Article

User Pairing for Delay-Limited NOMA-Based Satellite Networks with Deep Reinforcement Learning

Qianfeng Zhang [1,2], Kang An [3], Xiaojuan Yan [1,4,*], Hongxia Xi [1] and Yuli Wang [1]

1   Guangxi Key Laboratory of Ocean Engineering Equipment and Technology, Qinzhou 535011, China; qfzhang19@163.com (Q.Z.); xi15007778221@163.com (H.X.); wangyuli95@163.com (Y.W.)
2   Key Laboratory of Beibu Gulf Offshore Engineering Equipment and Technology (Beibu Gulf University), Education Department of Guangxi Zhuang Autonomous Region, Qinzhou 535011, China
3   Sixty-Third Research Institute, National University of Defense Technology, Nanjing 210007, China; ankang@nuaa.edu.cn
4   School of Information Science and Engineering, Southeast University, Nanjing 210096, China
*   Correspondence: yxj9609@163.com

Abstract: In this paper, we investigate a user pairing problem in power domain non-orthogonal multiple access (NOMA) scheme-aided satellite networks. In the considered scenario, different satellite applications are assumed with various delay quality-of-service (QoS) requirements, and the concept of effective capacity is employed to characterize the effect of delay QoS limitations on achieved performance. Based on this, our objective was to select users to form a NOMA user pair and utilize resource efficiently. To this end, a power allocation coefficient was firstly obtained by ensuring that the achieved capacity of users with sensitive delay QoS requirements was not less than that achieved with an orthogonal multiple access (OMA) scheme. Then, considering that user selection in a delay-limited NOMA-based satellite network is intractable and non-convex, a deep reinforcement learning (DRL) algorithm was employed for dynamic user selection. Specifically, channel conditions and delay QoS requirements of users were carefully selected as state, and a DRL algorithm was used to search for the optimal user who could achieve the maximum performance with the power allocation factor, to pair with the delay QoS-sensitive user to form a NOMA user pair for each state. Simulation results are provided to demonstrate that the proposed DRL-based user selection scheme can output the optimal action in each time slot and, thus, provide superior performance than that achieved with a random selection strategy and OMA scheme.

Keywords: NOMA-based satellite networks; delay QoS limitation; user pairing

1. Introduction

Due to the inherent nature of providing vast coverage and economic service, satellite communication has the ability to effectively supplement terrestrial networks during disasters and in rural and deserts areas; thus, it has been considered as an important component for next-generation wireless networks [1]. However, the dramatically increased demand for data access can result in even bigger challenges, including massive connectivity, limited power/spectral resources, and various quality of service (QoS) requirements, in future satellite networks. Recently, non-orthogonal multiple access (NOMA) schemes, including power domain NOMA [2] and code domain NOMA [3], featuring multiple access, high resource utilization efficiency, and user fairness, has become a promising solution to alleviate these challenges faced by future satellite networks. Of these two schemes, power domain NOMA (or simply NOMA for short) scheme, which has the ability to harmoniously integrate with orthogonal multiple access (OMA) techniques in existing satellite architectures, is the main motivation and focus of this article.

In a NOMA-based satellite network, a satellite/multiple users can simultaneously communicate with multiple users/a satellite in downlink/uplink transmissions by super-

posing various signals with different power levels in the same time/spectral block. To date, many works have investigated the performance enhancement of various NOMA-based satellite networks, such as the improved outage probability in integrated satellite terrestrial relay networks with perfect successive interference cancellation (SIC) [4], augmented erogodic capacity in uplink satellite communications [5], and increased network utility in the satellite-based internet of things [6]. An extension of work [4] to an imperfect SIC scenario with Alamouti space–time block coding was studied in [7]. Moreover, some works studied resource management of the NOMA-based satellite network from the perspective of increasing system resource efficiency, e.g., aiming to maximize long-term age of information, the authors in [8] utilized a ListNet algorithm and a particle swarm optimization algorithm to obtain an optimized power allocation solution in a satellite-based internet of things scenario. Similarly, the work in [9] proposed a joint subchannel assignment and power allocation algorithm to further optimize the sum rate of a secondary network in cognitive satellite–unmanned aerial vehicle–terrestrial networks. Although NOMA-based satellite networks can enhance spectrum/power utilization efficiency and a system's performance, we must note that these Shannon performance enhancements were achieved by selecting users with distinctive differences in channel gains to form a NOMA pair, which is only suitable for a system with delay-insensitive applications.

However, with rising technological developments in wireless communications, new satellite applications with diverse delay QoS requirements have occurred to facilitate our daily life and provide more efficient service, such as applications on smart grids, environmental monitoring and forecasting, navigation, smart cities, and telemedicine. Among these applications, telemedicine and smart grids are identical delay-critical scenarios, while environmental monitoring is a typical delay-tolerant scenario. Thus, Shannon capacity, which fails to take users' diverse delay QoS requirements into consideration, is no longer suitable to use in future satellite networks to characterize the performance of real-time and delay-sensitive applications/scenarios, and it is of paramount importance to study the achievable performance of satellite networks under heterogenous delay QoS requirements. Under these conditions, the concept of effective capacity, which was proposed in [10] as an effective performance metric to show the maximum constant arrival rate with a given delay QoS constraint, has been introduced in various satellite communication scenarios to show the adverse impact of delay QoS limitations on system performance [11–14], such as the authors in [11], who proposed an algorithm to schedule users in different time slots while guaranteeing users' delay QoS requirements in a satellite–terrestrial backhaul network. In cognitive satellite–terrestrial networks, the effective capacity was introduced to guarantee the delay requirement of a primary user [12], whose extension to study effective energy efficiency of the same networks was studied in [13]. Moreover, work [14] studied the achieved effective capacity of a NOMA-based satellite system with delay adhering to users' service requirements. Although these aforementioned works have shown the negative impact of delay QoS requirements on OMA-/NOMA-based satellite networks, how to select users in a NOMA based system, i.e., whether it is effective to only select users with big channel differences, to form a NOMA pair has not been investigated.

It is worth noting that, in addition to free space loss (FSL), antenna gain, fading severity, and location information in a beam spot can also influence the link budget of a satellite user, all of which, combined with users' various delay QoS requirements, make the user grouping in a NOMA-based system nontrivial, especially in satellite networks, which are highly applied in military and civilian fields. To solve this challenge, a supervised learning algorithm, with which solutions can be obtained without model-oriented analysis and design, as an effective solution for resource management has been widely used in several prior works, such as work [15], which proposed a genetic algorithm (GA)-improved support vector machine scheme to effectively pair users for NOMA-based satellite networks. A fully connected deep neural network-assisted approach was studied in [16,17] to facilitate efficient beam hopping and design beam illumination pattern in multibeam satellite systems, respectively. The work in [18] proposed an accurate forecasting method

by using deep neural networks for LEO satellite links. Notably, supervised learning, such as the algorithms used in [15–18], needs to learn characteristics from input data and desired output data, while a reinforcement learning (RL) algorithm, which is model-free and data-driven, has been extensively adopted in various wireless networks with different objectives. For example, based on *Q*-learning, an algorithm for jointly optimizing user pairing and power allocation was proposed in [19] to maximize the total sum rate of a satellite random access system. Considering large-scale low-earth orbit constellations, the work in [20] developed a low-complexity successive deep Q-learning algorithm for optimal satellite handover. The authors in [21] proposed a *Q*-learning NOMA-based random access scheme for time slot and channel allocation in satellite–terrestrial relay networks. In [22], the authors adopted a graph neural network and RL algorithms in a hybrid satellite–terrestrial network to optimize UAV trajectory and maximize the number of served users. In [23,24], the authors conducted resource management in a relay-aided network with the help of distributionally robust deep RL (DRL) and enhanced DRL algorithms, respectively.

Motivated by these observations, for the work herein, we leaned upon a DRL algorithm to pair users and provide services with various delay QoS requirements for future NOMA-based satellite networks (since this paper's aim was to pair users in delay-limited NOMA-based satellite networks with a DRL algorithm, while the impacts of low-density parity check codes [25] in NOMA-based satellite networks will be our follow-up research.). The main contributions of this work can be described as follows:

- The concept of effective capacity is employed to measure the rate achieved with a given delay QoS constraint, based on which, a power allocation coefficient is firstly obtained by ensuring the achieved capacity of users with sensitive delay QoS requirements is not less than that achieved with an OMA scheme, and then, the user pairing problem is formulated with the aim of maximizing the sum effective capacity of the considered system;

- Because various delay QoS requirements have varying negative impacts on users' capacity, user pairing in a NOMA-based network with various delay QoS constraints is different from that in traditional NOMA-based delay-insensitive system. In this condition, to maximize system capacity with the obtained power allocation factor, when the delay-critical user is fixed, a DRL approach is introduced to select one user who has relatively insensitive delay requirement and good link condition, compared to the other users, to optimize NOMA user pairing with low complexity;

- The proposed DRL-based NOMA user pairing strategy is compared to an OMA scheme and NOMA with a random user-selecting scheme, which reveal the superiority of introducing the NOMA scheme and DRL algorithm in the satellite networks from the perspective of performance enhancement. Specifically, the advantage of the proposed approach is achieved by selecting the most suitable delay tolerant user to pair with the delay-sensitive user and form a NOMA user group in each time slot.

The rest of this paper is outlined as follows. The system model is presented in Section 2. Section 3 introduces the concept of effective capacity, obtains the power allocation scheme by ensuring the achieved capacity of the user with sensitive delay QoS requirement is not less than that achieved with the OMA scheme, and formulates the user pairing problem for the delay-limited NOMA-aided satellite network. In Section 4, a DRL algorithm is described in detail and tested in the proposed system. Performance results are discussed and conclusions are given in Sections 5 and 6, respectively.

2. System Model

Consider a downlink NOMA-based satellite system that is designed to serve m ($m \geq 2$) users with the help of the NOMA scheme. These m users are randomly deployed in an area approximated as a circle of radius R with different channel statistical prosperities and delay QoS requirements. (In this paper, channel estimation errors, co-channel interference, complexity, and mobility constraints are not taken into consideration in the proposed system model; the influences of these parameters on user selection and system performance

will be a focus in our future works, based on the contributions in the current work.) Without loss of generality, users are ordered based on their link budgets, i.e., $Q_1 \leq Q_2 \cdots \leq Q_m$, where Q_j is the link budget of User j ($j = 1, 2, \cdots, m$). For simplicity, we further assume only the cth and t^{th} users ($1 \leq c < t \leq m$) are selected to form a NOMA group, and each user in the proposed model is equipped with a single antenna.

Thus, the received signal at User j ($j = c, t$) is

$$y_j = Q_j x + w_j, \tag{1}$$

where w_j denotes the noise at User j with zero mean and δ^2 variance, $x = \sum_{j=c,t} \sqrt{\alpha_j^p P_s} x_j$ is the superposed signal (with α_j^p being a fraction of the transmission power P_s allocated to User j and x_j ($E[|x_j|^2] = 1$) being the signal for User j), Q_j (including FSL, antenna gain, beam gain, and fading model) is the entire link budget from satellite to User j, which can be described as follows:

$$Q_j = \Phi_j G_s(\varphi_j)|g_j|^2, \tag{2}$$

where $\Phi_j = L_j G_j$, with L_j and G_j being the FSL and antenna gain at User j, respectively. $G_s(\varphi_j)$, which is the beam gain of User j, with φ_j denoting the angle between User j and beam center with respect to the satellite, can be approximated as [5]

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

with $G_{\max}$ representing the maximum antenna gain, $J_n(\cdot)$ being the Bessel function of first kind and n-th order, d_j being the distance from the beam center to User j, and $a = 2.07123/R$. $|g_j|^2$ is the channel power gain of the satellite link, which is assumed to follow a widely applied Shadowed Rician fading model [26–30]. According to [31], the probability density function (PDF) of $|g_j|^2$ is

$$f_{|g_j|^2}(x) = \alpha_j e^{-\beta_j x}{}_1F_1(m_j; 1; \delta_I x), \tag{4}$$

where $\alpha_j = \frac{(2b_j m_j)^{m_j}}{2b_j(2b_j m_j + \Omega_j)^{m_j}}$, $\delta_j = \frac{\Omega_j}{2b_j(2b_j m_j + \Omega_j)}$, $\beta_j = \frac{1}{2b_j}$ with $2b_j$ and Ω_j, respectively, being the average power of the multipath and the LoS components, m_j ($m_j > 0$) denoting the Nakagami-m fading parameter, and ${}_1F_1(a; b; c)$ representing the confluent hypergeometric function ([32], Equation (9.14.1)).

Based on the principle of the downlink NOMA scheme, decoding order is decided by users' channel qualities, i.e., the user with a worse link condition decodes its own information firstly and directly. Thus, the signal-to-interference-plus-noise ratio (SINR) of User c is

$$\gamma_c^{\mathrm{N}} = \frac{\alpha_c^p \gamma \Phi_c G_s(d_c)|g_c|^2}{\alpha_t^p \gamma \Phi_c G_s(d_c)|g_c|^2 + 1} = \frac{\alpha_c^p \gamma Q_c}{\alpha_t^p \gamma Q_c + 1}, \tag{5}$$

where $\alpha_c^p + \alpha_t^p = 1$ and $\gamma = P_s/\delta^2$ is the average transmission SNR. At the same time, the user with better channel quality, i.e., User t, adopts the SIC strategy to decode and remove the interference from User c; the decoding SINR can be derived as

$$\gamma_{t \to c}^{\mathrm{N}} = \frac{\alpha_c^p \gamma \Phi_t G_s(d_t)|g_t|^2}{\alpha_t^p \gamma \Phi_t G_s(d_t)|g_t|^2 + 1} = \frac{\alpha_c^p \gamma Q_t}{\alpha_t^p \gamma Q_t + 1}. \tag{6}$$

We can derive that $\gamma_c^{\mathrm{N}} < \gamma_{t \to c}^{\mathrm{N}}$, since $Q_c < Q_t$. Then, User t decodes its own information, and the achieved SINR is

$$\gamma_t^{\mathrm{N}} = \alpha_t^p \gamma \Phi_t G_s(d_t) |g_t|^2 = \alpha_t^p \gamma Q_t. \tag{7}$$

3. Effective Capacity and Power Allocation

3.1. Effective Capacity

To provide services with different delay QoS requirements, the concept of effective capacity is employed to characterize the effect of delay QoS limitation on achieved performance, characterized by θ ($\theta \geq 0$) [10]. In this paper, the uncorrelated service process across different slots is further assumed and the normalized effective capacity is adopted. Under these conditions, given a delay QoS exponent θ_j, the normalized effective capacity of User j in bps/Hz is

$$C_j(\theta_j) = \frac{-1}{\theta_j T_f B} \ln\left(\mathcal{E}\left\{e^{-\theta_j T_f B R_j}\right\}\right) = \frac{1}{\psi_j \ln 2} \ln\left(\mathcal{E}\left[(1 + \gamma_j)^{\psi_j}\right]\right), \tag{8}$$

where $\psi_j = -\theta_j T_f B / \ln 2$, with T_f and B being the frame duration and the occupied bandwidth, respectively, $R_j = \log_2(1 + \gamma_j)$ is User j's transmission rate, and $\mathcal{E}$ is the expectation operator. We note that a larger/smaller delay QoS exponent θ_j is required in a more critical/tolerant delay-limited scenario.

3.2. Power Allocation Strategy

To ensure the capacity achieved by the user with a critical delay QoS requirement using the NOMA scheme is always better than that with the TDMA scheme, the power allocation coefficient should be further constrained. In this section, a power allocation scheme is investigated for two cases, i.e., User c in Case 1 and User t in Case 2 are assumed to be delay-sensitive users.

For Case 1, $\theta_c > \theta_t$ is assumed and the power allocation factor is limited by $C_c^{\mathrm{N}}(\theta_c) \geq C_c^{\mathrm{T}}(\theta_c)$, where

$$C_c^{\mathrm{N}}(\theta_c) = \frac{1}{\psi_c \ln 2} \ln\left(\mathcal{E}\left[\left(1 + \gamma_c^{\mathrm{N}}\right)^{\psi_c}\right]\right), \tag{9}$$

and

$$C_c^{\mathrm{T}}(\theta_c) = \frac{1}{\psi_c \ln 2} \ln\left(\mathcal{E}\left[\left(1 + \gamma_c^{\mathrm{T}}\right)^{0.5\psi_c}\right]\right), \tag{10}$$

with $\gamma_j^{\mathrm{T}} = \gamma \Phi_j G_s(d_j) |g_j|^2 = \gamma Q_j$ being the SINR of User j ($j = c, t$) achieved with the TDMA scheme, and 0.5 owes to the loss in multiplexing in the TDMA system. By substituting (5) into (8), along with some manipulations, α_c^p can be derived as

$$\alpha_c^p \geq 1 - \frac{1}{\sqrt{1 + \gamma Q_c} + 1}, \tag{11}$$

which means that the value of α_c^p is decided by γ, location information, and fading severity of User c.

For Case 2, $\theta_t > \theta_c$ is considered, and factor α_c^p is limitied by restriction condition $C_t^{\mathrm{N}}(\theta_t) \geq C_t^{\mathrm{T}}(\theta_t)$, with

$$C_t^{\mathrm{N}}(\theta_t) = \frac{1}{\psi_t \ln 2} \ln\left(\mathcal{E}\left[\left(1 + (1 - \alpha_c^p)\gamma Q_t\right)^{\psi_t}\right]\right), \tag{12}$$

and

$$C_t^{\mathrm{T}}(\theta_t) = \frac{1}{\psi_t \ln 2} \ln\left(\mathcal{E}\left[(1 + \gamma Q_t)^{0.5\psi_t}\right]\right). \tag{13}$$

Then, we can obtain

$$\alpha_c^p \leq 1 - \frac{1}{\sqrt{1 + \gamma Q_t} + 1}. \tag{14}$$

Based on the power allocation coefficient obtained in (11) for Case 1 or (14) for Case 2, the effective capacity of User c can be given by

$$
\begin{aligned}
C_c^{\mathrm{N}}(\theta_c) &= \frac{1}{\psi_c \ln 2} \ln\left(\mathcal{E}\left[\left(1 + \gamma_c^{\mathrm{N}}\right)^{\psi_c}\right]\right) \\
&= \frac{1}{\psi_c \ln 2} \ln\left(\int_{\mathrm{R}_{jn}}^{\mathrm{R}_{jf}} \int_0^\infty \left(1 + \gamma_c^{\mathrm{N}}\right)^{\psi_c} f_{|g_c|^2}(x) f_{d_c}(y) \, dx \, dy\right),
\end{aligned} \tag{15}
$$

where $f_d(y) = \frac{2y}{\mathrm{R}_{jf}^2 - \mathrm{R}_{jn}^2}$ is the PDF of User j's location [4] if it distributes in an annular area with inner radius R_{jn} and outer radius R_{jf}. To evaluate (15), we first express $_1F_1(m_j; 1; \delta_j x)$ in (4) and $(1 + x)^a$ in terms of the Meijer G-functions from Equation (9.34.8) in [32] and binominals represented by Equation (1.11) in [32], as

$$_1F_1(m_j; 1; \delta_j x) = \frac{1}{\Gamma(m_j)} G_{1,2}^{1,1}\left[-\delta_j x \left|\begin{array}{c} 1 - m_j \\ 0, 0 \end{array}\right.\right], \tag{16}$$

and

$$(1 + x)^a = \sum_{k=0}^\infty \frac{\Gamma(a + k)}{k! \Gamma(a)} x^k, \tag{17}$$

where $G_{1,2}^{1,1}[\cdot|\cdot]$ ([32], Equation (9.301)) is the Meijer-G function and $\Gamma(\cdot)$ ([32], Equation (8.310.1)) is the Gamma function. Then, inserting (4), (5), (16)–(17) into (15) along with ([32], Equation (7.813.1)), we obtain the result as

$$
\begin{aligned}
C_c^{\mathrm{N}}(\theta_c) = \frac{1}{\varphi_c \ln 2} \ln\Bigg(&\alpha_c \sum_{k=0}^\infty \sum_{m=0}^\infty \frac{\left(1 - \alpha_c^p\right)^k \Phi_c^{m+k} \gamma^{m+k} G_{\max} \Gamma(m + \varphi_c) \Gamma(k - \varphi_c)}{\beta_c^{k+m+1} k! m! \Gamma(m_c) \Gamma(-\varphi_c) \Gamma(\varphi_c)} \\
&\times G_{2,2}^{1,2}\left[-\frac{\delta_c}{\beta_c} \left|\begin{array}{c} -k - m, 1 - m_c \\ 0, 0 \end{array}\right.\right] \int_{R_{cn}}^{R_{cf}} \left(\frac{J_1(ay)}{2ay} + \frac{J_3(ay)}{a^3 y^3}\right)^2 \frac{2y}{R_{cf}^2 - R_{cn}^2} \, dy\Bigg).
\end{aligned} \tag{18}
$$

By further defining Ψ_c to denote the integration part of (18) and, with the help of Equation (8.442.2) in [32], we obtain

$$\Psi_c = \sum_{n=0}^\infty \Theta(n) \frac{(-1)^n a^{2n} \left(R_{cf}^{2n+2} - R_{cn}^{2n+2}\right)}{n! 4^n (2n + 2) \left(R_{cf}^2 - R_{cn}^2\right)}, \tag{19}$$

where

$$\Theta(n) = \frac{F(-n, -1 - n; 2; 1)}{16 \Gamma(2 + n)} + \frac{F(-n, -1 - n; 4; 1)}{8 \Gamma(4) \Gamma(2 + n)} + \frac{F(-n, -3 - n; 4; 1)}{32 \Gamma(4) \Gamma(4 + n)}, \tag{20}$$

with $F(a, b; c; d)$ being the hypergeometric function ([32], Equation (9.100)). Finally, substituting (19) and (20) into (18), the desired result for the expression of $C_c^{\mathrm{N}}(\theta_c)$ can be obtained.

Similarly, the effective capacity of User t can be given by

$$C_t^{\mathrm{N}}(\theta_c) = \frac{1}{\psi_t \mathrm{ln}2} \ln\left(\int_{\mathrm{R}_{tn}}^{\mathrm{R}_{tf}} \int_0^\infty \left(1 + \gamma_t^{\mathrm{N}}\right)^{\psi_t} f_{|g_t|^2}(x) f_{d_t}(y) dx dy \right). \tag{21}$$

By substituting (4) and (7) into (21) and following with the similar steps as those in the derivation of (12), the effective capacity expression of User t can be derived as

$$C_t^{\mathrm{N}}(\theta_t) = \frac{1}{\varphi_t \ln 2} \ln\left[\sum_{k=0}^\infty \frac{\alpha_t \Gamma(\varphi_t + k)\left(\alpha_t^p \bar{\gamma}\Phi_t\right)^k}{k!\Gamma(\varphi_t)\Gamma(m_t)\beta_t^{k+1}} G_{2,2}^{1,2}\left(\frac{-\delta_t}{\beta_t} \bigg| \begin{matrix} -k, 1 - m_t \\ 0, 0 \end{matrix}\right) \right.$$
$$\left. \times \sum_{n=0}^\infty \Theta(n) \frac{(-1)^n a^{2n}\left(R_{cf}^{2n+2} - R_{cn}^{2n+2}\right)}{n!4^n(2n+2)\left(R_{cf}^2 - R_{cn}^2\right)} \right]. \tag{22}$$

Then, the sum effective capacity of the considered system can be given as $C^{\mathrm{N}} = C_c^{\mathrm{N}}(\theta_c) + C_t^{\mathrm{N}}(\theta_t)$.

3.3. Problem Formulation

Although the closed-form expression of sum rate for the considered system has been derived, we must note that the rate of User j ($j = c, t$) is influenced by many factors, such as delay exponent θ_j, transmission SNR γ, fading severity, location information d_j, and α_j^p. Thus, to expressively show the different impacts of these key parameters on the achieved performance, the normalized effective capacity of User j is plotted in Figure 1, where ILS, AS, and FHS are infrequent light shadowing, average shadowing, and frequent heavy shadowing, respectively.
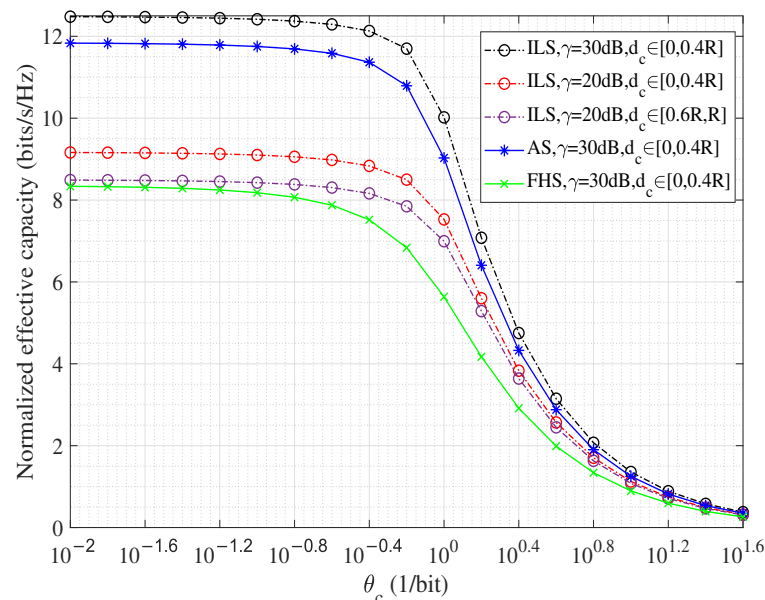
Figure 1. Normalized effective capacity versus delay exponent θ_j for various SNR γ, fading severity, and location information d_j, when $\alpha_j^p = 1$.

From Figure 1, we can directly observe that, when $\theta_j \to 0$, effective capacity converges to the ergodic capacity, since only delay-insensitive traffic is needed. However, when $\theta_j > 10$, even for case $\alpha_j^p = 1$, effective capacity reduces to 0 due to the required delay QoS being too stringent. Thus, the range of User j's delay limitation is assumed to be constrained as $\theta_j \in [0.5, 10]$ in this paper. In addition, an increased d_j, i.e., a worse fading severity, or decreased γ can degrade the capacity curves. Moreover, all capacity curves decrease

with increasing θ_j. This observation clearly indicates that the achieved performance suffers from a combination of these factors, although, in both Case 1 and Case 2, it seems like a user with the smallest delay QoS exponent, nearest location information, and best shadowing should be selected as User t/c and paired with the User c/t in Case 1/2 to maximize the sum performance of the considered system. Conversely, while in a spot beam, the user with the nearest location information or best fading condition may have a relatively large θ, or vice versa. Thus, how to select User t/c in Case 1/2 is a vital issue in a delay-limited scenario.

For simplicity, herein, we mainly focus on the user pairing in Case 1, which means that α_c^p must meet $C_t^N(\theta_t) \geq C_t^T(\theta_t)$. Then, the optimization problem is to find a user who can obtain the best power utilization efficiency, after taking into account link budget and delay QoS requirement, to be the User t. The mathematical formulation of this problem can be denoted by P1 and formulated as

$$P1 : \max_{\{d_t, Q_t, \theta_t\}} C_t^N(\theta_t)$$

$$s.t. \quad C1 : Q_t > Q_c, t \in 1, 2, \cdots, c-1, c+1, \cdots, m;$$
$$C2 : (11), \theta_c > \theta_t > 0;$$
$$C3 : d_{c,t} \leq R. \tag{23}$$

In the aforementioned problem, C1 ensures that the link budget of User t must be better than that of User c to successfully perform SIC; C2 denotes that, in Case 1, the resource allocation threshold in (11) must be ensured to guarantee the minimum data rate requirement of User c, and C3 implies that the limited location information of Users c and t.

4. DRL for Delay-Constrained User Pairing

The deep Q-network (DQN) algorithm, which combines the advantages of Q-learning and deep neural networks, is one of the most representative value-based method in the DRL family, with which the expected returns of actions can be predicted based on a certain environmental observation; a framework of applying such an algorithm in user pairing for the considered system is provided in Figure 2. (The DQN method is the classical approach in the DRL family, whose complexity analysis is not provided in this paper—the interested reader can refer to [33].) Although DRL deployment causes more delay, it is believed that this delay can be significantly decreased with the improvement of chip processing speed).

Since our objective in problem P1 is to choose an appropriate user to be User t at different time slots to maximize the power resource utilization, we thus define a tuple $\bar{M} := < S, A, R, \pi >$ to model this problem as a Markov decision process (MDP) for a stationary decision. Specifically, S means the state and observation space, A represents the set of actions, R means the designed reward, and π is the policy that makes the decision. Meanwhile, $Q_\pi(s_l, a_l)$ is defined as the Q-value obtained with policy π when the environment is in state s_l while adopting action a_l at the lth time slot. For the problem P1, key elements, such as the states, actions, and cost, in an MDP model are described in detail as follows:

- State S: At time slot l, a tuple denoted by $s_l = (P_s, \Phi_j, d_j, g_j, \theta_j)$, $s_l \in S$ is used to describe the system state, where $P_s, \Phi_j, d_j, g_j, \theta_j$ are transmission power, antenna gains, location information, fading severity, and delay QoS exponent of User j ($j = 1, 2, \cdots, c-1, c+1, \cdots, m$), as analyzed in Sections 2 and 3, respectively. Since s_l varies in different time slots, the agent is required to adjust its action in each slot accordingly;

- Action A: NOMA user pairing is important for NOMA-aided satellite networks with delay QoS constraints because it directly impacts the resource utilization efficiency. Thus, user selection should be designed based on current state; here, we set the action space as $[A = 1, 2, \cdots, c-1, c+1, \cdots, m]$, and then $a_l = m$ means the mth user is selected to be the User t;

- Reward design: Equation (11) must be satisfied to ensure that User *c*'s performance achieved with the NOMA scheme is not less than that achieved with the TDMA scheme. Based on this, our objective is to select a user to be User *t* who, with the remaining power resource, can achieve the largest effective capacity. Thus, if User *j* is selected at time slot *l*, the reward is assigned as

$$C_j^N(\theta_j, l) = C_j^N(\theta_j), (j = 1, 2, \cdots, c-1, c+1, \cdots, m) \tag{24}$$

As can be seen from Figure 2, the DQN algorithm has two phases. In the data-generation phase, *Q*-learning with experience pool *D* is used to generate data for the next network-training phase. In this process, the agent chooses an action a_l according to its observation s_l under policy π. To trade off between exploration and exploitation, ε-greedy exploration is used here, which means, for state s_l, a random action with probability ϵ ($0 < \epsilon < 1$) or the best action with probability $(1 - \epsilon)$ is chosen to be action a_l. With this ε-greedy policy, *Q*-value function $Q_\pi(s_l, a_l)$, which describes the expected $R_\pi(l)$, can be given by

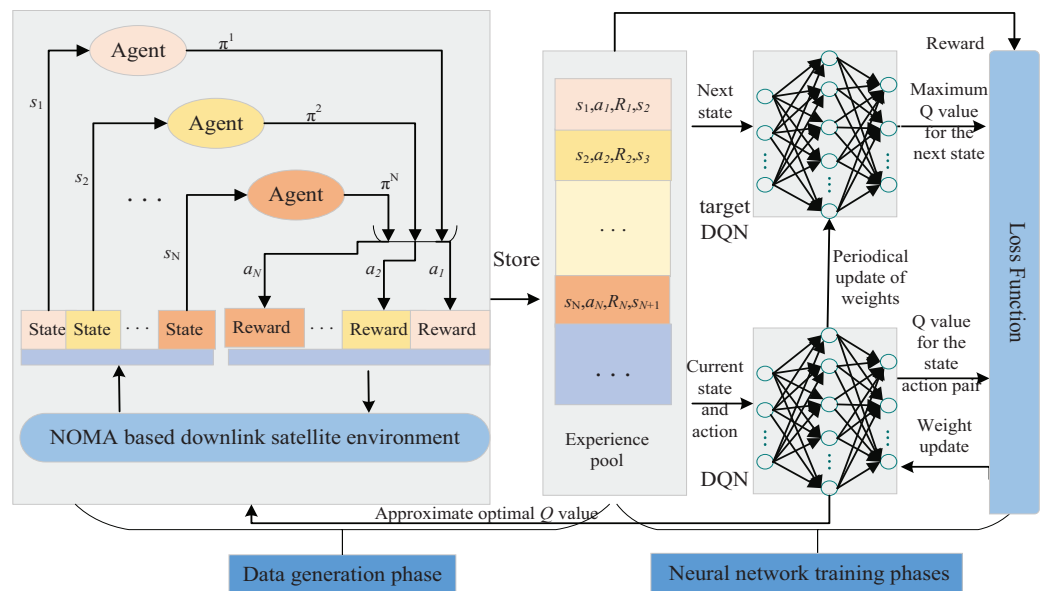$$Q_\pi(s_l, a_l) = E(R_\pi(l)|S = s_l, A = a_l). \tag{25}$$

Figure 2. DQN-based NOMA user pairing model.

This *Q*-value function is updated with

$$Q_\pi(s_l, a_l) = Q_\pi(s_l, a_l)(1 - \bar{\alpha}) + \bar{\alpha}\left(R_\pi(l) + \hat{\gamma} \max_{a_{l+1}} Q_\pi(s_{l+1}, a_{l+1})\right), \tag{26}$$

where $\bar{\alpha}$ and $\hat{\gamma}$ are the learning rate and discount factor, respectively. The best action can be written as $a_l^\star = \max_{a_l \in A} Q_\pi(s_l, a_l)$. Following the environmental transition resultant from variations in users' link budgets and delay QoS limitations, the tuple (s_l, a_l, R_l, s_{l+1}) at the lth time slot is collected and stored in the experience pool, in which the old tuple gives space to the newest tuple (if the pool is full).

Considering that the number of satellite users in a beam spot could be very large, the size and computation efficiency of *Q* values of (25) for all possible actions are large and low. In this context, deep neural networks parameterized by θ' and θ, called target DQN and training DQN, respectively, are used in the neural network training phase to estimate the *Q*-value by function approximations. As shown in Figure 2, the target of the

DQN is to estimate the maximum Q-value for the next state, i.e., $\max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}; \theta'\right)$. The training DQN network is deployed to make an action decision and estimate the Q-value for the current state, whose loss function can be written as

$$L(\theta) = E\left[\left(R_\pi(t) + \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1}; \theta'\right) - Q(s_t, a_t; \theta)\right)^2\right]. \tag{27}$$

Using stochastic gradient descent to minimize the function in (27), the correct weights of θ can be learned by the training DQN. The weights θ' are frozen for several steps and then updated by setting $\theta' = \theta$ for the goal of stabilizing the training. The specific steps for the training DQN to select one from many users to be User t is given in Algorithm 1.

Algorithm 1: DQN Algorithm-based NOMA User Pairing in Satellite Networks.

Initialization: $\bar{\alpha}$, $\hat{\gamma}$, pool capacity D, θ, θ', state s_0, l=k=1, parameter C.
while $k \leq K$ **do**
    for $l \leq T$ **do**
        Observe state s_l;
        Choose action a_l under policy π;
        Perform action a_l, observe reward $R_\pi(l)$, and next state s_{l+1};
        Store tuple $(s_l, a_l, R_\pi(l), s_{l+1})$ in pool;
        if *number of tuples in pool is larger than N_p* **then**
            Sample random mini-batch of tuples $(s_l, a_l, R_\pi(l), s_{l+1})$ from D;
            Update θ by performing stochastic gradient descent on (27);
            Set $\theta' = \theta$ in every C steps.
        end
    end
    k=k+1;
end
Output: In each time slot, a user who can achieve the biggest capacity with the power factor in (11) is selected to be User t.

---

5. Results

In this section, simulation results are provided to characterize the effects of users' specific delay QoS requirements on the power allocation scheme, user selection strategy, and system performance. Without loss of generality, we assumed $T_f B = 1$, the carrier frequency as 4 GHz, and radius $R = 125$ km [5,13]. Moreover, we set the number of users as 150, the fading severities, location information, and delay requirements of these users were randomly generated within [ILS, AS, FHS], [0, 1R], and [0.5, 10], respectively, to show the various channel conditions, locations, and application scenarios of different satellite users. The delay QoS exponent of the delay-sensitive user, i.e., User c, was set as $\theta_c = 9.38$, and the label (ILS/AS) denotes the link-shadowing severity of User t/User c in this paper.

We first conducted numerical simulations to show the impact of shadowing, γ, and d_c on the power allocation coefficient α_c^p, as illustrated in Figure 3. From this figure, we can clearly see that, when User c experiences a lighter shadowing, a higher γ, or a closer location information d_c, a larger α_c^p is needed to ensure that the performance achieved with the NOMA scheme is not less than that achieved with the TDMA scheme, which is consistent with the analytical result given in (11). In the following simulations, α_c^p was set to meet the condition of $C_c^N(\theta_c) = C_c^T(\theta_c)$ without other descriptions. Moreover, it can be observed that the analytical results were all consistent with the Monte Carlo simulations.

Then, simulations were conducted to illustrate the capacity of User t achieved with the NOMA scheme and TDMA scheme versus delay requirement θ_t, shown in Figure 4), from which we can clearly observe that the capacity curves all degrade with increasing θ_t. This is an expected result because a larger θ_t means a smaller tolerated delay outage and a lower supported constant arrival rate. Moreover, we find that the superiority of the

NOMA scheme gradually decreases with increasing delay limitation θ_t, i.e., when $\theta_t \geq 10^{0.4}$, the capacity gap between NOMA and TDMA curves almost disappears. The superiority of the NOMA scheme, for the case $\theta_t < 10^{0.4}$, is significantly upgraded for a larger γ, a lighter fading severity of User t, or a smaller d_t. This is because any of these factors corresponds to a more favourable condition. This phenomenon suggests that, in addition to the shadowing, d_t, and γ, θ_t must be taken into account to form a flexible NOMA user group and ensure the superiority of NOMA-based satellite networks.
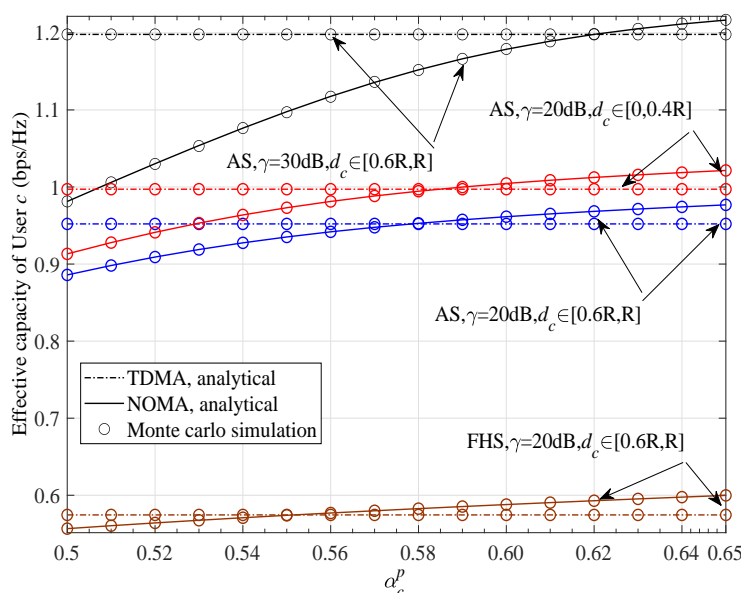
Figure 3. Effective capacity of User c achieved with TDMA and NOMA schemes versus α_c^p under various system parameters.
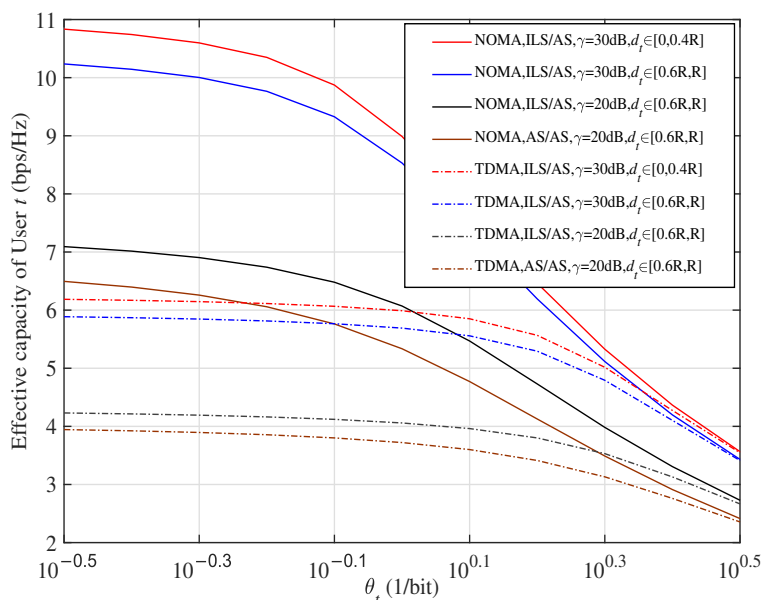

Figure 4. Effective capacity of User t for two access schemes versus θ_t with various γ, d_t, and fading severities, when $d_c \in [0.6R, R]$ and $\alpha_t^p = 1 - \alpha_c^p$.

Finally, the DQN algorithm was adopted to select one from many users to be User t and pair them with User c to form a NOMA user group. Specially, since the assumption that $Q_c < Q_t$ must be satisfied, only users with ILS/AS severity were viewed as candidates.

Meanwhile, $\alpha_t^p = 1 - \alpha_c^p$ varied with the location and fading severity of User c as well as the transmission average SNR γ, as shown in Figure 3.

The convergences of the proposed DQN algorithm with different learning rates are shown in Figure 5, from which we find that a smaller value of learning rate leads to a faster convergence, since a smaller learning rate means a lower newly acquired cost is accepted to adjust the evaluated $Q_\pi(s_l, a_l)$. Thus, $\bar{\alpha} = 0.01$ was set in our algorithm. Figure 6 compares the effective capacity of selected user achieved with NOMA and TDMA schemes under the proposed strategy and random selection strategy. It can be seen from Figure 6 that curves with the proposed NOMA scheme are superior to those with the TDMA scheme for all cases, demonstrating the advantages of employing the NOMA scheme in delay QoS-constrained satellite communication networks. Moreover, since the proposed DQN-based user selection scheme can find the optimal action for each state, and, thus, it can provide superior performance as well as a much bigger performance difference between NOMA and TDMA schemes than those achieved with a random selection strategy in each time slot.
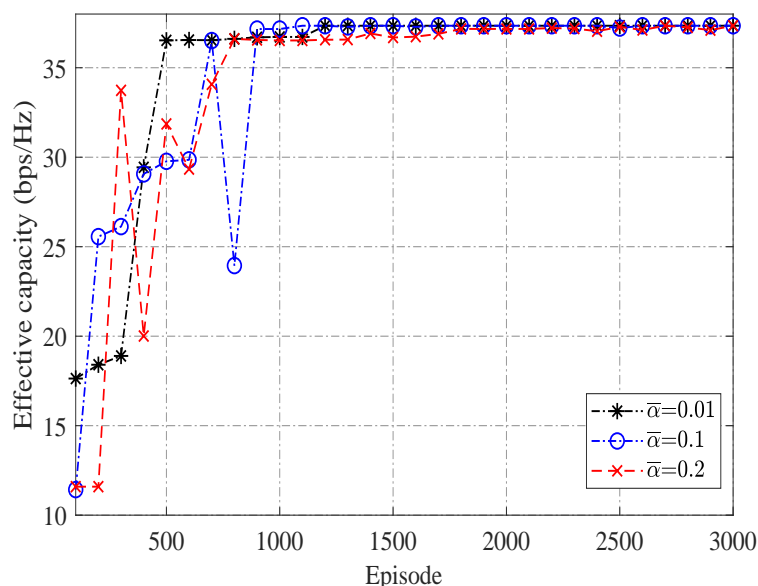

Figure 5. Convergences of the proposed DQN user selection algorithm with different learning rates.
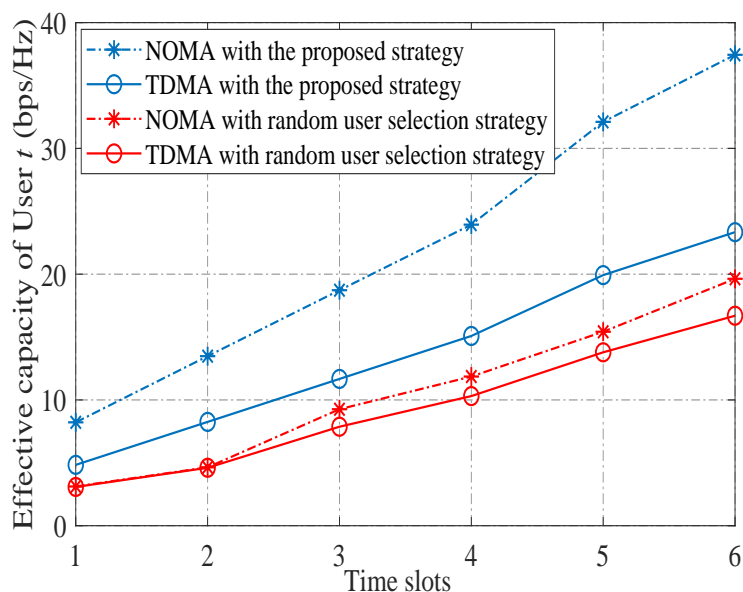

Figure 6. Effective capacity of selected user achieved with two access schemes under the proposed strategy and random selection strategy.

6. Conclusions

In this paper, we have proposed a user pairing scheme in NOMA-based satellite networks with delay QoS constraints. With the objective of maximizing the sum effective capacity without degrading the performance of the delay-sensitive user, the user pairing problem was formulated. In particular, we designed the power allocation strategy to make sure that the performance of the delay-sensitive user achieved with the NOMA scheme was not less than that achieved with an OMA scheme. Based on this, the DRL algorithm was adopted to select a user from many users to pair with the delay-sensitive user and form a NOMA group. Simulation results have been provided to validate those performance analyses, show the effects of key parameters on system performance and the user selection strategy, and demonstrate that the DRL algorithm can significantly improve the system performance by finding the optimal action for each state.

Author Contributions: Conceptualization, Q.Z. and K.A.; methodology, X.Y. and K.A.; validation, X.Y. and K.A.; investigation, Q.Z., H.X., and Y.W.; writing—original draft preparation, X.Y.; writing—review and editing, X.Y., K.A., and Q.Z. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wang, C.-X.; You, X.; Gao, X.; Zhu, X.; Li, Z.; Zhang, C.; Wang, H.; Huang, Y.; Chen, Y.; Haas, H.; et al. On the road to 6G: Visions, requirements, key technologies and testbeds. *IEEE Commun. Surv. Tutor.* 2023, *in press*. [CrossRef]
2. Kodheli, O.; Lagunas, E.; Maturo, N.; Sharma, S.K.; Shankar, B.; Montoya, J.F.M.; Duncan, J.C.M.; Spano, D.; Chatzinotas, S.; Kisseleff, S.; et al. Satellite communications in the new space era: A survey and future challenges. *IEEE Commun. Surv. Tut.* **2021**, *23*, 70–109. [CrossRef]
3. Tang, J.; Bian, D.; Li, G.; Hu, J.; Cheng, J. Resource allocation for LEO beam-hopping satellites in a spectrum sharing scenario. *IEEE Access* **2021**, *9*, 56468–56478. [CrossRef]
4. Guo, K.; Dong, C.; An, K. NOMA-based cognitive satellite terrestrial relay network: Secrecy performance under channel estimation errors and hardware impairments. *IEEE Internet Things J.* **2022**, *9*, 17334–17347. [CrossRef]
5. Yan, X.; Xiao, H.; An, K.; Zheng, G.; Chatzinotas, S. Ergodic capacity of NOMA-based uplink satellite networks with randomly deployed users. *IEEE Syst. J.* **2020**, *14*, 3343–3350. [CrossRef]
6. Jiao, J.; Sun, Y.; Wu, S.; Wang, Y.; Zhang, Q. Network utility maximization resource allocation for NOMA in satellite-based internet of things. *IEEE Internet Things J.* **2020**, *7*, 3230–3242. [CrossRef]
7. Toka, M.; Vaezi, M.; Shin, W. Outage analysis of alamouti-NOMA scheme for hybrid satellite–terrestrial relay networks. *IEEE Internet Things J.* **2023**, *10*, 5293–5303. [CrossRef]
8. Jiao, J.; Hong, H.; Wang, Y.; Wu, S.; Lu, R.; Zhang, Q. Age-optimal downlink NOMA resource allocation for satellite-based IoT network. *IEEE Trans. Veh. Technol.* 2023, *in press*. [CrossRef]
9. Liu, R.; Guo, K.; An, K.; Zhou, F.; Wu, Y.; Huang, Y.; Zheng, G. Resource allocation for NOMA-enabled cognitive satellite-UAV-terrestrial networks with imperfect CSI. *IEEE Trans. Cogn. Commun. Netw.* 2023, *in press*. [CrossRef]
10. Wu, D.; Negi, R. Effective capacity: A wireless link model for support of quality of service. *IEEE Trans. Wireless Commun.* **2003**, *2*, 630–643.
11. Ji, Z.; Cao, S.; Wu, S.; Wang, W. Delay-aware satellite-terrestrial backhauling for heterogeneous small cell networks. *IEEE Access*, **2020**, *8*, 112190–112202. [CrossRef]
12. Ruan, Y.; Li, Y.; Wang, C.-X.; Zhang, R.; Zhang, H. Effective capacity analysis for underlay cognitive satellite-terrestrial networks. In Proceedings of the 2017 IEEE International Conference on Communications, Paris, France, 21–25 May 2017.
13. Ruan, Y.; Li, Y.; Wang, C.-X.; Zhang, R.; Zhang, H. Energy efficient power allocation for delay constrained cognitive satellite terrestrial networks under interference constraints. *IEEE Trans. Wireless Commun.* **2019**, *18*, 4957–4969. [CrossRef]

14. Yan, X.; An, K.; Li, D.; Xi, H.; Wang, Y.; Li, X.; Chen, H. Delay-limited performance analysis of NOMA-enabled satellite internet of things. In Proceedings of the 2021 IEEE/CIC International Conference on Communications in China, Xiamen, China, 28–30 July 2021.

15. Yan, X.; An, K.; Wang, C.-X.; Zhu, W.-P.; Li, Y.; Feng, Z. Genetic algorithm optimized support vector machine in NOMA-based satellite networks with imperfect CSI. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020.

16. Lei, L.; Lagunas, E.; Yuan, Y.; Kibria, M.G.; Chatzinotas, S.; Ottersten, B. Deep learning for beam hopping in multibeam satellite systems. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2020.

17. Lei, L.; Lagunas, E.; Yuan, Y.; Kibria, M.G.; Chatzinotas, S.; Ottersten, B. Beam illumination pattern design in satellite networks: Learning and optimization for efficient beam hopping. *IEEE Access* **2020**, *8*, 136655–136667. [CrossRef]

18. Homssi, B.A.; Chan, C.C.; Wang, K.; Rowe, W.; Allen, B.; Moores, B.; Csurgai-Horváth, L.; Fontxaxn, F.P.; Keepan, S.; Al-Hourani, A. Deep learning forecasting and statistical modeling for Q/V-band LEO satellite channels. *IEEE Trans. Mach. Learn. Commun. Netw.* 2023, *in press*. [CrossRef]

19. Zhao, B.; Dong, X.; Ren, G.; Liu, J. Optimal user pairing and power allocation in 5G satellite random access networks. *IEEE Trans. Wireless Commun.* **2022**, *21*, 4085–4097. [CrossRef]

20. Liu, H.; Wang, Y.; Wang, Y. A successive deep Q-learning based distributed handover scheme for large-scale LEO satellite networks. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring), Helsinki, Finland, 19–22 June 2022.

21. Tubiana, D.A.; Farhat, J.; Brante, G.; Souza, R.D. Q-learning NOMA random access for IoT-satellite terrestrial relay networks. *IEEE Trans. Wireless Lett.* **2022**, *11*, 1619–1623. [CrossRef]

22. Chen, Y.-J.; Chen, W.; Ku, M.-L. Trajectory design and link selection in UAV-assisted hybrid satellite-terrestrial network. *IEEE Wireless Commun. Lett.* **2022**, *26*, 1643–1647. [CrossRef]

23. Zhao, J.; Yu, L.; Cai, K.; Zhu, Y.; Han, Z. RIS-aided ground aerial NOMA communications: A distributionally robust DRL approach. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 1287–1301 [CrossRef]

24. Deng, B.; Jiang, C.; Yao, H.; Guo, S.; Zhao, S. The next generation heterogeneous satellite communication networks: Integration of resource management and deep reinforcement learning. *IEEE Wirel. Commun.* **2020**, *27*, 105–111. [CrossRef]

25. Shao, S.; Hailes, P.; Wang, T.-Y.; Wu, J.-Y.; Maunder, R.G.; Al-Hashimi, B.M.; Hanzo, L. Survey of turbo, LDPC, and polar decoder ASIC implementations. *IEEE Commun. Surveys Tuts.* **2019**, *21*, 2309–2333. [CrossRef]

26. Lin, Z.; Lin, M.; Champagne, B.; Zhu, W.-P.; Al-Dhahir, N. Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks. *IEEE Trans. Commun.* **2021**, *69*, 6345–6360. [CrossRef]

27. Lin, Z.; An, K.; Niu, H.; Hu, Y.; Chatzinotas, S.; Zheng, G.; Wang, J. SLNR-based secure energy efficient beamforming in multibeam satellite systems. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 2085–2088. [CrossRef]

28. An, K.; Chatzinotas, S.; Hu, Y.; Lin, Z.; Niu, H.; Wang, Y.; Zheng, G. Refracting RIS aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 3717–3724.

29. Tegos, S.A.; Diamantoulakis, P.D.; Xia, J.; Fan, L.; Karagiannidis, G.K. Outage performance of uplink NOMA in land mobile satellite communications. *IEEE IEEE Trans. Wireless Lett.* **2020**, *7*, 1710–1714. [CrossRef]

30. Chu, J.; Chen, X.; Zhong, C.; Zhang, Z. Robust design for NOMA-based multibeam LEO satellite internet of things. *IEEE Internet Things J.* **2021**, *8*, 1959–1970. [CrossRef]

31. Abdi, A.; Lau, W.; Alouini, M.-S.; Kaveh, M. A new simple model for land mobile satellite channels: First and second order statistics. *IEEE Trans. Wireless. Commun.* **2003**, *2*, 519–528. [CrossRef]

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

33. Zhong, R.; Liu, Y.; Mu, X.; Chen, Y.; Song, L. AI empowered RIS-assisted NOMA networks: Deep learning or reinforcement learning? *IEEE J. Sel. Areas Commun.* **2021**, *40*, 182–196. [CrossRef]