Delay Constrained Resource Allocation for NOMA Enabled Satellite Internet of Things with Deep Reinforcement Learning

Xiaojuan Yan, *Member, IEEE*, Kang An, Qianfeng Zhang, Gan Zheng, *Fellow, IEEE*, Symeon Chatzinotas, *Senior Member, IEEE*, and Junfeng Han

***Abstract*—With the ever increasing requirement of transferring data from/to smart users within a wide area, satellite internet of things (S-IoT) networks has emerged as a promising paradigm to provide cost-effective solution for remote and disaster areas. Taking into account the diverse link qualities and delay quality-of-service (QoS) requirements of S-IoT devices, we introduce a power domain non-orthogonal multiple access (NOMA) scheme in the downlink S-IoT networks to enhance resource utilization efficiency and employ the concept of effective capacity to show delay-QoS requirements of S-IoT traffics. Firstly, resource allocation among NOMA users is formulated with the aim of maximizing sum effective capacity of the S-IoT while meeting the minimum capacity constraint of each user. Due to the intractability and non-convexity of the initial optimization problem, especially in the case of large-scale user-pair in NOMA enabled S-IoT. This paper employs a deep reinforcement learning (DRL) algorithm for dynamic resource allocation. Specifically, channel conditions and/or delay-QoS requirements of NOMA users are carefully selected as state according to exact closed-form expressions as well as low-SNR and high-SNR approximations, a deep Q network is first adopted to yet reward and output the optimum power allocation coefficients for all users, and then learn to adjust the allocation policy by updating the weights of neural networks using gained experiences. Simulation results are provided to demonstrate that with a proper discount factor, reward design, and training mechanism, the proposed DRL based power allocation scheme can output optimal/near-optimal action in each time slot, and thus, provide superior performance than that achieved with a fixed power allocation strategy and orthogonal multiple access (OMA) scheme.**

***Index Terms*—Power-domain non-orthogonal multiple access (NOMA), satellite internet of things (S-IoT), resource allocation, deep reinforcement learning (DRL).**

I. Introduction

SATELLITE internet of things (S-IoT), in which the special characteristics of satellite networks are utilized to achieve the ubiquitous coverage, is capable of providing remote detection and swift reaction with destructive disasters as well as environmental monitor and communicate other [1], [2]. Although S-IoT can realize anything and anyone communications in anytime and anywhere, the increasingly growing number of IoT devices will reach 24.1 billion by 2030 as predicted [3], the ever continuously emerging of new applications and evolving of former applications to scenarios with more stringent reliability/latency/data rate requirements, all directly lead to a vital design constraint in S-IoT because of limited spectrum resource.

Non-orthogonal multiple access (NOMA) scheme, which can be perfectly integrated with the existing orthogonal multiple access (OMA) scheme by exploiting the power domain for multiple access within each resource block, has been considered as a promising approach in S-IoT to increase spectrum and energy resource utilization [4]. In recent years, many efforts have investigated the performance of NOMA based S-IoT networks from various performance metrics. Particularly, the network utility maximization were investigated in [5] and [6] with joint optimization and deep learning methods, respectively. Considering the channel phase uncertainty, the authors in [7] proposed two robust beamforming algorithms to minimize the total power consumption in NOMA-based multi-beam S-IoT networks, and proved that the power consumed with NOMA scheme was far less than that with OMA scheme. In addition, the authors in [8] conducted outage performance investigation of NOMA users with fixed power allocation strategy in millimeter-wave band S-IoT networks, where the direct access were unavailable and multiple antennas were deployed at the relay node.

While these aforementioned works have investigated various S-IoT scenarios assisted with the NOMA scheme is superior to that with OMA scheme, the main limitation of those works is that only the fixed power allocation is assumed. Since the interference caused by superposition coding at the transmit side is cancelled by using successive interference cancellation (SIC) at the receiver side [9]–[14], the power allocated to one NOMA user crucially affect its ability to remove inter-interference and observe its own signal. Therefore, power allocation strategy has a significant impact on NOMA users' performance and

This work is supported by the NSFC under grant 61901502, Guangxi Natural Science Foundation under grant 2020GXNSFBA159051, the Scientific Research Foundation of Beibu Gulf University under grant 2019KYQD40, and the Basic Ability Enhancement Program for Young and Middle-aged Teachers of Guangxi 2020KY10020. (Corresponding author: K. An.)

X. Yan, Q. Zhang, and J. Han are with the College of Mechanical, Naval Architecture and Ocean Engineering, Beibu Gulf University, Qinzhou 535011, China (e-mail: yxj9609@163.com, qfzhang19@163.com, and xz@bbgu.edu.cn).

K. An is with the Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China (e-mail: ankang89@nudt.edu.cn).

G. Zheng is with the Wolfson School of Mechanical, Electrical, and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: g.zheng@lboro.ac.uk).

S. Chatzinotas is with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg (e-mail: symeon.chatzinotas@uni.lu).

the superiority of the NOMA scheme. Fixed power allocation policy, as a result, can not well suit users' channel diversities and effectively provide an improved spectral efficiency, even at the cost of increased complexity [15].

Moreover, in most existing studies on NOMA based S-IoT networks, the delay quality-of-service (QoS) requirements of users were not taken into consideration. Although the S-IoT receivers may located in the same beam spot coverage area, receivers still have various delay-QoS requirements [16], [17] and channel qualities due to their application scenarios and location environments, i.e., smart grid in remote locations is an identical delay-critical scenario, while environmental monitoring is a typical delay-tolerant scenario. In this regard, the authors in [5] and [6] conducted a jointly network stability and power allocation optimization problem for long-term network utility, which failed to take into consideration the specific delay-QoS requirement of each S-IoT user. By using a one-dimensional numerical search (NS) strategy, the achievable system performance for a two user NOMA system under delay-limited QoS constraint was studied in [18] and [19] over κ-μ shadowed fading and in single-input multiple-output scenarios, respectively. However, the difficulty of this NS strategy to precisely design the search ranges of power factors' increases with the number of NOMA users. Under this condition, optimum power allocation for NOMA based S-IoT with various delay-QoS requirements are required to meet the urgent requirement of improving the resource utilization efficiency and application scenarios with more stringent reliability/latency/data rate requirements.

Due to the combinatorial feature of delay-QoS requirement and the non-convex property of power allocation in NOMA systems, it is nontrivial to find a global optimum solution, especially in S-IoT with complex compelling application in military and civilian fields. To tackle this issue, several prior works turned to machine learning (ML) tools to achieve an effective solution for resource allocation, where efficient solutions can be obtained without model oriented analysis and design. In recent years, two main branches of ML, namely, supervised learning, such as neural network and support vector machine, and reinforcement learning (RL), i.e., Q-learning and SARSA, have been incorporated in various wireless networks with different objectives, i.e., the authors in [20] proposed a genetic algorithm (GA) improved support vector machine scheme to effectively pair users for NOMA based satellite networks. In multibeam satellite systems, the work in [21] proposed a fully connected deep neural network assisted approach to facilitate efficient beam hopping. A neural network improved GA was proposed in [22] to study the issue of satellite data downlink replanning problem for IoT internet connection. In [23], the authors proposed a convolutional neural network based approach to detect anomalous network activity and improve the traffic control performance for space-air-ground integrated networks. It is noted that supervised learning, such as algorithm used in [20]-[23], need a certain amount of labeled data to infer a function. Since RL is model free and data driven by learning from interaction with the environment, it has been extensively adopted in wireless networks for dynamic and low-latency design without the knowledge of accurate mathematical models. For example, based on Q-learning, a long-term optimal capacity allocation algorithm was proposed in [24] to optimize the long term utility of a multi-layer satellite network. Considering an energy constrained S-IoT, the work in [25] applied a RL based approach for optimal channel allocation. The authors in [26] proposed a spatial anti-jamming scheme using Stackelberg game and RL for heterogeneous internet of satellites to minimize anti-jamming routing cost. In [27], the authors adopted a deep reinforcement learning (DRL) in heterogeneous satellite networks to allocate resource more flexibly and efficiently among different satellite systems. In [28] and [29], the authors conducted resource allocation in multi-user cellular network with the help of DRL and enhanced DRL algorithms, respectively. Simulations of these prior works have shown that ML in wireless networks can help to achieve optimal or near-optimal performance with reduced computational complexity.

Motivated by these observations, here we resort to DRL algorithms to effectively allocate resource, such as power allocation among NOMA users, to provide services with various delay-QoS requirements and high resource utilization efficiency for future S-IoT systems. Then main contributions of this work are follows:

- Both users' delay-QoS requirements and minimum rate limitations are taken into account in the proposed resource allocation scheme. Particularly, we employ the concept of effective capacity, which is firstly introduced in [30] as the maximum constant arrival rate that can be supported under a given delay constraint, to investigate the effect of each user's delay-limited QoS constraint on the performance of proposed system. Then, we formulate an optimization problem to obtain an optimal power allocation scheme to maximize sum effective capacity while satisfying minimum rate limitations for each NOMA user.
- By deriving exact closed-form expressions and approximated low-SNR as well as high-SNR expressions for effective capacity of each NOMA user, we study in detail that how key parameters, such as power allocation factor and delay-QoS requirement, impact the performance of each user. On this basis, the state space of the DRL algorithm is carefully designed according to transmission condition. Moreover, to ensure the superiority of the NOMA scheme, the reward is set as zero if any user's performance is smaller than that achieved with a time-division multiple access (TDMA).
- The proposed DRL based power allocation approach is compared to the TDMA scheme, NOMA with fixed power allocation strategy [31], and NOMA with numerical search strategy [18], which reveal the superiority of introducing the NOMA scheme and DRL algorithm in the S-IoT from the perspective of performance enhancement and reduced compute complexity. Specifically, the proposed approach is proved to be superior to the TDMA and NOMA with fixed power allocation strategy, by selecting an optimum/near optimum action in each time slot.

The rest of this paper is outlined as follows. The system model including related channel model and signal model is

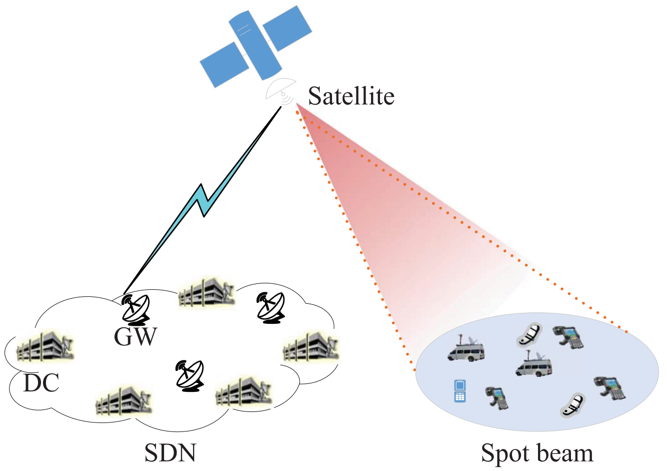

Fig. 1. Architecture of NOMA based S-IoT networks

presented in Section II. Section III introduces the concept of effective capacity, derives the exact capacity expressions as well as approximated low-SNR and high-SNR capacity expressions for each NOMA user, and formulates the resource allocation problem for the delay-constrained NOMA based downlink S-IoT networks under minimum rate constraints. In Section IV, DRL algorithm is introduced in detail and tested in the proposed system. Simulation results and discussions are provided and conclusions are made in Sections V and VI, respectively.

II. System model

As shown in Fig. 1, using Software Defined Network (SDN) technology, a Date Center (DC) transfers signal to a Gateway (GW) who has a good link quality [32]. Then, with the help of the NOMA scheme, signals from DCs are superposed at the GW by allocating different power to each user, the linear superposition of these signals is subsequently broadcasted via a satellite to its corresponding smart users. Here, we assume the number of users is $M (M \geq 2)$ and these users are uniform deployed in the same spot beam but with different locations and channel statistical prosperities. Moreover, all nodes in the proposed model are assumed to be equipped with a single antenna for simplicity. Before introducing the proposed QoS-delay guaranteed resource allocation strategy, the link model and the signal model with NOMA scheme are introduced as follows:

A. Link model

From satellite to users, the entire link budget of User j $(j = 1, 2, \cdots, M)$ can be modeled as

$$Q_j = G_j L_j G_s(\varphi_j) |g_j|^2 L_{pj}, \tag{1}$$

where

- G_j and L_j are the antenna gain and free space propagation loss at User j, respectively. Due to the fact that NOMA S-IoT users are served within the same frequency and spot beam coverage area, we assume that $G_j = G$ and $L_j = L$ in this paper for simplicity.

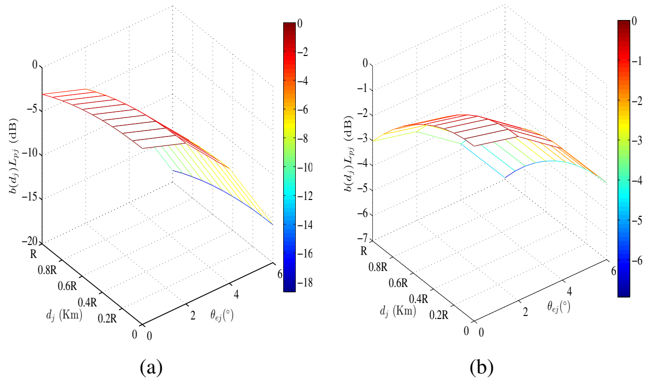

Fig. 2. $b(d_j) L_{pj}$ versus d_j and θ_{ej} with $f_j = 4$ GHz for various D: (a) $D = 1$ m and (b) $D = 0.5$ m.

- $G_s(\varphi_j)$: The beam gain of User j, here φ_j denotes the angle between User j and beam center with respect to the satellite, which is closely related to the location of User j and approximated as [9]

$$G_s(\varphi_j) \approx G_{\max} \left(\frac{J_1(ad_j)}{2ad_j} + 36 \frac{J_3(ad_j)}{a^3 d_j^3} \right)^2 = G_s(d_j), \tag{2}$$

where $G_{\max}$ is the maximum antenna gain, $J_n(\cdot)$ is the Bessel function of first kind and n-th order [33], d_j denotes the distance from the beam center to User j, and $a = 2.07123/R$ with R being the radius of the beam spot, whose coverage area is approximated as a circle.

- $|g_j|^2$: The channel power gain of satellite link is assumed to follow a Shadowed-Rician fading model, which is mathematically tractable and widely applied in various fixed and mobile satellite services for a variety of frequency bands, such as the UHF-band, L-band, S-band, and Ka-band [34]–[36]. In this case, the probability density function (PDF) of $|g_j|^2$ is given by [37]

$$f_{|g_j|^2}(x) = \alpha_j e^{-\beta_j x} {}_1F_1(m_j; 1; \delta_j x), \tag{3}$$

where $\alpha_j = \frac{(2b_j m_j)^{m_j}}{2b_j (2b_j m_j + \Omega_j)^{m_j}}$, $\delta_j = \frac{\Omega_j}{2b_j (2b_j m_j + \Omega_j)}$, $\beta_I = \frac{1}{2b_j}$ with $2b_j$ and Ω_j being the average power of the multipath and the LoS components, respectively, m_j $(m_j > 0)$ being the Nakagami-m fading parameter, and ${}_1F_1(a; b; c)$ being the confluent hypergeometric function [33, Eq. (9.100)]. Moreover, the authors in [37] have also associated parameters b_j, m_j, and Ω_j to elevation angles θ_j when $20^0 \leq \theta_j \leq 80^0$, as

$$\begin{aligned} b_j(\theta_j) &= -4.7943 \times 10^{-8} \theta_j^3 + 5.5784 \times 10^{-6} \theta_j^2 \\ &\quad -2.1344 \times 10^{-4} \theta_j + 3.271 \times 10^{-2}, \\ m_j(\theta_j) &= 6.3739 \times 10^{-5} \theta_j^3 + 5.8533 \times 10^{-4} \theta_j^2 \\ &\quad -1.5973 \times 10^{-1} \theta_j + 3.5156, \\ \Omega_j(\theta_j) &= 1.4428 \times 10^{-5} \theta_j^3 - 2.3798 \times 10^{-3} \theta_j^2 \\ &\quad +1.2702 \times 10^{-1} \theta_j - 1.4864. \end{aligned} \tag{4}$$

- L_{pj}: Note that elevation angle θ_j given in (4) is an ideal angle without any antenna-pointing error taken into account, which is unavailable in practice due to the mobility of satellite terminal and/or satellite perturbation caused by Moon, Sun, and atmospheric drag. Thus, depointing loss L_{pj} is inevitable, according to [38], L_{pj} in dB can be written as $L_{pj} = 27.23 \times 10^{-3} f_j^2 D^2 \theta_{ej}^2$ with θ_{ej} being the pointing error angle at User j and D the diameter of antenna aperture.

Fig. 2 illustrates the effects of parameters d_j, D, and θ_{ej} on the link budget of User j, here $\theta_j = 50°$ and $b\left(d_j\right) = G_s(d_j)/G_{\max}$. As shown in this figure, links with good levels when User j in perfect condition, i.e., d_j and/or θ_{ej} is small, and bad level with severe losses due to the d_j, the D, or the θ_{ej} increases are visible. Thus, we can make a conclusion that significant differences in S-IoT users' link budgets can be clearly observed even if they are assumed to experience the same antenna gain and free space loss. To facilitate performance evaluation in following sections, we assume the channel qualities of these M S-IoT users are in an ascending order, i.e., $G_s(d_i)\left|g_i\right|^2 L_{pi} < G_s(d_j)\left|g_j\right|^2 L_{pj}$ for $i < j$ if without other description.

B. Signal model

After modeling the link budget, this subsection gives the signal models of the downlink S-IoT system with the OMA and NOMA schemes in the following:

1) OMA: For the OMA scheme, such as the TDMA commonly applied in S-IoT, a specified time slot is allocated to User j, within which its unit energy signal, x_j, is transmitted from satellite with transmission power P_s. The received signal at User j is $y_j = \sqrt{P_s Q_j} x_j + w$ where w denotes the noise at the User j with zero mean and δ^2 variance. Thus, the signal-to-interference-plus-noise ratio (SINR) of User j is

$$\gamma_j^{\mathrm{T}} = \frac{P_s Q_j}{\delta^2} = \Theta_j \bar{\gamma} |g_j|^2, \tag{5}$$

where $\Theta_j = L_j G_j G_s\left(d_j\right) L_{pj}$ and $\bar{\gamma} = P_s/\delta^2$ is the transmission average SNR.

2) NOMA: Based on the NOMA scheme, the satellite can broadcast a superposed signal x ($x = \sum_{j=1}^{M} \sqrt{\alpha_j^p P_s x_j}$) to multiple users over the same time/frequency block, where α_j^p is a fraction of the transmission power P_s allocated to User j and x_j ($\mathrm{E}\left[|x_j|^2\right] = 1$) is the signal for User j. The received signal at User j is $y_j = \sqrt{Q_j} x + w$. According to the principle of the NOMA scheme, user with the worst link condition decodes its own information directly, thus, the instantaneous SINR of User 1 is

$$\gamma_1^N = \frac{\alpha_1^p \bar{\gamma} \Theta_1 |g_1|^2}{\left(1 - \alpha_1^p\right) \bar{\gamma} \Theta_1 |g_1|^2 + 1}. \tag{6}$$

For User k ($1 < k < M$), SIC strategy will be used to decode and remove the interference from users with worse link conditions. Since the ascending order of the link budgets, the use of SIC can be always guaranteed at the User k [31]. Then, the achieved SINR at User k can be given by

$$\gamma_k^N = \frac{\alpha_k^p \bar{\gamma} \Theta_k |g_k|^2}{\sum_{i=k+1}^{M} \alpha_i^p \bar{\gamma} \Theta_k |g_j|^2 + 1}. \tag{7}$$

While for user with the best channel gains, by conducting the SIC, its own information can be observed and written as

$$\gamma_M^N = \alpha_M^p \bar{\gamma} \Theta_M |g_M|^2 . \tag{8}$$

The link budget and signal model will be subsequently used in the effective capacity evaluation and associated optimization.

III. Effective capacity and problem formulation

In this section, we present the optimal problem formulation for the considered delay-QoS constrained downlink S-IoT systems. Firstly, the concept of effective capacity is presented. Then, the analytical expressions for the effective capacity of the considered system with both OMA (for comparison) and NOMA scheme are, respectively, derived. Finally, the resource optimum allocation problem for the delay-constrained NOMA based downlink S-IoT networks under minimum rate constraints is formulated.

A. Effective capacity

S-IoT system aims at providing a wide range of services with various QoS-delay requirements. For this reason, we adopt the concept of effective capacity, which provides a measure for the maximum constant supportable source rate for a given delay exponent requirement characterized by θ ($\theta \geq 0$). Different from Shannon capacity without any requirements on QoS-delay, effective capacity guarantees a latency violation probability for the incoming user's data traffic in the wireless network. Specifically, for delay-critical application such as smart grid, a stringent latency should be ensured and the effective capacity turns to be the outage capacity. While for delay-tolerant application such as environmental monitoring which concerns more on data throughput, a loosen latency is needed and the effective capacity tends to be the ergodic capacity [39]. Given a QoS-delay exponent θ, the normalized effective capacity for an independent and identically distributed (i.i.d.) block fading channel can be given by

$$C\left(\theta\right) = \frac{-1}{\theta T_f B} \ln\left(E\left\{e^{-\theta T_f B R}\right\}\right), \tag{9}$$

where T_f denotes the length of each fading block, B is the system bandwidth, $R = \log\left(1 + \gamma^{T/N}\right)$ is the transmission rate, and $E\left[\cdot\right]$ denotes the expectation operator. It is worth noting that a larger QoS-delay exponent θ is needed for a more critical latency requirement application.

B. Effective capacities with OMA and NOMA schemes

After introducing the fundamental concept, the effective capacities for both OMA based and NOMA based downlink S-IoT networks are discussed as follows:

1) OMA: By substituting (5) into (9), the effective capacity of User j with the TDMA scheme can be given by

$$C_j^T\left(\theta_j\right) = \frac{-\ln\left(E\left\{e^{-\theta_j T_f B \log\left(1 + \Theta_j \bar{\gamma} |g_j|^2\right)/M}\right\}\right)}{\theta_j T_f B}, \tag{10}$$

where the factor M appears because of the time resources needed with the TDMA scheme are M times of that with the

$$C_j^T\left(\theta_j\right) = \frac{-1}{a_j \ln 2} \ln \alpha_j \sum_{k=0}^{\infty} \frac{\Gamma\left(\frac{-a_j}{M} + k\right)}{k!\Gamma\left(-a_j/M\ \right)\Gamma(m_j)} \Theta_j^k \bar{\gamma}^k \beta_j^{-k-1} G_{2,2}^{1,2}\left(\frac{-\delta_j}{\beta_j}\left|\begin{matrix}-k, 1-m_j\\ 0,0\end{matrix}\right.\right). \tag{14}$$

$$C_1^N\left(\theta_1\right) = \frac{-1}{a_1 \ln 2} \ln\left(\alpha_1 \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{\left(1-\alpha_1^p\right)^k}{\Theta_1^{-k}\bar{\gamma}^{-k}} \frac{\Theta_1^n \bar{\gamma}^n \Gamma\left(n-a_1\right)\Gamma\left(k+a_1\right)}{k!n!\Gamma\left(m_1\right)\Gamma\left(-a_1\right)\Gamma\left(a_1\right)} \beta_1^{-k-n-1} G_{2,2}^{1,2}\left[-\frac{\delta_1}{\beta_1}\left|\begin{matrix}-k-n, 1-m_1\\ 0,0\end{matrix}\right.\right]\right). \tag{18}$$

$$C_k^N\left(\theta_k\right) = \frac{-1}{a_k \ln 2} \ln\left(\alpha_k \sum_{l=0}^{\infty}\sum_{n=0}^{\infty} \frac{\Gamma\left(n-a_k\right)\Gamma\left(l+a_k\right)}{n!l!\Gamma(a_k)\Gamma(-a_k)\Gamma(m_k)} \left(\sum_{i=k+1}^{M} \frac{\alpha_i^p\Theta_k}{\bar{\gamma}^{-1}}\right)^l \left(\sum_{i=k}^{M} \frac{\alpha_i^p\Theta_k}{\bar{\gamma}^{-1}}\right)^n \beta_k^{-l-n-1} G_{2,2}^{1,2}\left[\frac{-\delta_k}{\beta_k}\left|\begin{matrix}-l-n, 1-m_j\\ 0,0\end{matrix}\right.\right]\right). \tag{19}$$

$$C_M^N\left(\theta_M\right) = \frac{-1}{a_M \ln 2} \ln\left(\alpha_M \sum_{k=0}^{\infty} \frac{\Gamma\left(-a_M+k\right)}{k!\Gamma\left(-a_M\right)\Gamma\left(m_M\right)} (\alpha_M^p)^k \Theta_M^k \bar{\gamma}^k \beta_M^{-k-1} G_{2,2}^{1,2}\left(\frac{-\delta_M}{\beta_M}\left|\begin{matrix}-k, 1-m_M\\ 0,0\end{matrix}\right.\right)\right). \tag{20}$$

NOMA scheme [18], [19]. By defining $a_j = \theta_j T_f B/\ln 2$, we get

$$\begin{aligned} C_j^T\left(\theta_j\right) &= \frac{-1}{a_j \ln 2} \ln\left(E\left(\left(1+\Theta_j\bar{\gamma}|g_j|^2\right)^{-\frac{a_j}{M}}\right)\right) \\ &= \frac{-1}{a_j \ln 2} \ln\left(\int_0^{\infty} \left(1+\Theta_j\bar{\gamma}x\ \right)^{-\frac{a_j}{M}} f_{|g_j|^2}\left(x\right)dx\right). \end{aligned} \tag{11}$$

To evaluate (11), we respectively express $\left(1+\Theta_j\bar{\gamma}x\right)^{\frac{-a_j}{M}}$ and ${}_1F_1\left(m_j;1;\delta_j x\right)$ in (3) into Binominals representations with [33, Eq. (1.11)] and Meijer-G functions with [33, Eq. (9.14.1)] as

$$\left(1+\Theta_j\bar{\gamma}x\ \right)^{\frac{-a_j}{M}} = \sum_{k=0}^{\infty} \frac{\Gamma\left(\frac{-a_j}{M}+k\right)}{k!\Gamma\left(-a_j/M\ \right)} \Theta_j^k \bar{\gamma}^k x^k, \tag{12}$$

and

$${}_1F_1\left(m_j;1;\delta_j x\right) = \frac{1}{\Gamma\left(m_j\right)} G_{1,2}^{1,1}\left[-\delta_j x\left|\begin{matrix}1-m_j\\ 0,0\end{matrix}\right.\right], \tag{13}$$

where $G_{1,2}^{1,1}\left[\cdot|\cdot\right]$ [33, (9.301)] is the Meijer-G function and $\Gamma\left(\cdot\right)$ [33, (8.310.1)] is the Gamma function. Inserting (3), (12), and (13) into (11) along with [33, (7.813.1)], the desired result for the effective capacity of User j with the TDMA scheme can be derived as (14). Then, we can get the sum effective capacity of the considered system with the TDMA scheme as $C^T = \sum_{j=1}^{M} C_j^T\left(\theta_j\right)$.

2) Exact capacity expressions for NOMA users: By substituting (6)–(8) into (9) and following similar steps as that in the derivation of (11), the individual effective capacity for Users 1, k $(1 < k < M)$, and M in a NOMA pair can be given by

$$C_1^N\left(\theta_1\right) = \frac{-1}{a_1 \ln 2} \ln\left(E\left(\frac{\Theta_1\bar{\gamma}|g_1|^2+1}{\left(1-\alpha_1^p\right)\Theta_1\bar{\gamma}|g_1|^2+1}\right)^{-a_1}\right), \tag{15}$$

$$C_k^N(\theta_k) = \frac{-1}{a_k \ln 2} \ln\left(E\left(1+\frac{\alpha_k^p\Theta_k\bar{\gamma}|g_k|^2}{\sum_{i=k+1}^{M}\alpha_i^p\Theta_k\bar{\gamma}|g_k|^2+1}\right)^{-a_k}\right), \tag{16}$$

$$C_M^N\left(\theta_M\right) = \frac{-1}{a_M \ln 2} \ln\left(E\left(1+\alpha_M^p\bar{\gamma}\Theta_M|g_M|^2\right)^{-a_M}\right). \tag{17}$$

Similarly, with the help of Binominals and Meijer-G functions, the accurate closed-form of effective capacity of S-IoT users with the NOMA scheme can be derived as (18)–(20). Note that these expressions are computationally expensive to evaluate and difficult to directly obtain the impacts of power allocation factor and delay-QoS requirement on the effective capacity of each NOMA user. In the following subsections, the approximated capacity expressions for NOMA users when $\bar{\gamma} \to 0$ and $\bar{\gamma} \gg 1$ are derived.

3) Low SNR Approximated capacity expressions for NOMA users ($\bar{\gamma} \to 0$): When $\bar{\gamma} \to 0$, by using a first order Taylor series expansion, i.e., $C_j^T\left(\theta_j\right) \approx C_j^T\left(\theta_j\right)|_{\bar{\gamma}\to 0} + \bar{\gamma}\dot{C}_j^T\left(\theta_j\right)|_{\bar{\gamma}\to 0}$ with $\dot{C}_j^T\left(\theta_j\right)$ being the first-order of $C_j^T\left(\theta_j\right)$ respect to $\bar{\gamma}$, along with some simple manipulations, approximated capacity expression for User j $(j = 1, 2, \cdots, M)$ can be expressed as

$$C_j^N\left(\theta_j\right) \approx \frac{1}{\ln 2}\int_0^{\infty} \alpha_j^p\Theta_j\bar{\gamma}x f_{|g_j|^2}\left(x\right)dx. \tag{21}$$

Combining (3), (13), and (21) in conjunction with [33, (7.813.1)], we have

$$C_j^N\left(\theta_j\right) \approx \frac{\bar{\gamma}}{\ln 2}\alpha_j^p\Theta_j\beta_j^{-2} G_{2,2}^{1,2}\left(\frac{-\delta_j}{\beta_j}\left|\begin{matrix}-1, 1-m_j\\ 0,0\end{matrix}\right.\right). \tag{22}$$

Based on the above derived results, result for the approximate expression of the sum effective capacity when $\bar{\gamma} \to 0$ can be straightforwardly evaluated as $\sum_{j=1}^{M} C_j^N(\theta_j)$. It is interesting to find from (21) and (22) that when $\bar{\gamma} \to 0$, effective capacity of each NOMA user only depends on the power allocation coefficients and fading severity, and has nothing to do with delay-QoS requirement.

4) High SNR Approximated capacity expressions for NOMA users ($\bar{\gamma} \gg 1$): For the case when $\bar{\gamma} \gg 1$, the approximated effective capacities of Users 1, k, and M can be respectively given by

$$C_1^N\left(\theta_1\right) \approx \frac{-1}{\ln 2} \ln\left(1-\alpha_1^p\right), \tag{23}$$

$$C_k^N(\theta_k) \approx \frac{1}{\ln 2} \ln\left(1+\frac{\alpha_k^p}{\sum_{i=k+1}^M \alpha_i^p}\right), \tag{24}$$

$$C_M^N(\theta_M) \approx \frac{1}{\ln 2}\left[\ln(\alpha_M^p \bar{\gamma}\Theta_M) - a_M^{-1}\ln\left(\int_0^\infty x^{-a_M} f(x) dx\right)\right]. \tag{25}$$

After inserting (3) and (13) into (25) and using [33, (7.813.1)], the closed-form expression for User M can be derived as

$$\begin{aligned} &C_M^N\left(\theta_M\right) \approx \frac{1}{\ln 2}\ln\left(\alpha_M^p \bar{\gamma}\Theta_M\right) + \\ &\frac{-1}{a_M \ln 2}\ln \alpha_M \beta_M^{a_M-1} G_{2,2}^{1,2}\left(\frac{\delta_M}{\beta_M}\middle| \begin{matrix} -a_M, 1-m_M \\ 0,0 \end{matrix}\right). \end{aligned} \tag{26}$$

Finally, by adding each NOMA user's effective capacity, the approximated sum effective capacity of the considered system for $\bar{\gamma} \gg 1$ can be obtained. Specially, from (23)–(26), we find that except user with the best link condition, i.e., User M, capacities of other users mainly depend on power allocation strategy and become independent of fading severity as well as delay-QoS requirement.

In light of these above discussions, we find that power allocation strategy has a significant impact on the performance of each NOMA user, while the delay-QoS requirement can greatly affect sum capacity performance of the proposed system. Moreover, these closed-form expressions show that it is difficult to directly find the optimum power allocation factors to maximize the sum effective capacity for the considered system due to high computational complexity.

C. Problem formulation

Our design objective is to maximize the sum effective capacity of users in a NOMA pair while guaranteeing that each user's performance of NOMA scheme is superior of that of the OMA scheme. To this end, the satellite needs to optimally design the power allocation factor according to user's link quality and delay-QoS requirement, as well as the transmission average SNR of the considered system. Therefore, the optimization problem for the proposed NOMA based S-IoT system, denoted by **P1**, can be formulated as

$$\mathbf{P1}: \max \sum_{j=1}^M C_j^N\left(\theta_j\right) \tag{27a}$$

$$s.t. \quad \sum_{j=1}^M \alpha_j^p = 1,\ 0 < \alpha_j^p < 1, \tag{27b}$$

$$C_j^N\left(\theta_j\right) \geq C_j^T\left(\theta_j\right),\ \theta_j \geq 0, \tag{27c}$$

the constraint (27b) means limited total resource budget, the constraint (27c) represents the delay-QoS and minimum capacity requirements for each NOMA user, which further limits the power allocation factor in a certain range and ensures the superiority of the NOMA scheme in S-IoT networks. Note that, problem **P1** is NP hard and the sum capacity of the proposed system tightly depends on the power allocation strategy. Under this consideration, the aim of this paper is to incorporate the DRL algorithm to allocate power resource and realize two goals simultaneously. One goal is to meet the delay-QoS requirement on each user and the other is to further improve the resource utilization efficiency of the satellite with the NOMA scheme.

IV. DEEP REINFORCEMENT LEARNING

Deep Q-network (DQN) is the most representative value-based method. Through combining the advantages of Q-learning and deep neural network, DQN with a single or multiple agents can learn to predict the expect returns of all actions for a given environment observation. A frameworks of applying DQN in resource allocation for the considered S-IoT is provided as shown in Fig. 1. The satellite acts as an agent and interacts with the unknown environment to gain experiences, learning from which, policy π and decision making are then conceived. This agent is trained by exploring and exploiting the environment and refining power resource allocation strategy based on its own observations of the environment state.

As shown in Fig. 3, the DQN based algorithm can be divided into two phases, i.e., the data generating and the neural network training phases. In the data generating phase, the agent selects an action a_t, according to policy π based on the observed state s_t. The system then moves into a new state s_{t+1} with a certain probability influenced by the system's inherent transitions. Meanwhile, the agent receives a reward R_t from the system. In the network training phase, data stored in the experience pool is randomly chosen to train two neural networks to approximate the Q-value, and thus, replace the need for a table to store the expect returns (Q value). Key elements of the considered DRL based power resource allocation among NOMA users are described in detail as follows.

A. State and observation space S

In the considered DQN framework, we assume that the satellite can sense the environment sate based on its own observation. Moreover, we assume that there is T states need to be taken into account. As analyzed in Section III.B, key parameters, i.e., link qualities and delay-QoS requirements of NOMA users, which influence users' performance by a multiplicity of different system's transmission conditions. Thus, we set state $s_t \in S$ according to system condition at time slot t. For example, when $\bar{\gamma}(t) \to 0$, the s_t is defined as $s_t = \{Q_1\left(t\right), Q_2\left(t\right), \cdots, Q_M\left(t\right)\}$, while when $\bar{\gamma}(t) \gg 1$, state s_t is $s_t = \{Q_1\left(t\right), Q_2\left(t\right), \cdots, Q_M\left(t\right), \theta_M\}$. Otherwise, state s_t is

$$\begin{aligned} s_t = \{&Q_1\left(t\right), Q_2\left(t\right), \cdots, Q_M\left(t\right), \\ &\theta_1\left(t\right), \theta_2\left(t\right), \cdots, \theta_M\left(t\right)\}. \end{aligned} \tag{28}$$

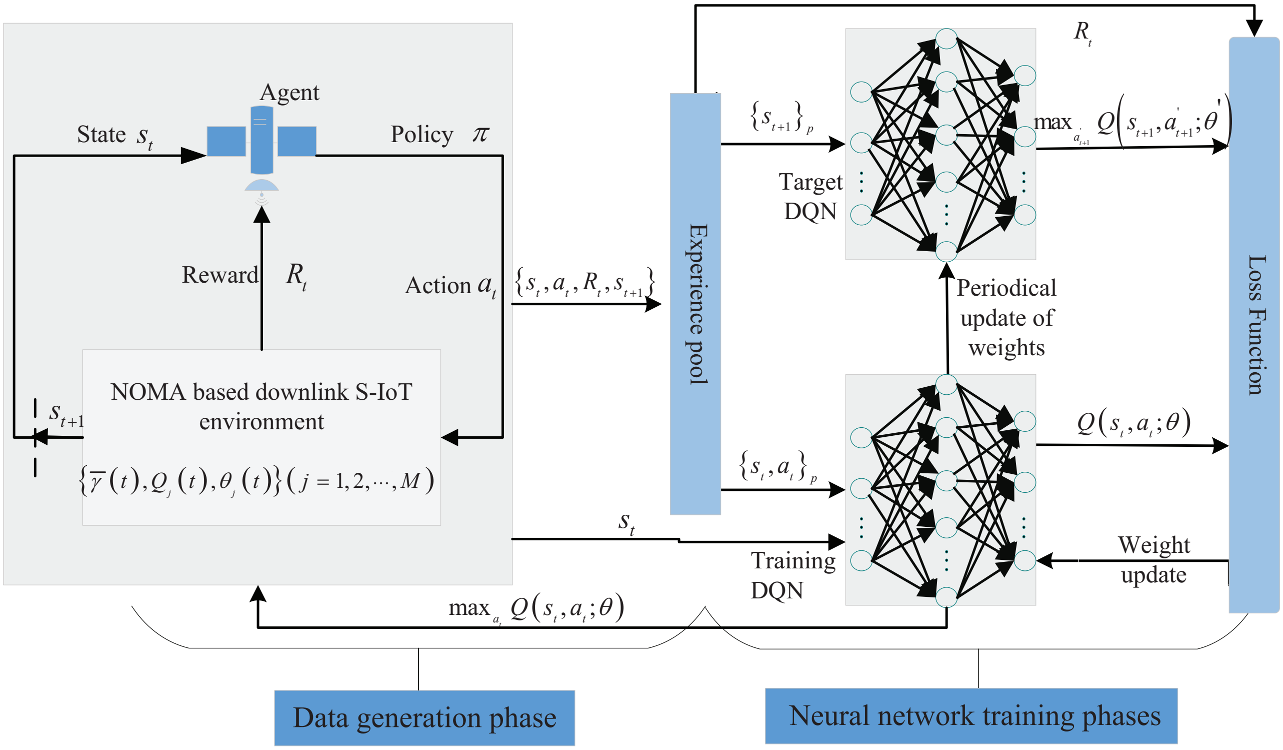
Agent
State s_t
Policy π
Reward R_t
Action a_t
s_{t+1}
NOMA based downlink S-IoT environment
$\{\bar{\gamma}(t), Q_j(t), \theta_j(t)\}(j=1,2,\cdots,M)$
$\{s_t, a_t, R_t, s_{t+1}\}$
Experience pool
$\{s_{t+1}\}_p$
Target DQN
R_t
$\max_{a_{t+1}} Q(s_{t+1}, a'_{t+1}; \theta')$
Periodical update of weights
$Q(s_t, a_t; \theta)$
$\{s_t, a_t\}_p$
s_t
Training DQN
Weight update
Loss Function
$\max_{a_t} Q(s_t, a_t; \theta)$
Data generation phase
Neural network training phases

Fig. 3. DQN based power allocation model.

States in different time slots are determined by variations in average SNR, link budget, and delay-QoS requirements, which need agent to adjust its action in each slot accordingly.

B. Action space A

Power allocation is an important research topic for NOMA based communications, as discussed in prior Sections, power allocation scheme not only affects the performance of each NOMA user and the proposed system, but also impacts the resource utilization efficiency. Thus, power allocated among NOMA users should be designed carefully to contain all the possible power allocation decision. In this paper, considering that power resource at the satellite is continuous and limited by the maximum transmission power P_s, we quantized 1 to N_a numbers and thus form the action space as $A = \alpha_1^p, \alpha_2^p, \cdots, \alpha_{N_a}^p$. Then, at time slot t, the action a_t ($a_t \in A$) is $a_t = \alpha_1^p(t), \alpha_2^p(t), \cdots, \alpha_M^p(t)$, $\sum_{j=1}^{M} \alpha_j^p(t) \leq 1$, where $\alpha_j^p(t)$ is the corresponding power allocation coefficient for User j.

C. Reward design

What makes deep learning algorithm particularly attractive for obtaining solutions in the NP hard and/or hard-to-optimize problems is the flexibility in reward design. If the reward signal at each time slot is designed to have a close association with the desired objective, the system performance can be significantly improved. In the investigated NOMA based S-IoT power allocation problem described in Section III, our objectives are twofold: 1) ensure each user's performance with the NOMA scheme is not less than that achieved with the TDMA scheme and 2) maximize the sum effective capacity of the considered system.

For the first objective, after the agent performs an action, we set User j's achieved effective capacity at each time slot t equals to zero until the constraint in (27c) is satisfied, with which the delay-QoS requirement of User j and the superiority of introducing the NOMA scheme in the S-IoT networks are both ensured. Thus, the effective capacity of User j at time slot t can be rewritten as

$$C_j^N(\theta_j, t) = \begin{cases} C_j^N(\theta_j), if\ C_j^N(\theta_j) \geq C_j^T(\theta_j) \\ 0, \qquad otherwise \end{cases}. \tag{29}$$

Moreover, the sum capacity of all NOMA users will be set to zero if any user's achievable rate with the NOMA scheme is smaller than that achieved with the TDMA scheme.

For the second objective, RL is adopted here to obtain flexible and dynamic power allocation strategy to maximize the achievable system performance and resource utilization efficiency. Since the goal of RL algorithm is to find a policy π (a mapping from states to actions), to guide agent choose a specific action in a certain state to maximize the achievable return $R_\pi(t)$, where $R_\pi(t)$ is defined as the corresponding cumulative discounted rewards, given by

$$R_\pi(t) = \sum_{k=0}^{\infty} \gamma^k \sum_{j=1}^{M} C_j^N(\theta_j, t+k), \tag{30}$$

where γ $(0 \leq \gamma \leq 1)$ is a discount factor used here to trade off between immediate and future rewards. Since that users in S-IoT scenario are assumed to be randomly located, there we suppose little relation between link qualities of these users in different time slots. In this case, we prefer immediate rewards over future rewards and a low discount factor γ value.

D. Learning algorithm

1) Data generation phase: We leverage Q-learning with experience pool of capacity D to generate data for next network training phase. During this process, to tradeoff between exploration and exploitation, some ways, such as ε-greedy exploration, is always used to choose an action at random with probability ϵ $(0 < \epsilon < 1)$ or the best action as deemed by current policy with probability $(1 - \epsilon)$. With this ε-greedy strategy, following policy π, action a_t will be selected in state s_t and Q value function, which describes the expected $R_\pi(t)$, is given by

$$Q_\pi(s_t, a_t) = E\left(R_\pi(t) | S = s_t, A = a_t\right). \tag{31}$$

Updating this action-value function with

$$Q_\pi(s_t, a_t) = Q_\pi(s_t, a_t)(1-\alpha) + \alpha\left(R_\pi(t) + \gamma \max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1})\right), \tag{32}$$

where α denotes the learning rate. The optimal policy $\pi^\star$ can be easily obtained by selecting the highest valued action in each state, i.e., $Q_{\pi^\star}(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t)$. Following the environment transition caused by variations in users' channel qualities and/or delay-QoS requirements, the satellite agent collects and stores the tuple (s_t, a_t, R_t, s_{t+1}) at time slot in the experience pool. Since the size of pool is limited to D, old tuple will be removed to give space for the newest tuple if the pool is full.

2) Neural network training phases: For power allocation task proposed in this paper is a sequential decision problem, and the size of Q values of (31) for all possible actions may be large. In this case, it is challenging to model the Q-learning process efficiently. Thus, deep neural networks parameterized by $\theta^{'}$ and θ, called target DQN and training DQN, respectively, are used to estimate Q value by function approximations. As shown in Fig. 3, with random batches of experiences selected from experienced pool, the target DQN is trained to generate the maximum Q value for next state, i.e., $\max_{a^{'}_{t+1}} Q\left(s_{t+1}, a^{'}_{t+1}; \theta^{'}\right)$. While the training DQN network is for estimating Q values for current state action pairs and making a action decision for state s_t. With parameter θ, the loss function of the DQN network is

$$L(\theta) = \left(R_\pi(t) + \max_{a^{'}_{t+1}} Q\left(s_{t+1}, a^{'}_{t+1}; \theta^{'}\right) - Q(s_t, a_t; \theta)\right)^2. \tag{33}$$

By using stochastic gradient descent to minimize the loss function given in (33), the training DQN can learn the correct weights of θ. The weights $\theta^{'}$ of target DQN are frozen for several time slots and then updated by copying the weights from the training DQN network, for the goal of stabilizes the training. The main steps of training procedure is given in **Algorithm 1**.

Algorithm 1 Resource Allocation in NOMA based S-IoT with DQN Algorithm.

```
Initialize experience pool to capacity $D$, initialize DQN with weight $\theta$ and target DQN with weight $\theta^{'} = \theta$, the link qualities and delay-QoS requirements of all NOMA users at the first time slot are initialized as state $s_1$.
for time $t$ in 1 to $T$ do
    Observe state information $s_t$;
    Choose an action $a_t$ by using $\epsilon$-greedy policy;
    After performing action $a_t$, agent gets reward $R(t)$, and moves to the next state $s_{t+1}$ as users moving and/or delay-QoS requirement changing;
    Stack experience tuple $(s_t, a_t, R_t, s_{t+1})$ into $D$;
    if size of tuples in pool is larger than $N_p$ then
        Sample a mini-batch of $m$ tuples from $D$;
        DQN network updates $\theta$ by minimizing the loss function given in (33) with stochastic gradient descent.
        Update target DQN network by setting $\theta^{'} = \theta$ in every $C$ steps.
    end if
end for
```

V. Numerical results

In this section, simulations are provided to evaluate the performance of the proposed resource allocation scheme and show the superiority of the NOMA based strategy in the S-IoT networks, compared to the TDMA scheme, NOMA with fixed power allocation strategy [31], and NOMA with numerical search strategy [18]. Specially, the number of users in a NOMA pair is set as 3 and they are in an ascending order and assumed to experience corresponding elevation angles, such as 20°, 50°, 70° for Users 1, 2, and 3. Moreover, we consider BT_f = 1 [16]–[19], the carrier frequency as 4 GHz, R = 125 Km, D = 1 m, G= 3.5 dBi, and G_{max} = 52.1 dBi [40]. Simulations all runs under the particular software and hardware environments of i9 CPU, 512G RAM, Win10 operating system, Spyder, and TensorFlow 2.0. The power allocation among NOMA users in the DQN algorithm simulation is trained with a set of training-target neural networks, within which 150 neurons is assumed for each hidden layer.

We first conduct numerical simulations to show the low and high SNR approximated effective capacities for each user in the NOMA based downlink S-IoT in Figs 4 and 5, respectively. Here, we adopt the fixed power allocation strategy proposed in [31] to allocate part of transmission power to user with best channel quality, i.e., User 3, and the remaining is equally divided between the first and second user. From these two figures. we can clearly see that approximated results computed by (21) and (23)–(26) all match well with the exact results, confirming the validation of the derived closed-form expressions. Meanwhile, as shown in Fig. 4, user's performance can be enhanced if more power resource is allocated and/or a loosen latency is needed. This observation clearly indicates the impact of delay-QoS requirement on each user's performance, and the urgent need of taking users' delay-QoS requirement into consideration to develop a comprehensive resource allocation strategy.

Then, we study the convergence of the DQN algorithm with different discount factor γ in Fig. 6. Specially, the number

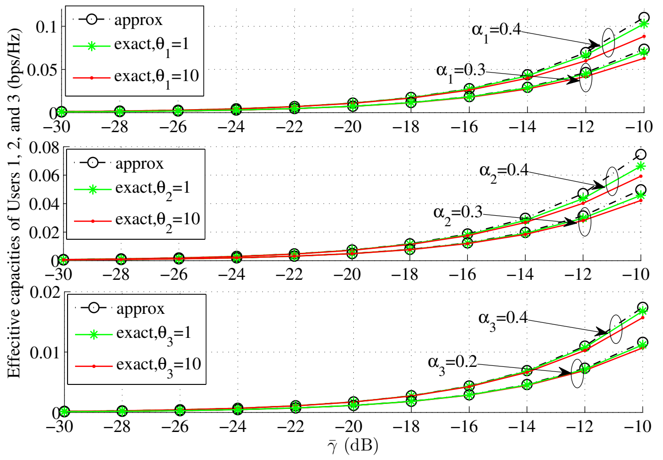

Fig. 4. Low SNR approximation of the effective capacity for each NOMA user.

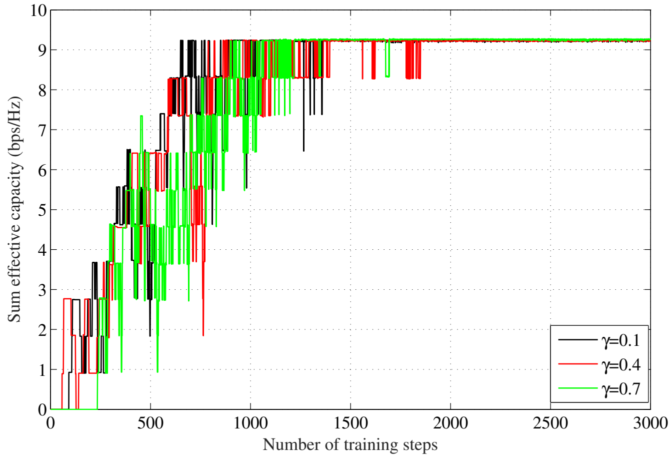

Fig. 6. Convergence of DQN in the considered system with different γ.

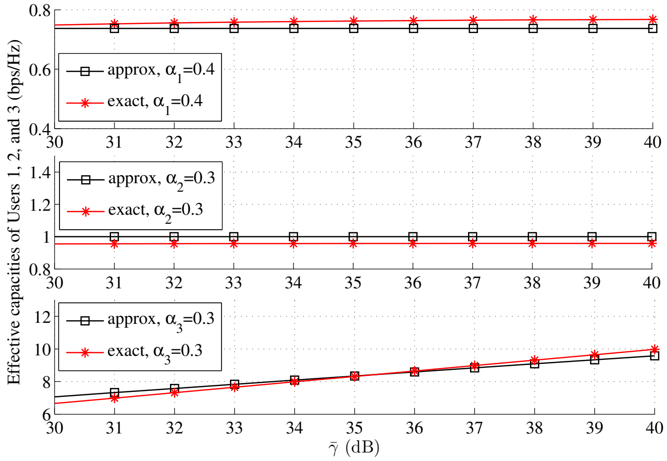

Fig. 5. High SNR approximation of the effective capacity for each NOMA user with $\theta = 1$.

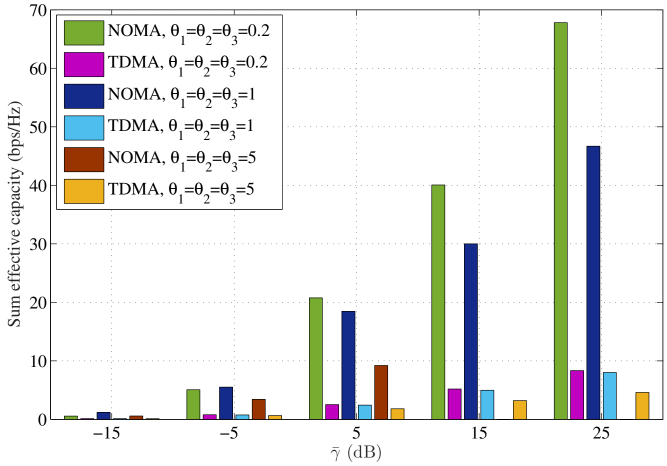

Fig. 7. Sum effective capacity of the DQN based downlink S-IoT vs $\bar{\gamma}$.

of training episode is set as 3000, time slots $T = 5$, and transmission power $P_s = 5$ dB. It can be observed that capacity curves first huge fluctuate and then become converse gradually as the training episode increases. A low discount factor, i.e., $\gamma = 0.1$, make the DQN network converse quickly, since the states are assumed to be randomly configured in each time slot and little impacts of future rewards on the cumulative discounted rewards. Thus, $\gamma = 0.1$ is considered in following simulations if without other descriptions.

The achievable sum effective capacity of the considered NOMA enabled S-IoT with DQN algorithm at different $\bar{\gamma}$ is shown in Fig. 7. As the average SNR $\bar{\gamma}$ increases, the capacity bars achieved with the NOMA and TDMA schemes increase. Moreover, the performance with the NOMA scheme under most $\bar{\gamma}$ is far larger than that achieved with the TDMA scheme. However, we find that for case $\theta_1 = \theta_2 = \theta_3 = 5$ when $\bar{\gamma} = 15$ and $\bar{\gamma} = 25$, the achievable performances with the NOMA scheme become 0. This is due to the fact that when $\bar{\gamma}$ increases, performance of each user with the TDMA scheme increases correspondingly. While the increasing of θ means more power resource is needed for each NOMA user to meet limitation set in (29), for example, sum capacity is 0 means that there has at least one user whose rate limitation is not met. Thus, under the joint impact of $\bar{\gamma}$ and θ, the number of users in a NOMA pair needs to be further studied to maximize the resource utilization efficiency.

Finally, Fig. 8 compares the sum effective capacities of users achieved with NOMA and TDMA schemes with different delay-QoS exponents. We can clearly see from this figure that curves with the NOMA scheme are superior to those with the TDMA scheme in all cases, showing the advantage of introducing the NOMA scheme in the downlink S-IoT. Besides, we find that performance degrades for both schemes as the QoS exponent increases, implying that besides channel qualities, various delay-QoS requirements of users must also be taken into account to meet the more stringent reliability/latency/data demand for future S-IoT networks.

Fig. 9 compares the achievable sum effective capacity of the considered system with different power allocation strategies. As seen in this figure, the sum capacity of the considered system is degraded when any latency requirement of these users becomes more stringent. The reason behind this phenomena is that a tighter QoS-delay constraint means a shorter

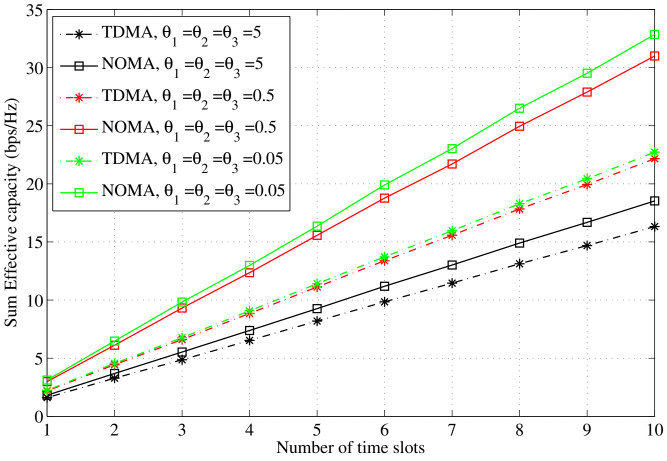

Fig. 8. Sum effective capacity comparison between the NOMA and TDMA schemes with $P_s = 5$ dB.

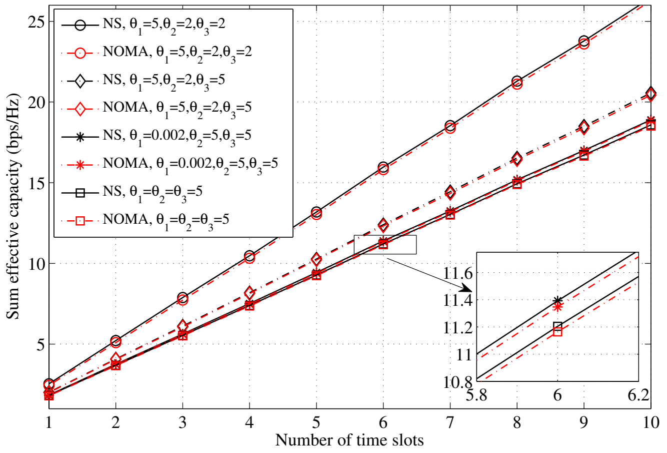

Fig. 9. Sum effective capacity comparison between NOMA with the DRL and NS power allocation strategies for $P_s = 5$ dB.

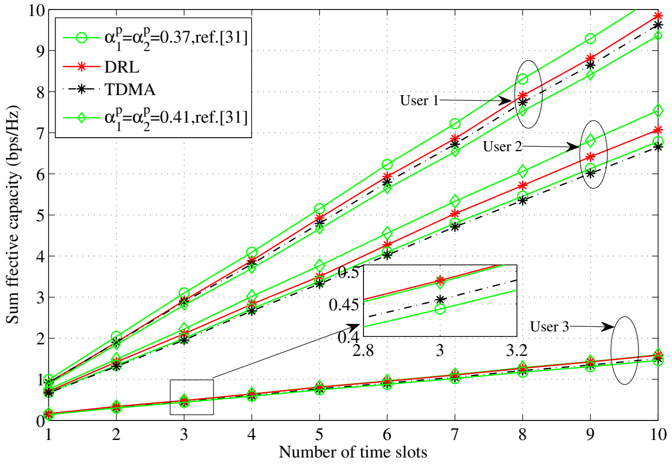

Fig. 10. Effective capacities of each users comparison between NOMA with various power allocation strategies and the TDMA scheme.

transmission delay and a lower supported constant arrival rate. Moreover, we can clearly find that capacity curves achieved with the proposed DRL based power allocation scheme closely follow with that achieved with the NS strategy [18], [19] in all time slots, which means the optimal/near-optimal action can be selected with the trained DRL algorithm. This observation indicates that the NOMA based S-IoT employing the DRL algorithm can achieve a high resource utilization efficiency but with reduced computational complexity.

Fig. 10 conducts simulations to compare the achievable effective capacity of each user with power allocation strategies proposed in this paper and ref.[31], and the TDMA scheme. Here, we set $\theta_1 = \theta_2 = \theta_3 = 5$ and $P_s = 5$ dB. It can be observed that with the strategy proposed in [31], better capacity performance can be achieved when larger power allocated to Users 1 and 2, while at the same time, capacity curve of User 3 degrades. For example, compare to $\alpha_1^p = \alpha_2^p = 0.37$, performances with the NOMA scheme of Users 1 and 2 are improved when $\alpha_1^p = \alpha_2^p = 0.41$ and better than that achieved with the TDMA scheme, while the performance of User 3 is decreased and inferior to that achieved with the TDMA. This phenomenon reveals the important effects of power allocation strategy on the NOMA based systems' performance and a flexible and optimal power allocation strategy on ensuring the rate limitation of each user. Meanwhile, the curves with the DQN based power allocation schemes are superior to those with the TDMA scheme for all users, demonstrating the advantages of employing the DRL algorithm in the NOMA based S-IoT networks.

VI. Conclusions

In this paper, we have developed a dynamic power allocation strategy for NOMA based S-IoT system with delay-QoS constraints by DRL algorithm, with which optimum/near-optimum power allocation factors are selected for NOMA users to maximize the sum effective capacity of the proposed system. Simulations have been provided to validate that, with such a mechanism, NOMA scheme is effective in encouraging spectrum sharing among multiple users to further improve system level performance although decision making is near-optimum in some conditions. This paper mainly focuses on the resource dynamic allocation of a NOMA pair within a single beam spot scenario. With the increasing S-IoT devices, we will explore the optimum resource allocation with the DRL algorithm for multiple NOMA pairs in multiple beam spots in our future work.

References

[1] C. Liu, W. Feng, Y. Chen, C. X. Wang, and N. Ge, "Cell-free satellite-UAV networks for 6G wide-area internet of things," *IEEE J. Sel. Areas Commun.*, to be published.

[2] Z. Na, M. Zhang, M. Jia, M. Xiong, and Z. Gao, "Joint uplink and downlink resource allocation for the Internet of things," *IEEE Access*, vol. 7, pp. 15758–15766, 2019.

[3] T. Li, J. Yuan, and M. Torlak, "Network throughput optimization for random access narrowband cognitive radio internet of things (NB-CRIoT)," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1436C1448, 2018.

[4] Z. Na, Y. Liu, J. Shi, *et al.*, "UAV-supported clustered NOMA for 6G-enabled internet of things: trajectory planning and resource allocation," *IEEE Internet Things J.*, to be published.

[5] J. Jiao, Y. Sun, S. Wu, *et al.*, "Network utility maximization resource allocation for NOMA in satellite-based internet of things," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3230–3242, April 2020.

[6] Y. Sun, Y. Wang, J. Jiao, S. Wu, and Q. Zhang, "Deep learning-based long-term power allocation scheme for NOMA downlink system in S-IoT," *IEEE Access*, vol. 7, pp. 86288–86296, 2019.

[7] J. Chu, X. Chen, C. Zhong, and Z. Zhang, "Robust design for NOMA-based multi-beam LEO satellite internet of things," *IEEE Internet Things J.*, to be published.

[8] Y. He, J. Jiao, X. Liang, *et al.*,"Outage performance of millimeter-wave band NOMA downlink system in satellite-based IoT," in *prob. ICCC'19*, Changchun, China, 2019, pp. 356–361.

[9] X. Yan, H. Xiao, K. An, G. Zheng, and S. Chatzinotas, "Ergodic capacity of NOMA-based uplink satellite networks with randomly deployed users," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3343-3350, Sept. 2020.

[10] X. Yan *et al.*, "The application of power-domain non-orthogonal multiple access in satellite communication networks." *IEEE ACCESS*, vol. 7, pp. 63531-63539, May 2019.

[11] X. Yan, H. Xiao, C.-X. Wang, K. An, A. T. Chronopoulos, and G. Zheng, "Performance analysis of NOMA-based land mobile satellite networks," *Proc. IEEE ACCESS*, Vol. 6, pp. 31327–31339, Jun. 2018.

[12] X. Yan, H. Xiao, K. An, G. Zheng, and W. Tao, "Hybrid satellite terrestrial relay networks with cooperative non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 978–981, May 2018.

[13] X. Yan, H. Xiao, C.-X. Wang, and K. An, "Outage performance of NOMA-based hybrid satellite-terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 538–541, Aug. 2018.

[14] X. Yan, H. Xiao, C.-X. Wang, and K. An, "On the ergodic capacity of NOMA-based cognitive hybrid satellite terrestrial networks," in *Proc. IEEE ICCC'17*, Qingdao, China, 2017, pp. 1-5.

[15] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE VTC'13*, Dresden, Germany, 2013, pp. 1–5.

[16] Y. Ruan, Y. Li, C. Wang, R. Zhang, and H. Zhang, "Energy efficient power allocation for delay constrained cognitive satellite terrestrial networks under interference constraints," *IEEE Trans. Wireless. Commun.*, vol. 18, no. 10, pp. 4957-4969, Oct. 2019.

[17] Y. Ruan, Y. Li, C. Wang, R. Zhang, and H. Zhang, "Effective capacity analysis for underlay cognitive satellite-terrestrial networks," in *Proc. IEEE ICC'17*, Paris, France, 2017, pp. 1-6.

[18] V. Kumar, B. Cardiff, S. Prakriya, and M. F. Flanagan, "Effective rate of downlink NOMA over κ-μ shadowed fading with integer fading parameters," in *Proc. IEEE ICC'20*, Dublin, Ireland, 2020, pp. 1-7.

[19] V. Kumar, B. Cardiff, S. Prakriya, and M. F. Flanagan, "Link-layer capacity of downlink NOMA with generalized selection combining receivers,," in *Proc. IEEE ICC'20*, Dublin, Ireland, 2020, pp. 1-7.

[20] X. Yan, K. An, C. X. Wang, *et al.*, "Genetic algorithm optimized support vector machine in NOMA-based satellite networks with imperfect CSI," in *Proc. IEEE ICASSP'20*, Barcelona, Spain, 2020, pp. 8817-8821.

[21] L. Lei, E. Lagunas, Y. Yuan, M. G. Kibria, S. Chatzinotas, and B. Ottersten, "Deep learning for beam hopping in multibeam satellite systems," in *Proc. IEEE VTC'20*, Antwerp, Belgium, 2020, pp. 1-5.

[22] Y. Song, B. Song, Z. Zhang, and Y. Chen, "The satellite downlink replanning problem: a BP neural network and hybrid algorithm approach for IoT internet connection," *IEEE Access*, vol. 6, pp. 39797-39806, July 2018.

[23] N. Kato, Z. M. Fadlullah, F. Tang, B. Mao, et al., "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wirel. Commun.*, vol. 26, no. 4, pp. 140–147, Aug. 2019.

[24] C. Jiang and X. Zhu, "Reinforcement learning based capacity management in multi-layer satellite networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4685-4699, July 2020.

[25] B. Zhao, J. Liu, Z. Wei, and I. You, "A deep reinforcement learning based approach for energy-efficient channel allocation in satellite internet of things," *IEEE Access*, vol. 8, pp. 62197-62206, Mar. 2020.

[26] C. Han, L. Huo, X. Tong, H. Wang, and X. Liu, "Spatial anti-jamming scheme for internet of satellites based on the deep reinforcement learning and stackelberg game," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5331–5342, May 2020.

[27] B. Deng, C. Jiang, H. Yao, S. Guo, and S. Zhao, "The next generation heterogeneous satellite communication networks: integration of resource management and deep reinforcement learning," *IEEE Wirel. Commun.*, vol. 27, no. 2, pp. 105–111, April 2020.

[28] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach," in *Proc. IEEE ICC'19*, Shanghai, China, 2019, pp. 1-6,

[29] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255-6267, Oct. 2020.

[30] Dapeng Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," in *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[31] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[32] T. Li, H. Zhou, H. Luo, Q. Xu, and Y. Ye, "Using SDN and NFV to implement satellite communication networks," in *Proc. IEEE NaNA'16*, Hakodate, Japan, 2016, pp. 131-134.

[33] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York, NY, USA: Academic, 2007.

[34] 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.

[35] K. An *et al.*, "Outage performance of cognitive hybrid satellite-terrestrial networks with interference constraint," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9397-9404, Nov. 2016.

[36] K. Guo *et al.*, "On the Performance of the Uplink Satellite Multiterrestrial Relay Networks With Hardware Impairments and Interference," *IEEE Systems Journal*, in press.

[37] A. Abdi, W. Lau, M.-S. Alouini, and M. Kaveh, "A new simple model for land mobile satellite channels: first and second order statistics," *IEEE Trans. Wireless. Commun.*, vol. 2, no. 3, pp. 519–528, May 2003.

[38] E. Lutz, M. Werner, and A. Jahn, *Satellite systems for personal and broadband communications*, Springer, Berlin, Germany, 2000.

[39] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for low-latency vehicular communications: an effective capacity perspective," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 905-917, April 2019.

[40] W. Lu, K. An, and T. Liang, "Robust beamforming design for sum secrecy rate maximization in multibeam satellite systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1568–1572, Jun. 2019.