



# Article Constrained DRL for Energy Efficiency Optimization in RSMA-Based Integrated Satellite Terrestrial Network

Qingmiao Zhang<sup>1</sup>, Lidong Zhu<sup>1,\*</sup>, Yanyan Chen<sup>2</sup> and Shan Jiang<sup>3</sup>

- <sup>1</sup> National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China; qingmiao.zhang@outlook.com
- <sup>2</sup> School of Computer Science and Information Engineering, Xiamen Institute of Technology, Xiamen 361021, China; chenyanyan@xit.edu.cn
- <sup>3</sup> China Mobile (Jiangxi) Communications Group Co., Ltd., Yichun 336000, China; jiangshan@jx.chinamobile.com
- \* Correspondence: zld@uestc.edu.cn

Abstract: To accommodate the requirements of extensive coverage and ubiquitous connectivity in 6G communications, satellite plays a more significant role in it. As users and devices explosively grow, new multiple access technologies are called for. Among the new candidates, rate splitting multiple access (RSMA) shows great potential. Since satellites are power-limited, we investigate the energyefficient resource allocation in the integrated satellite terrestrial network (ISTN)-adopting RSMA scheme in this paper. However, this non-convex problem is challenging to solve using conventional model-based methods. Because this optimization task has a quality of service (QoS) requirement and continuous action/state space, we propose to use constrained soft actor-critic (SAC) to tackle it. This policy-gradient algorithm incorporates the Lagrangian relaxation technique to convert the original constrained problem into a penalized unconstrained one. The reward is maximized while the requirements are satisfied. Moreover, the learning process is time-consuming and unnecessary when little changes in the network. So, an on-off mechanism is introduced to avoid this situation. By calculating the difference between the current state and the last one, the system will decide to learn a new action or take the last one. The simulation results show that the proposed algorithm can outperform other benchmark algorithms in terms of energy efficiency while satisfying the QoS constraint. In addition, the time consumption is lowered because of the on-off design.

**Keywords:** integrated satellite terrestrial network; rate splitting multiple access; energy efficiency; constrained deep reinforcement learning; soft actor-critic

# 1. Introduction

As users and devices explosively grow in the next generation mobile communication systems and Internet of Things (IoT), it is more difficult to provide ubiquitous and reliable massive access with conventional terrestrial networks [1]. Since satellite communication has greatly developed over the years, it is considered a promising solution as a complement for terrestrial communications with its extensive coverage and continuous service [2]. However, it is still not enough to meet the demands with satellites alone; new multiple access (MA) technologies are also called for [3].

In long-term evolution (LTE), orthogonal frequency division multiple access (OFDMA) and singe carrier frequency division multiple access (SC-FDMA) are adopted to boost the capacity in downlink and uplink, respectively. By assigning users orthogonal sub-carriers or discrete Fourier transform (DFT) precoders, multiple access among them becomes possible. However, the orthogonality requirement limits their capacity. The massive machine type communication (mMTC) in 5G and 6G cannot be supported by them. Several non-orthogonal multiple access (NOMA) technologies are proposed to further improve the



Citation: Zhang, Q.; Zhu, L.; Chen, Y.; Jiang, S. Constrained DRL for Energy Efficiency Optimization in RSMA-Based Integrated Satellite Terrestrial Network. *Sensors* **2023**, 23, 7859. https://doi.org/10.3390/ s23187859

Academic Editor: Mario Marchese

Received: 18 August 2023 Revised: 12 September 2023 Accepted: 12 September 2023 Published: 13 September 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). spectral efficiency. Among them, power domain NOMA (PD-NOMA) has been extensively studied over the years. Different transmission powers are allocated based on users' channel conditions. By utilizing the successive interference cancellation (SIC) technique, multiple users could share the same frequency band and time slot, resulting in capacity boost. Nonetheless, it requires user ordering and significant channel discrepancy for better performance. The increase of users and devices also makes it more difficult and complex for user paring, ordering and power allocation, causing performance deterioration. Recently, a novel rate splitting multiple access (RSMA) has drawn a lot attention. It divides user's messages into a public part and a private part, assigning them corresponding precoders. Users also share the same frequency and time resources. But it does not need user paring and ordering or great channel differences to perform well. It has higher capacity compared with OFDMA, SC-FDMA, and PD-NOMA with relatively low receiver complexity [4,5].

The idea of rate splitting was first presented in a single input single output (SISO) system [6]. Over the years, it has been extended for multiple input single output (MISO) and multiple input multiple output (MIMO) systems [7,8]. In the groundbreaking works of rate splitting (RS), the authors thoroughly study this method in terms of sum-rate theoretically; besides the obvious larger achievable rate region, the results prove that RS not only significantly decrease the complexity of precoder design and user scheduling but also exhibits robustness with imperfect channel state information (CSI) compared with conventional MA schemes [7,8]. For the classical weighted sum-rate (WSR) maximization in RSMA, usually this non-convex problem is converted to a convex weighted minimum mean square error (WMMSE) to solve [9]. But recently, Ref. [10] has adopted a new deep learning approach called deep unfolding to simplify and improve the WMMSE algorithm in an integrated satellite terrestrial network (ISTN), while [11] proposes to use the blockcoordinate descent algorithm for an intelligent reflecting surface (IRS) aided system, both of which could achieve a better performance. Another hot topic of RSMA is the maxmin fairness in a multi-group or multi-beam scenario. Reference [12] uses semi-definite relaxation and a convex-concave procedure to attain better fairness. Another reference [13] introduces LogSumExp approximation and a generalized power iteration framework for beamforming. Lately, the application of RSMA for simultaneous wireless information and power transfer (SWIPT) has gained more attention due to exponential energy consumption growth. In a multigroup multicast system, the authors of [14] uses the successive convex approximation (SCA) algorithm to minimize the transmission power.

Since the satellite is powered by solar panels, its power consumption cannot be neglected. Even though we could adopt RSMA to increase the capacity of ISTN, it is also crucial to complete the task with the lowest power possible. As for the energy efficiency optimization for RSMA, some of the literature proposes SCA-based algorithms for this problem [15,16]. However, these approaches are either too complex or only approximate to optimal solutions. Deep reinforcement learning, on the other hand, is a model-free method that can find the optimal solution with no need of prior knowledge. Additionally, the ability to learn is considered a fundamental requirement in the next-generation communication systems. Naturally, DRL is a hopeful candidate solution.

Using DRL for energy efficiency optimization in communications and other systems has been researched over the years. These studies fall into two categories: one with "discrete DRL" and one with "continuous DRL". The former could only deal with discrete action/state space problems, e.g., deep Q-network (DQN) and state action reward state action (SARSA). However, most of the time, energy efficiency optimization is a continuous problem. Therefore, to apply "discrete DRL" to it, the action/state space should be discretized first. For example, Ref. [17] proposes to combine centralized DQN with multiagent DQN in a 5G cognitive heterogeneous network, Ref. [18] utilizes a SARSA-based algorithm for ultra-dense networks, a multi-armed bandit RL is used for the energy-efficient trajectory design of unmanned aerial vehicles (UAVs). The second kind of RL could tackle continuous action/state space tasks, like deep deterministic policy gradient (DDPG) and proximal policy optimization (PPO). So there is no discretization requirement in these

schemes. Ref. [19] applies DDPG to jointly optimize computing and caching; however, the constraints in it are simplified and not shown in the algorithm design. To allocate resources for 5G radio access network (RAN), asynchronous advantage actor-critic (A3C) is exploited. To satisfy the constraints, Ref. [20] uses many modules, which increases the complexity.

Besides DRL, other approaches—model-based or model-free—are also widely used for energy efficiency optimization in various systems. In [21], the authors adopt supervised machine learning to forecast traffic load and evaluate energy efficiency in Beyond 5G (B5G) networks. In wireless sensor networks (WSNs), Ref. [22] proposes a reliable clustering and routing algorithm considering different factors to save energy. The authors of [23] designed a cluster-based hierarchical routing protocol forming minimum spanning trees among communicating nodes, reducing energy wastage and prolonging the network lifetime in WSNs.

Energy efficiency optimization problems in communication networks are always limited by some conditions. Therefore, DRL, which could meet the constraints and tackle continuous tasks, is needed. Fortunately, constrained DRL (CDRL) has been proposed for this kind of problems. Generally, there are three main classes of CDRL:

- 1. Primal-dual approach incorporates Lagrangian relaxation to transform the original constrained problem to an unconstrained dual problem [24,25];
- 2. Policy optimization approach directly uses a more trackable function to replace original objective or constraint [26,27];
- 3. Penalty function approach designs an unconstrained problem by adding penalty in the objective [28].

CDRL has been applied for network slicing recently. In [29], interior-point policy optimization (IPO) is adopted while [30,31] utilize the soft actor-critic (SAC) method. Another reference [32] gives an example of optimal power flow using constrained DDPG.

Nonetheless, there is no prior research on the energy efficiency optimization of RSMA adopting CDRL. So we propose to use SAC with a Lagrangian relaxation technique to improve the energy efficiency with the QoS constraint in ISTN. The contributions of our work are summarized as follows:

- We formulate the constrained Markov decision process (CMDP) framework of energy efficiency optimization, presenting a dual unconstrained problem adopting Lagrangian relaxation;
- We propose a constrained SAC algorithm whose objective function includes QoS constraint and entropy as a penalty;
- An on-off mechanism is introduced to avoid unwanted learning process when little changes in ISTN.

#### 2. System Model

#### 2.1. RSMA Architecture

To better understand how RSMA works, we use a *k*-user MIMO system to demonstrate the process, as shown in Figure 1. First, all users' messages are divided into two parts: public and private parts, which would be  $\{M_1^{\text{pub}}, M_2^{\text{pub}}, \dots, M_k^{\text{pub}}\}$  and  $\{M_1^{\text{pri}}, M_2^{\text{pri}}, \dots, M_k^{\text{pri}}\}$ in the figure. Then, all the common messages are combined together to obtain one new message  $M^{\text{pub}}$  whereas *k* private messages remain untouched. By encoding, we can obtain k + 1 streams  $\{s_{\text{pub}}, s_1, \dots, s_k\}$ . Stream  $s_{\text{pub}}$  will be encoded again by a public precoder  $w_{\text{pub}}$  shared among all users. As for the private streams, each of them will be encoded by its own precoder  $\{w_1, w_2, \dots, w_k\}$  accordingly. Finally, by superimposition, we will get the transmitted signal  $w_{\text{pub}}s_{\text{pub}} + w_1s_1 + \dots + w_ks_k$ .

At the receiver end, since the public precoder  $w_{pub}$  is known to all users, they can directly decode the signal to restore their public messages. Performing one-time successive interference cancellation (SIC) and decoding, one can restore its own private message by discarding all others' private streams. After combination, the original message is obtained.

According to reference [9], RSMA is the bridge that connects spacial division multiple access (SDMA) and PD-NOMA. In SDMA, users are assigned with different beamforming precoders. Multi-user interference is discarded as noise. The performance of SDMA heavily depends on the orthogonality of users' channels and the number of antenna.



Figure 1. RSMA architecture.

In PD-NOMA, users are distinguished by different power. And in a multi-user system, the *k*-th user needs to perform (k - 1) times SIC to restore its own message. That means multi-user interference needs to be fully decoded. PD-NOMA's performance highly relies on the discrepancy of users' channel gains and the perfect SIC operation.

Unlike these two extreme approaches, RSMA partially decodes the public stream and partially discards the private streams interference. This design allows it to achieve larger capacity. Moreover, the performance of RSMA depends on the precoder design and does not need channels to be orthogonal or significantly different. Therefore, RSMA is less limited than the other two. In addition, the one-time SIC in RSMA greatly reduces the complexity at receiver end. The advantages of RSMA could be summarized as follows.

- It has a larger achievable rate region, which means it allows more connections;
- It has no requirements for users' channel condition, so it could be applied to more scenarios;
- It does not need user pairing or ordering for SIC, which reduces the complexity and delay;
- It only needs one-time SIC operation. Complexity is reduced again and the error accumulation is avoided.

#### 2.2. Network Model

The integrated satellite terrestrial network we investigate is depicted in Figure 2. In the downlink scenario, a satellite serves multiple users within its coverage. It could be a low earth orbit (LEO) satellite, a medium earth orbit (MEO) satellite or a geosynchronous (GEO) satellite. It is equipped with *L* antennas to provide extensive coverage. Each user is equipped with a single antenna. In this network, the RSMA scheme is adopted for multi-user transmission. Using different precoders, it can simultaneously serve multiple users and reuse the entire spectrum effectively.



Figure 2. Network model.

To better describe the link between satellite and terrestrial users in real life, we employ the shadowed-Rician fading as our channel model [33]. Assume  $h_k$  is the channel fading coefficient between satellite and *k*-th user. The probability density function (PDF) is listed below:

$$f_{|h_k|^2}(x) = \alpha_k e^{-\beta_k x} {}_1 F_1(m_k; 1; \delta_k x), \tag{1}$$

where  $\alpha_k = (2b_k m_k / (2b_k m_k + \Omega_k))^{m_k} / 2b_k$ ,  $\beta_k = 1/2b_k$ , and  $\delta_k = \Omega_k / 2b_k (2b_k m_k + \Omega_k)$ .  $\Omega_k$  and  $2b_k$  indicates the average power of line-of-sight (LoS) and multipath components, respectively.  $m_k$  is the Nakagami-*m* parameter with range  $[0, \infty)$ .  ${}_1F_1(a; b; x)$  is the confluent hypergeometric function given by

$${}_{1}F_{1}(a;b;x) = \sum_{i=0}^{\infty} \frac{(a)_{i}}{(b)_{i}} \frac{x^{i}}{i!},$$
(2)

where  $(a)_i = \Gamma(a+i)/\Gamma(a)$  is the Pochhammer symbol and  $\Gamma(a)$  is the Gamma function [34].

According to [8], RSMA's performance relies on the quality of CSI. In ISTN, the CSIs of satellites are continuously collected to monitor network condition and adjust precoders. However, perfect CSIs are unlikely to be obtained due to various reasons. A sensor could be introduced to assist CSI collection, but its placement will affect the results. Several papers [35–37] have studied this problem and proposed various algorithms. In this paper, for simplicity, we assume CSIs are acquired by terrestrial terminals.

The errors in CSI have various sources, such as Doppler shift [38] and satellite attitude dynamics [39,40]. Taking all these factors into account will cause rapid expansion of state space in DRL, which will lead to unstable learning and incapability to converge. Fortunately, the influence of these factors can be reflected in CSI. To achieve a tradeoff between complexity and accuracy, we add noise sampled from complex Gaussian distribution into the channel to simulate errors.

The satellite in this network divides the messages for all users and encodes them with RSMA precoders. After broadcasting, terrestrial users use a shared public precoder and their own private ones to restore the messages for them. To maximize the capacity with minimum power consumption, we will investigate the energy efficiency optimization in the next section.

#### 3. Problem Formulation

As mentioned before, we define the vector of user streams as  $s \triangleq [s_{\text{pub}}, s_1, \cdots, s_K] \in \mathbb{C}^{K+1}$  and the precoder matrix is  $\mathbf{w} \triangleq [w_{\text{pub}}, w_1, \cdots, w_K] \in \mathbb{C}^{L \times (K+1)}$ . The final linearly precoded signal that satellite broadcasts to K users is

$$\mathbf{x} = \mathbf{w}_{\mathrm{pub}}s_{\mathrm{pub}} + \sum_{k=1}^{K} w_k s_k, \tag{3}$$

where  $w_{\text{pub}}, w_k \in \mathbb{C}^L$  are used to execute simultaneous downlink transmissions. The received signal has the following form:

$$\mathbf{y} = \mathbf{h}^H \mathbf{x} + \boldsymbol{n},\tag{4}$$

where  $\mathbf{h} = [h_1, h_2, \dots, h_K]$  is the downlink channel matrix,  $\mathbf{n} \sim C\mathcal{N}(0, \mathbf{I}_K)$  is the additive white Gaussian noise (AWGN) vector, which is normalized for brevity. For better understanding, we take user *k* as an example, its received signal will be

$$y_{k} = h_{k}^{H} w_{\text{pub}} s_{\text{pub}} + h_{k}^{H} w_{k} s_{k} + h_{k}^{H} \sum_{i \neq k} w_{i} s_{i} + n_{k},$$
(5)

where item  $h_k^H \sum_{i \neq k} w_i s_i$  is the multi-user interference and  $n_k$  is the noise.

Without loss of generality, we assume the power of users' streams are normalized as well, which is  $\mathbb{E}[\mathbf{ss}^H] = \mathbf{I}$ . This will help us to obtain simpler and clearer mathematical expressions. Therefor, the SINR of the public part of the signal for user *k* is

$$\gamma_{k}^{\text{pub}} = \frac{\left|h_{k}^{H}w_{\text{pub}}\right|^{2}}{\sum_{i=1}^{K}\left|h_{i}^{H}w_{i}\right|^{2} + 1}.$$
(6)

Usually, we can now acquire the rate with Shannon's formula. But in this scenario, we want to guarantee that all *K* users could successfully receive the signal. That means the rate of the public part of the signal should not exceed the user with lowest public part rate, which yields the following equation:

$$R_k^{\text{pub}} = \min_k \left\{ \log_2 \left( 1 + \gamma_k^{\text{pub}} \right) \right\}.$$
(7)

Similarly, the SINR of private part of the signal for user *k* will be

$$\gamma_k^{\text{pri}} = \frac{\left|h_k^H w_{\text{pub}}\right|^2}{\sum_{i \neq k}^K \left|h_i^H w_i\right|^2 + 1}.$$
(8)

For the private part of the signal, user *k* will discard others' private streams. So there is no rate requirement like the public part. The rate of private part of the signal for user *k* is

$$R_k^{\rm pri} = \log_2\left(1 + \gamma_k^{\rm pri}\right). \tag{9}$$

Now we can get the final rate of user *k*, which is

$$R_k = R_k^{\rm pub} + R_k^{\rm pri}.$$
 (10)

The total power consumption of the broadcast signal is  $|w_{pub}s_{pub}|^2 + \sum_k |w_k s_k|^2$ . Since we have normalized the power of user streams, we can obtain the power consumption as tr(**ww**<sup>H</sup>), in which tr(·) is the trace of a matrix. By the definition of energy efficiency, the problem we aim to solve is given as

$$\max_{\mathbf{w}} EE = \frac{\sum_{k=1}^{K} R_k}{\operatorname{tr}(\mathbf{w}\mathbf{w}^H) + P_c}.$$
(11)

s.t. 
$$\operatorname{tr}\left(\mathbf{w}\mathbf{w}^{H}\right) \leq P_{t},$$
 (12)

$$R_k \ge R^{\text{th}},$$
 (13)

where  $P_c$  is the circuit power consumed in the satellite,  $P_t$  is the maximum transmission power of the satellite and  $R^{\text{th}}$  is the required minimum rate.

Equation (11) is the sum energy efficiency in the network. Constraint (12) is the power limit, indicating that the transmit power cannot exceed the maximum power. Constraint (13) is the QoS constraint to guarantee successful transmission. Since this problem is non-convex and NP-hard, conventional approaches use SCA-based methods to approximate the optimal solution. However, the complexities of these approaches are very high, and the optimal solution sometimes cannot be obtained. So we consider adopting a model-free DRL method for this problem.

To satisfy (12), we can simply use the projection operation to limit the precoders, which can be expressed as:

$$Proj(\mathbf{w}) = \begin{cases} \mathbf{w} & \text{if } \operatorname{tr}(\mathbf{w}\mathbf{w}^{H}) \leq P_{t}, \\ \mathbf{w}/\|\mathbf{w}\| & \text{otherwise.} \end{cases}$$
(14)

Nonetheless, Constraint (13) is difficult to incorporate in a conventional DRL algorithm. Naturally, we want to adopt constrained DRL to tackle this problem, which utilizes different methods to satisfy the constraint. As the soft actor-critic algorithm has shown excellent performance and has been able to deal with continuous action/state space tasks, we combine it with Lagrangian relaxation technique to optimize the energy efficiency.

#### 4. Constrained SAC Algorithm

4.1. Constrained Markov Decision Process Formulation

In a normal DRL framework, the problem should be first defined by a turple (S, A, P, r) to solve, which consists of state space S, action space A, transition probability P and reward r. Similarly, a constrained DRL also needs this turple but with an extra cost function C. So, we give the following definitions.

State space is a set of representations of the environment observed by the agent. These observations have all the relevant information the agent needs to make a decision. In the ISTN, the SINRs, rates and energy efficiency can be calculated according to the channel fading coefficients, thus, we select them as the state. The current state space is defined as:

$$s^{(t)} = \left\{ h_1^{(t)}, h_2^{(t)}, \cdots, h_K^{(t)} \right\}.$$
(15)

Action space is a set of all valid actions of choices available to the agent. In this scenario, precoders are allocated to users to calculate energy efficiency with channel fading coefficients. So the current action space is given as

$$a^{(t)} = \{\mathbf{w}\},\tag{16}$$

where **w** is defined in Section 3.

Reward is a feedback from the environment based on the last action that agent took. It utilizes this feedback to improve its action policy. Usually, we can directly use energy efficiency as a reward; however, the QoS constraint cannot be incorporated in this case. Therefore, we introduce the Lagrangian relaxation method to convert the original problem into an unconstrained one.

According to references [41,42], the following constrained optimization

$$\max_{x} f(x). \tag{17}$$

$$\text{s.t. } g_i(x) \ge 0, \tag{18}$$

is equivalent to the problem

$$\min_{\lambda>0} \max_{x} \hat{f}(x, \lambda, \rho) = f(x) + \sum_{i} \lambda_{i} g_{i}(x),$$
(19)

where  $\hat{f}(x, \lambda)$  is the Lagrangian relaxed form of original f(x). The Lagrangian multiplier  $\lambda_i$  adjusts the relative importance of constraint  $g_i(x)$ , which is called a cost function as well, against the objective function. Then, with the following equations at the *t*-th iteration, it can be gradually updated.

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \mu_\lambda g_i\left(x^{(t)}\right),\tag{20}$$

where  $\mu_{\lambda}$  is the learning rate of the Lagrangian multiplier.

With the above information, the current instantaneous reward in Lagrangian relaxed form is defined as:

$$r(s^{(t)}, a^{(t)}) = EE + \sum_{k=1}^{K} \lambda_k^{(t)} \left( R_k^{(t)} - R^{\text{th}} \right),$$
(21)

in which the cost function is denoted as  $c_k^{(t)} = R_k^{(t)} - R^{\text{th}}$ . The reward will decline when the QoS constraint is not satisfied.

### 4.2. The Proposed Constrained SAC Algorithm

Standard SAC is a policy gradient approach which uses the actor network to evaluate the policy and critic network to improve the policy. Additionally, the entropy of the policy is used as a penalty to improve its exploration, which yields fast convergence speed [43]. The maximum objective with entropy in it could be achieved when the optimum policy is found, which can be expressed as:

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{T} \mathbb{E}_{s^{(t)}, a^{(t)}} \left[ r\left(s^{(t)}, a^{(t)}\right) + \mu_{\pi} \mathcal{H}\left(\pi\left(a^{(t)} \middle| s^{(t)}\right)\right) \right],$$
(22)

where  $\mu_{\pi}$  is the temperature parameter that regulates the relative importance of the entropy, which could increase the variance of the policy distribution so that more actions are available to be chosen. The Shannon entropy of the policy is defined as:

$$\mathcal{H}\left(\pi\left(a^{(t)}\middle|s^{(t)}\right)\right) = -\sum_{a^{(t)}}\pi\left(a^{(t)}\middle|s^{(t)}\right)\log\pi\left(a^{(t)}\middle|s^{(t)}\right).$$
(23)

The architecture of the proposed constrained SAC is illustrated in Figure 3. As shown in the picture, mainly three different components with five neural networks are used in this scheme, which are elaborated below.



Figure 3. Architecture of constrained SAC.

To evaluate the soft state value, a deep neural network (DNN) with a target network is adopted, which are represented by parameters  $\psi$  and  $\bar{\psi}$ , respectively. The soft state value is used for soft Q-value approximation, both of which are utilized to evaluate received return in the future and are defined as:

$$V(s^{(t)}) = \mathbb{E}_{a^{(t)} \sim \pi} \Big[ Q(s^{(t)}, a^{(t)}) - \log \pi \big( a^{(t)} \big| s^{(t)} \big) \Big],$$
(24)

$$Q\left(s^{(t)}, a_t\right) = r\left(s^{(t)}, a_t\right) + \rho \mathbb{E}_{s^{(t+1)} \sim P}\left[V\left(s^{(t+1)}\right)\right],\tag{25}$$

where  $\rho$  is the discount factor. With the objective and its gradient below, we can update the soft value network using the gradient descent method. This network is trained to minimize the square residual error with a target network to stabilize the training.

$$\mathcal{J}_{V}(\psi) = \mathbb{E}_{s^{(t)}} \left[ \frac{1}{2} \left( V_{\psi}\left(s^{(t)}\right) - \mathbb{E}_{a_{t}} \left[ Q_{\theta}\left(s^{(t)}, a^{(t)}\right) - \log \pi_{\phi}\left(a^{(t)} \left| s^{(t)}\right) \right] \right)^{2} \right],$$
(26)

$$\nabla \mathcal{J}_{V}(\psi) = \nabla_{\psi} V_{\psi}\left(s^{(t)}\right) \left[ V_{\psi}\left(s^{(t)}\right) - Q_{\theta}\left(s^{(t)}, a^{(t)}\right) + \log \pi_{\phi}\left(a^{(t)} \middle| s^{(t)}\right) \right].$$
(27)

The soft Q-value is estimated by double soft Q-value network. The two-Q-network design is meant to avoid overestimation. Parameters in this network could be updated by minimizing the soft Bellman residual. The objective and its gradient are expressed as:

$$\mathcal{J}_{Q}(\theta) = \mathbb{E}_{s^{(t)}, a^{(t)}} \left[ \frac{1}{2} \left( Q_{\theta} \left( s^{(t)}, a^{(t)} \right) - \left( r \left( s^{(t)}, a^{(t)} \right) + \rho \mathbb{E}_{s^{(t)} \sim P} \left[ V_{\bar{\psi}} \left( s^{(t+1)} \right) \right] \right) \right)^{2} \right], \quad (28)$$

$$\nabla \mathcal{J}_{Q}(\theta) = \nabla_{\theta} Q_{\theta} \left( s^{(t)}, a^{(t)} \right) \left[ Q_{\theta} \left( s^{(t)}, a^{(t)} \right) - r \left( s^{(t)}, a^{(t)} \right) - \rho V_{\bar{\psi}} \left( s^{(t+1)} \right) \right].$$
(29)

Finally, the policy network provides the action for agent based on the optimal policy it finds. By minimizing the Kullback–Leibler (KL) divergence between the policy and soft Q-value, the parameters could be learned. The objective and its gradient are listed below.

$$\mathcal{J}_{\pi}(\phi) = \mathbb{E}_{s^{(t)}, a^{(t)}} \Big[ \log \pi_{\phi} \Big( a^{(t)} \Big| s^{(t)} \Big) - Q_{\theta} \Big( s^{(t)}, a^{(t)} \Big) \Big], \tag{30}$$

$$\nabla \mathcal{J}_{\pi}(\phi) = \nabla_{\phi} \pi_{\phi} \left( a^{(t)} \left| s^{(t)} \right) + \left[ \nabla_{a^{(t)}} \log \pi_{\phi} \left( a^{(t)} \left| s^{(t)} \right) - \nabla_{a^{(t)}} Q_{\theta} \left( s^{(t)}, a^{(t)} \right) \right] \nabla_{\phi} f_{\phi} \left( \epsilon_{t}; s^{(t)} \right), \tag{31}$$

where  $a^{(t)} = f_{\phi}(\epsilon_t; s^{(t)})$  is the reparameterization, which could lower the variance estimation.  $\epsilon_t$  is an input noise sampled from Gaussian distribution. The update process of  $\lambda$  could also be realized by neural networks; however, this will slow the learning process and destablize the network. So, we use the gradient ascend method for the update.

Moreover, since the learning process is time-consuming, if CSI changes little in the network, it is wasteful to learn a new action that almost equals to the last one. So we introduce an on–off mechanism to avoid unnecessary learning. When the difference between the current state and the last state is small enough, we will adopt the last action instead of learning a new one. With all the information above, we present our constrained SAC in Algorithm 1.

## Algorithm 1: Constrained SAC algorithm.

- 1: initialize network parameters:  $\psi$ ,  $\theta$ ,  $\phi$
- initialize Lagrangian multiplier and its learning rate, temperature parameter and discount factor: λ, μ<sub>λ</sub>, μ<sub>π</sub>, ρ
- 3: empty replay memory  $\mathcal{D}$ , initialize last state  $\bar{s} = 0$  and initialize last action  $\bar{a} = 0$
- 4: copy parameter to target network:  $\bar{\psi} \leftarrow \psi$
- 5: for each episode  $(t = 1, 2, \dots, T)$  do
- 6: observe environment  $s^{(t)}$
- 7: compute the difference between  $s^{(t)}$  and  $\bar{s}$ :  $\Delta = |s^{(t)} \bar{s}|$ .
- 8: **if**  $\Delta$  is greater than the threshold **then**
- 9: sample action  $a^{(t)}$  from policy network:  $a^{(t)} \sim \pi_{\phi}(\cdot|s^{(t)})$
- 10: set  $\bar{s} = s^{(t)}$  and  $\bar{a} = a^{(t)}$
- 11: **else**
- 12: take action  $a^{(t)} = \bar{a}$
- 13: end if
- 14: calculate reward according to Equation (21)
- 15: update replay memory:  $\mathcal{D} \leftarrow \mathcal{D} \cup \left(s^{(t)}, a^{(t)}, r, s^{(t+1)}, c_k^{(t)}\right)$
- 16: **if** it is time to update **then**
- 17: sample a mini batch from  $\mathcal{D}$ .
- 18: update  $\psi$  via stochastic gradient ascent on (27)
- 19: update  $\theta$  via stochastic gradient ascent on (29)
- 20: update  $\phi$  via stochastic gradient ascent on (31)
- 21: update  $\lambda$  according to (20)
- 22: update  $\bar{\psi}$  by  $\bar{\psi} \leftarrow \tau \psi + (1 \tau) \bar{\psi}$

```
23: end if
```

24: end for

### 5. Simulation Results

In order to verify the performance of the proposed constrained SAC algorithm, we test it in various scenarios with different algorithms, including a DDPG approach incorporating penalty as constraint in its reward function, unconstrained SAC, and conventional SCAbased algorithm.

#### 5.1. Simulation Parameters

In ISTN, multiple users are served by an LEO satellite within its coverage. The free space loss coefficient is given as  $\eta = (1/4f\pi d)$ , in which *f* and *d* are the frequency and distance between satellite and user.

The neural networks (NNs) in the proposed algorithm all consist of three fully connected hidden layers with ReLU as the activation function and Adam optimizer. The first two hidden layers of NN in the soft value network has 64 neurons while the last layer has 1 neuron. The hidden layers of NNs in the policy network and the soft Q network all have 64 neurons. The learning rate of the Adam optimizer is  $10^{-3}$  while the update rate of target networks is 0.005. Their loss functions are given in Equations (21), (22) and (27), respectively. Replay memory size and batch size are 200 and 32, respectively, and the episode is 5000 with 200 timestamps. Since all users are equally "important" in the ISTN, their QoS constraints contribute to the reward equally as well. So, we initialize the Lagrangian multipliers to 1. The initial value of the learning rate of the Lagrangian multiplier is usually small. So we set it to be  $10^{-3}$ . The tuning of them could be performed by introducing additional NNs with different optimizers; however, we fix their values to avoid extra complexity. The parameters are listed in Table 1.

Parameter	Value	Note
b	0.158	Average power of multi-path component
Ω	1.29	Average power of LoS component
т	19.4	Nakagami- <i>m</i> parameter
f	12 GHz	Satellite frequency
d	1000 km	Distance between satellite and users
K	2, 3, 4, 5	Number of users
$P_t$	40 W	Transmission power of satellite
Pc	5 W	Circuit power
Δ	0.1	Difference between last state and current state
R <sup>th</sup>	4 bps/Hz	Rate constraint
ρ	0.98	Discount factor
λ	1	Lagrangian multiplier
$\mu_{\lambda}$	$10^{-3}$	Learning rate of $\lambda$
$\mu_{\pi}$	0.2	Temperature parameter for entropy

Table 1. Simulation parameters.

#### 5.2. Results and Discussion

We evaluate the average reward of constrained SAC with different users, which is depicted in Figure 4. Clearly, it converges in all four scenarios around 200 to 300 episodes, and its average reward increases as users grow in the ISTN. Since we initialize the Lagrangian multiplier as 1 and the rate constraint as 4. The start points of four lines are -8, -12, -16 and -20, respectively. Two things need to be noted. The first is that the fluctuation becomes greater when users grow. The reason for this is that when more users served in the network, the action space becomes larger, making it more difficult to design the precoders

for each user. Another thing is that the gap between two lines shrinks as users increase. Given the transmission power of the satellite, the achievable rate of each user will slightly decrease when more users join the network, resulting in slowing the growth speed of the average reward.



Figure 4. Average reward of constrained SAC with different users.

The energy efficiency performance of constrained SAC with different users is shown in Figure 5. We can see that it shares some similarities with Figure 4, such as the greater fluctuation and smaller gap with more users. The reasons are the same—more users enlarge the action space. Since energy efficiency is the reward without constraint, the start points of these four lines are all 0.



Figure 5. Energy efficiency of constrained SAC with different users.

Next, we verify whether the rate constraint is satisfied with the proposed algorithm. In Figure 6, we can see that with different users, constrained SAC meets all the demands. For clarity, we add four straight lines to indicate the rate constraints in four scenarios. The capacity generated by the proposed algorithm all exceed the constraints. With these three



results, the constrained SAC can maximize the energy efficiency under the rate constraint. It provides an alternative solution for constrained optimization problem.

Figure 6. Capacity of constrained SAC with different users.

To evaluate the performance of constrained SAC (denoted as CSAC in the figures) with other benchmark algorithms, we choose the original SAC with no constraint and DDPG incorporating constraint into its reward function (denoted as DDPG in the figures) [44]. The existing model-based SCA optimization (denoted as SCA in the figures) is compared with these model-free DRL methods as well [15,16].

$$r_{\rm DDPG}^{(t)} = EE\left(1 - \eta^{(t)}\right),\tag{32}$$

where  $\eta(t)$  is the penalty defined as

$$\eta^{(t)} = \frac{1}{K} \sum_{i=1}^{K} \mathbb{I} \Big( R_k^{(t)} - R^{\text{th}} \Big),$$
(33)

in which  $\mathbb{I}(\cdot)$  is a indicator function given by

$$\mathbb{I}(x) = \begin{cases} 1 & x < 0\\ 0 & x \ge 0 \end{cases}$$
(34)

Whenever the rate is lower than the threshold, the reward of DDPG will be penalized. Therefore, to achieve the maximum reward, the rate will meet the requirement.

In the three users scenario, since the original SAC is unconstrained, this means its reward is the energy efficiency. Therefore, it looks like a straight line compared with the other two algorithms that have a larger reward in Figure 7. Constrained DDPG, however, uses the constraint as penalty in its reward function, showing a similar look to constrained SAC. But it is obvious that its convergence speed is slower. The reason for this is that constrained SAC has entropy in the reward function and noise in its action for more exploration, which will eventually improve the convergence speed. Therefore, CSAC has a better exploration performance than DDPG, which only adds noise in the action. And the reward DDPG achieves is a bit lower than that of our algorithm.



Figure 7. Average reward with different algorithms.

The energy efficiency performance comparison is depicted in Figure 8. Original SAC has the lowest performance because its rate requirement cannot be satisfied. DDPG once again presents relatively slow convergence speed. Its energy efficiency is higher than original SAC but lower than constrained SAC. SCA converges after five iterations. Its energy efficiency is even lower than DDPG, demonstrating that DRL could achieve better performance than the model-based optimization approach without any prior knowledge. However, unlike the "black-box" characteristic of DRL, SCA has more explicit expression. CSAC outperforms the other three with the highest performance, which validates its effectiveness.



Figure 8. Energy efficiency with different algorithms.

We also compare the capacities of four algorithms to check if the constraint satisfied. In Figure 9, the straight line represents the capacity requirement in the three users scenario, which is 12 bps/Hz. CSAC, DDPG, and SCA meet the capacity requirement with their different approaches. Original SAC, on the other hand, fails to satisfy the constraint, even

Iteration Capacity (bps/Hz) CSAC DDPG - SCA SAC Episodes

though its fluctuation goes beyond the straight line in rare cases. CSAC achieves the highest capacity among them, which results in the highest energy efficiency in Figure 8.

Figure 9. Capacity with different algorithms.

We investigate the effect of the on-off mechanism. In Figure 10, we vary thresholds to observe the energy efficiency change of the three algorithms. As the threshold increases, it is more likely that the last action will be taken instead of finding a new one. Consequently, the energy efficiency will decrease as the action does not perfectly apply to the new state. According to [9], when the strength difference of two channels is 0.3, additional 5 dB loss will be introduced. Although less learning process with a bigger threshold will cost less time, the tradeoff between time cost and performance should be considered carefully. Normally, the threshold needs to be less than 0.3, or it will deteriorate the capacity and energy efficiency.



Figure 10. Energy efficiency with different on-off mechanism thresholds.

Figure 11 shows the time cost performance of three algorithms. All of their time consumption is lowered when adopting an on–off mechanism because some of the learning process is substituted by the actions from the last states. Among them, the original SAC has the lowest time consumption because it does not need to compute the Lagrangian multiplier or penalty. DDPG has a simpler architecture and constraint in its reward function than CSAC, thus the lower time cost.



Figure 11. Average time cost per execution with different on-off mechanism thresholds.

Finally, the energy efficiency performance of CSAC and DDPG with different constraint requirements is illustrated in Figure 12. As the rate constraint grows, it is more difficult for CSAC and DDPG to find the suitable precoders, so their energy efficiency declines. However, the penalty design in DDPG is more sensitive to requirement change, and it dropps faster than CSAC.



Figure 12. Energy efficiency with different constraint requirements.

# 6. Conclusions

We investigate the energy efficiency optimization of rate splitting multiple access in an integrated satellite terrestrial network, where multiple users are served by a satellite with a QoS constraint. This problem is non-convex and NP-hard; therefore, we adopt a soft actor-critic deep reinforcement learning approach as it can deal with continuous state/action space tasks. To satisfy the QoS constraint, the Lagrangian relaxation technique is incorporated to convert the original problem into a constrained one. Also, an on-off mechanism is introduced to avoid frequent unnecessary learning processes if the state changes little. The simulation results validate that the proposed constrained SAC algorithm achieves a better performance while meeting the QoS requirement comparing with other algorithms in terms of convergence speed and energy efficiency. Time cost is reduced as well when the on-off mechanism is adopted. Future work will focus on CSAC improvement, such as exploration and sample efficiency, and on RSMA in a more complicated system like integrated space air terrestrial networks and satellite-based sensor networks.

Author Contributions: Conceptualization, Q.Z.; methodology, Q.Z. and L.Z.; software, Q.Z., Y.C. and S.J.; validation, Q.Z., Y.C. and S.J.; formal analysis, Q.Z.; investigation, L.Z.; writing—original draft preparation, Q.Z.; writing—review and editing, Q.Z. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data sharing not applicable.

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

#### References

- Wang, C.X.; You, X. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surv. Tutor.* 2023, 25, 905–974. [CrossRef]
- Xiao, Z.; Yang, J. LEO Satellite Access Network (LEO-SAN) towards 6G: Challenges and Approaches. *IEEE Wirel. Commun. Lett.* 2022, 1–8. [CrossRef]
- Liu, Y.; Zhang, S. Evolution of NOMA Toward Next Generation Multiple Access (NGMA) for 6G. *IEEE J. Sel. Areas Commun.* 2022, 40, 1037–1071. [CrossRef]
- Yang, Z.; Shi, J.; Li, Z.; Chen, M.; Xu, W. Energy Efficient Rate Splitting Multiple Access (RSMA) with Reconfigurable Intelligent Surface. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020; IEEE: Piscataway Township, NJ, USA, 2020; pp. 1–6. [CrossRef]
- Yang, Z.; Chen, M.; Saad, W. Optimization of Rate Allocation and Power Control for Rate Splitting Multiple Access (RSMA). *IEEE Trans. Commun.* 2021, *9*, 5988–6002. [CrossRef]
- 6. Han, T.; Kobayashi, K. A New Achievable Rate Region for the Interference Channel. *IEEE Trans. Inf. Theory* **1981**, 27, 49–60. [CrossRef]
- Joudeh, H.; Clerckx, B. Robust Transmission in Downlink Multiuser MISO Systems: A Rate-Splitting Approach. *IEEE Trans.* Signal Process. 2016, 64, 6227–6242. [CrossRef]
- 8. Dai, M.; Clerckx, B. A Rate Splitting Strategy for Massive MIMO With Imperfect CSIT. *IEEE Trans. Wirel. Commun.* 2016, 15, 4611–4624. [CrossRef]
- Clerckx, B.; Mao, Y. Rate-Splitting Unifying SDMA, OMA, NOMA, and Multicasting in MISO Broadcast Channel: A Simple Two-User Rate Analysis. *IEEE Wirel. Commun. Lett.* 2020, *9*, 349–353. [CrossRef]
- Zhang, Q.; Zhu, L. Deep Unfolding for Cooperative Rate Splitting Multiple Access in Hybrid Satellite Terrestrial Networks. *China Commun.* 2022, 19, 100–109. [CrossRef]
- 11. Katwe, M.; Singh, K. Rate Splitting Multiple Access for Sum-Rate Maximization in IRS Aided Uplink Communications. *IEEE Trans. Wirel. Commun.* 2023, 22, 100–109. [CrossRef]
- 12. Lee, B.; Shin, W. Max-Min Fairness Precoder Design for Rate-Splitting Multiple Access: Impact of Imperfect Channel Knowledge. *IEEE Trans. Veh. Technol.* 2023, 72, 1355–1359. [CrossRef]
- Kim, D.; Choi, J. Max–Min Fairness Beamforming With Rate-Splitting Multiple Access: Optimization Without a Toolbox. *IEEE Wirel. Commun. Lett.* 2023, 12, 232–236. [CrossRef]

- Li, T.; Zhang, H. Full-Duplex Cooperative Rate-Splitting for Multigroup Multicast With SWIPT. *IEEE Trans. Wirel. Commun.* 2022, 21, 4379–4393. [CrossRef]
- 15. Xing, F.; He, S. Energy Efficiency Optimization for Rate-Splitting Multiple Access-Based Indoor Visible Light Communication Networks. *IEEE J. Sel. Areas Commun.* 2022, 40, 1706–1720. [CrossRef]
- Katwe, M.; Singh, K. Rate Splitting Multiple Access for Energy Efficient RIS-aided Multi-user Short-Packet Communications. In Proceedings of the 2022 IEEE Globecom Workshops (GC Wkshps), Janeiro, Brazil, 4–8 December 2022; pp. 644–649. [CrossRef]
- 17. Giannopoulos, A.; Spantideas, S. Deep Reinforcement Learning for Energy-Efficient Multi-Channel Transmissions in 5G Cognitive HetNets: Centralized, Decentralized and Transfer Learning Based Solutions. *IEEE Access* **2021**, *9*, 129358–129374. [CrossRef]
- Malta, S.; Pinto, P. Using Reinforcement Learning to Reduce Energy Consumption of Ultra-Dense Networks With 5G Use Cases Requirements. *IEEE Access* 2023, 11, 5417–5428. [CrossRef]
- 19. Arani, A.H.; Azari, M.M. Reinforcement Learning for Energy-Efficient Trajectory Design of UAVs. *IEEE Internet Things J.* 2022, *9*, 9060–9070. [CrossRef]
- Azimi, Y.; Yousefi, S. Energy-Efficient Deep Reinforcement Learning Assisted Resource Allocation for 5G-RAN Slicing. *IEEE Trans. Veh. Technol.* 2022, 71, 856–871. [CrossRef]
- Thantharate, A.; Tondwalkar, A.V.; Beard, C.; Kwasinski, A. ECO6G: Energy and Cost Analysis for Network Slicing Deployment in Beyond 5G Networks. Sensors 2022, 22, 8614. [CrossRef]
- 22. El-Fouly, F.H.; Khedr, A.Y.; Sharif, M.H.; Alreshidi, E.J.; Yadav, K.; Kusetogullari, H.; Ramadan, R.A. ERCP: Energy-Efficient and Reliable-Aware Clustering Protocol for Wireless Sensor Networks. *Sensors* **2022**, *22*, 8950. [CrossRef]
- 23. Nasirian, S.; Pierleoni, P. Pizzza: A Joint Sector Shape and Minimum Spanning Tree-Based Clustering Scheme for Energy Efficient Routing in Wireless Sensor Networks. *IEEE Access* 2023, *11*, 68200–68215. [CrossRef]
- 24. Tessler, C.; Mankowitz, D.J. Reward Constrained Policy Optimization. arXiv 2018, arXiv:1805.11074.
- Ding, D.; Zhang, K. Natural Policy Gadient Primal-dual Method for Constrained Markov Decision Processes. Adv. Neural Inf. Process. Syst. 2020, 33, 8378–8390.
- Achiam, J.; Held, D. Constrained policy optimization. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, NSW, Australia, 6–11 August 2017; pp. 22–31. [CrossRef]
- 27. Yang, T.Y.; Rosca, J. Projection-based Constrained Policy Optimization. arXiv 2020, arXiv:2010.03152.
- Liu, Y.; Ding, J. IPO: Interior-point Policy Optimization under Constraints. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 Febraury–1 March 2020; Volume 34, pp. 4940–4947. [CrossRef]
- Liu, Y.; Ding, J. CLARA: A Constrained Reinforcement Learning Based Resource Allocation Framework for Network Slicing. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 1427–1437. [CrossRef]
- Zhou, X.; Zhang, X. Constrained Soft Actor-Critic for Energy-Aware Trajectory Design in UAV-Aided IoT Networks. *Electronics* 2022, 11, 1414–1418. [CrossRef]
- Qi, Q.; Lin, W. Augmented Lagrangian-Based Reinforcement Learning for Network Slicing in IIoT. *Electronics* 2022, 11, 3385. [CrossRef]
- 32. Yan, Z.; Xu, Y. Real-Time Optimal Power Flow: A Lagrangian Based Deep Reinforcement Learning Approach. *IEEE Trans. Power Syst.* **2020**, *35*, 3270–3273. [CrossRef]
- Jung, D.H.; Ryu, J.G. Performance Analysis of Satellite Communication System Under the Shadowed-Rician Fading: A Stochastic Geometry Approach. *IEEE Trans. Commun.* 2022, 70, 2707–2721. [CrossRef]
- 34. Gradshteyn, S.; Ryzhik, I.M. Table of Integrals, Series and Products, 7th ed.; Academic Press: New York, NY, USA, 2007.
- Yang, C.; Lu, W.; Liang, K.; Zhang, X.; Geng, X. Sensor placement algorithm for structural health monitoring with redundancy elimination model based on sub-clustering strategy. *Mech. Syst. Signal Process.* 2019, 124, 369–387. [CrossRef]
- Yang, C.; Ouyang, H. A novel load-dependent sensor placement method for model updating based on time-dependent reliability optimization considering multi-source uncertainties. *Mech. Syst. Signal Process.* 2022, 165, 108386. [CrossRef]
- Yang, C.; Xia, Y. A novel two-step strategy of non-probabilistic multi-objective optimization for load-dependent sensor placement with interval uncertainties. *Mech. Syst. Signal Process.* 2022, 176, 109173. [CrossRef]
- Zhang, Y.; Wu, Y.; Liu, A. Deep learning-based channel prediction for LEO satellite massive MIMO communication system. *IEEE Wirel. Commun. Lett.* 2021, 10, 1835–1839. [CrossRef]
- 39. Yang, C.; Lu, W.; Xia, Y. Reliability-constrained optimal attitude-vibration control for rigid-flexible coupling satellite using interval dimension-wise analysis. *Reliab. Eng. Syst. Saf.* **2023**, 237, 0951–8320. [CrossRef]
- 40. Yang, C.; Lu, W.; Xia, Y. Uncertain optimal attitude control for space power satellite based on interval Riccati equation with non-probabilistic time-dependent reliability. *Aerosp. Sci. Technol.* **2023**, *139*, 1270–9638. [CrossRef]
- Li, J.; Fridovich-Keil, D. Augmented Lagrangian Method for Instantaneously Constrained Reinforcement Learning Problems. In Proceedings of the 2021 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, 14–17 December 2021; pp. 2982–2989. [CrossRef]
- 42. Andreani, R.; Birgin, E.G. On Augmented Lagrangian Methods with General Lower-level Constraints. *SIAM J. Optim.* 2008, 18, 1286–1309. [CrossRef]

- 43. Haarnoja, T.; Zhou, A. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1861–1870.
- 44. Hieu, N.Q.; Hoang, D.T. Optimal Power Allocation for Rate Splitting Communications With Deep Reinforcement Learning. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 2820-2823. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.