.

[3] J. Jiao, et al., "Intelligent hybrid non-orthogonal multiple access relaying for vehicular networks in 6G," *IEEE Internet of Things Journal*, vol. 8, no. 19, pp. 14773–14786, 2021.

[4] S. Kaul, et al., "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2731-2735.

[5] A. Gong, et al., "Age-of-information-based scheduling in multiuser uplinks with stochastic arrivals: A POMDP approach," in *2020 IEEE Global Communications Conference*, 2020, pp. 1-6.

[6] Y. P. Hsu, et al., "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Transactions on Mobile Computing*, vol. 19, no. 12, pp. 2903-2915, 1 Dec. 2020.

[7] A. M. Bedewy, et al., "Optimal sampling and scheduling for timely status updates in multi-source networks," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4019-4034, June 2021.

[8] S. Liao, et al., "Age-optimal power allocation scheme for NOMA-based S-IoT downlink network," in *IEEE ICC 2021*, Virtual/Montreal, Canada, 2021.

[9] A. Maatouk, et al., "Minimizing the age of information: NOMA or OMA?" in *2019 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019.

[10] C. Guo, et al., "Age-optimal power allocation policies for NOMA and hybrid NOMA/OMA systems," in *IEEE International Conference on Communications*, 2021.

[11] Q. Wang, et al., "Minimizing age of information via hybrid NOMA/OMA," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1753-1758.

[12] L. You, et al., "Massive MIMO transmission for LEO satellite communications," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1851-1865, 2020.

[13] D. Goto, et al., "LEO-MIMO satellite systems for high capacity transmission," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1-6.

[14] J. Jiao, et al., "Fairness-improved and QoS-guaranteed resource allocation for NOMA-based S-IoT network," *Science China-Information Sciences*, vol. 64, p. 169306, 2021.

[15] Z. Xiang, et al.,"Secure transmission for NOMA systems with imperfect SIC," *China Communications*, vol. 17, no. 11, pp. 67-78, Nov. 2020.

[16] V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[17] H. Hasselt, et al., "Deep reinforcement learning with double Q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2016.

[18] I. Kadota, et al., "Scheduling policies for minimizing age of information in broadcast wireless networks" *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2637–2650, Dec. 2018.