energies

MDPI

Article

Dynamic User Resource Allocation for Downlink Multicarrier NOMA with an Actor–Critic Method

Xinshui Wang *, Ke Meng, Xu Wang, Zhibin Liu and Yuefeng Ma

School of Computer Science, Qufu Normal University, Rizhao 276826, China
* Correspondence: wxinshui@126.com

Abstract: Future wireless communication systems require higher performance requirements. Based on this, we study the combinatorial optimization problem of power allocation and dynamic user pairing in a downlink multicarrier non-orthogonal multiple-access (NOMA) system scenario, aiming at maximizing the user sum rate of the overall system. Due to the complex coupling of variables, it is difficult and time-consuming to obtain an optimal solution, making engineering impractical. To circumvent the difficulties and obtain a sub-optimal solution, we decompose this optimization problem into two sub-problems. First, a closed-form expression for the optimal power allocation scheme is obtained for a given subchannel allocation. Then, we provide the optimal user-pairing scheme using the actor–critic (AC) algorithm. As a promising approach to solving the exhaustive problem, deep-reinforcement learning (DRL) possesses higher learning ability and better self-adaptive capability than traditional optimization methods. Simulation results have demonstrated that our method has significant advantages over traditional methods and other deep-learning algorithms, and effectively improves the communication performance of NOMA transmission to some extent.

Keywords: NOMA; deep-reinforcement learning; actor–critic; power allocation; user pairing

Citation: Wang, X.; Meng, K.; Wang, X.; Liu, Z.; Ma, Y. Dynamic User Resource Allocation for Downlink Multicarrier NOMA with an Actor–Critic Method. *Energies* **2023**, *16*, 2984. https://doi.org/10.3390/en16072984

Academic Editor: Alicia Triviño-Cabrera

Received: 19 February 2023
Revised: 14 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

1. Introduction

With rapid advances in wireless communication and mobile access technologies, the forthcoming digital society will experience an increasing number of mobile applications. This unprecedented growth in technologies and applications requires sixth-generation wireless systems (6G) with enhanced bandwidth, highly reliable low-latency communication, and massive machine interconnection. Meanwhile, mobile users are also demanding higher transmission capacity and lower network latency [1,2].

Several vital 6G-oriented technologies are being investigated to meet the high requirements of mobile users and the emerging communication technology such as Multiple-Input Multiple-Output (MIMO) [3], Spatial Multiplexing Technology, Intelligent Reflecting Surface (IRS) [4], Unmanned Aerial Vehicle (UAV) communication [5–7], T-Hz communication, etc. In traditional methods, orthogonal resource blocks are allocated to different users to reduce interference between them. In NOMA schemes, signals from different users are coded and modulated at the NOMA system transmitter and directly superimposed together, in the same block of time and frequency resource, and then demodulated using serial interference cancellation (SIC) for the sub-user signals at the receiver [8–10]. Therefore, NOMA can effectively enhance spectral efficiency, enabling massive connectivity and the combining of new technologies.

Machine learning (ML) is used in various fields as an emerging technology in information generation. In a broad sense, ML refers to giving a machine the ability to learn so that it can accomplish functions that cannot be done by programming directly. The key is the machine model, which can make predictions in a practical sense, but before that, the model needs to be trained using data. Reinforcement learning (RL) is a branch of ML, usually described by Markov decision processes (MDP), but it is different from ML.

Mainly reflected in the absence of specific training data, the reward signal is not real-time. The research data of RL are mainly reflected in time series rather than independently distributed data, and the behavior of the current training choice will affect subsequent data. RL is the interaction between a machine and its environment, where the machine receives a desired state and receives a reward from the environment, which changes as a result of the machine's actions [11].

According to the description in Figure 1, the whole RL is an iterative update process. The intelligent agent interacts with the environment while performing a task, and the agent itself generates actions to change the environment. The agent continuously updates its action strategy based on the reward feedback value from the environment, and eventually achieves the task requirements. According to what agents focus on, RL algorithms can be classified into two categories: the optimal strategy (policy-based), the optimal cumulative reward (value-based) and the optimal action at each step (action-based).

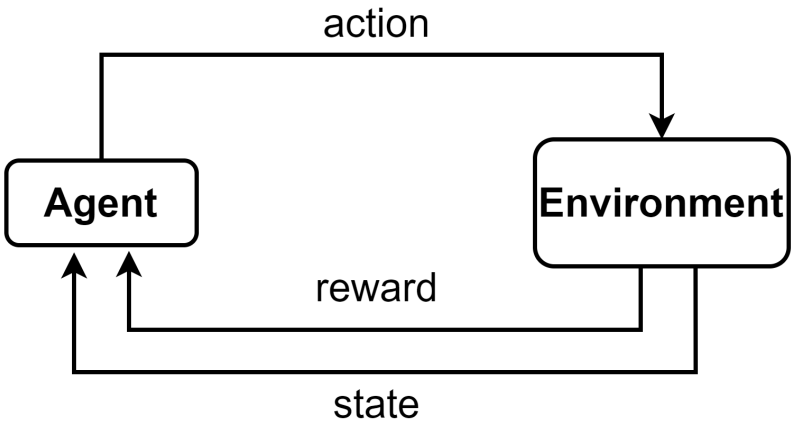

Figure 1. The basic framework of reinforcement learning.

One factor to consider when using this method is how large a problem could be handled. Practical tasks tend to be more complicated, have higher dimensionality, possess more states, and behave in a continuous pattern. In this situation, RL cannot handle complex high-dimensional continuous tasks due to the limitations of its own state space and action space. DRL combines traditional RL with a network structure, maintaining the superiority of RL while increasing the dimension of the input, making it the optimal choice for handling high-dimensional continuous tasks. DRL combines DL's perceptual capability with RL's decision-making capability, and is an artificial intelligence approach that is closer to the human way of thinking [12]. DRL has derived many related algorithms, e.g., Q-learning, Deep Q-Network (DQN), AC, etc. The AC algorithm can perform continuous-action learning, which enables single-step updates, and has higher accuracy and faster discovery of the highest cumulative reward strategy.

1.1. Related Work

In recent years, the issue of resource allocation has attracted increased attention in NOMA scenarios, including the optimization problem of power allocation, user pairing, etc. Numerous pieces of literature are devoted to optimizing transmission power allocation and channel assignment using different optimization methods. For instance, in [13], by analyzing Karush–Kuhn–Tucker (KKT) conditions in the presence of channel error, an optimal allocation solution aiming to maximize performance was derived. A mixed-integer programming problem of joint resource allocation was considered to optimize the communication performance with Quality of Service (QoS) constraint for each user [14]. A joint optimization method was proposed to enhance the sum rate of multiple users in NOMA uplinks, derived closed-form optimal solutions for decoding order and user power allocations, and obtained global optimal solutions using exhaustive search for user

grouping schemes, significantly improving system performance in [15]. For the NOMA downlink with imperfect SIC, an adaptive user-matching algorithm was derived in [16], which used the difference between the signal-to-noise ratio (SNR) of the users to pair users and improve the overall system sum rate. Generally, most channel-allocation schemes have been proven to be NP-hard problems [17]. The computational complexity of these problems is very high and hence it is difficult to obtain their global optimal solution. In multicarrier NOMA scenarios, under the constraints of maximum fairness, maximum weighted sum rate and QoS, a low-complexity joint channel allocation and power distribution algorithm was proposed in [18]. In [19], a channel-to-noise ratio outage threshold was defined, and a global optimal solution for the non-convex optimization problem was derived using the branch-and-bound method.

Traditional schemes lack adaptivity, the computational complexity is relatively high, and the method's efficiency is low. For example, the user-pairing problem is an exhaustive problem, and using exhaustive algorithms consumes more time and costs. Therefore we look for another way to reduce the time complexity. As a potential approach to solve the above NP-hard challenges, DRL has been widely used in communication scenarios or to solve optimization problems [20–22]. Ref. [23] evaluated and tested self-partitioning MIMO cell-free network architecture, performed network segmentation using DRL methods, and implemented a hybrid beamforming model using a new hybrid DRL convex optimization method. Three resource-allocation joint frameworks based on discrete and continuous DQN were proposed to solve the non-convex optimization problem in [24], and deep deterministic policy-gradient (DDPG) network was introduced to overcome the discretization loss, which effectively improved the learning efficiency and reduced the computing time. Ref. [25] optimized the user-pairing problem and the power allocation problem in two steps. First, a confident channel allocation was given to obtain the optimal power allocation scheme, and then DQN was used to find the optimal user-pairing scheme. The DRL framework was used to realize the allocation of network resources and effectively reduce the system transmission energy consumption in [26]. In a hybrid network multi-user scenario, ref. [27] used a distributed DRL algorithm for user power allocation, with fast convergence and higher user reachable rate. The authors in [28] proposed a DRL method with high stochasticity and adaptability for adaptive scheduling and processing of large-scale data, demonstrating the clear advantages of DRL in complex situations.

1.2. Contributions

The channel assignment and user-pairing optimization are combination optimization problems in multicarrier NOMA scenarios. Since the user could not communicate with the base station (BS) solely, and the user-pairing problem is a dynamic assignment and exhaustive problem, the choice of pairing scheme will directly affect the information reachable rate of all users, and then affect the performance of the whole system. Inspired by previous research, we propose a new dynamic pairing scheme for a downlink multicarrier NOMA system, and the summary of this paper is as follows:

- Solving the combinatorial optimization problem in the case of two users in a subchannel of multicarrier NOMA, and we obtain a closed-form solution to represent the optimal resource allocation for users on each subchannel in the corresponding system scenario.
- In the AC framework, we use temporal difference estimation with the addition of baseline as the advantage function in the update gradient to improve the convergence efficiency of the algorithm. Then, we build a NOMA downlink communication scene and embed the DRL algorithm in this scene.
- For the dynamic user-pairing problem, we use the AC algorithm to obtain the optimal pairing scheme that maximizes the total communication rate of all user equipment (UEs). The DRL method provides a new scheme to solve the traditional user-pairing optimization problem. Simulation results have shown that our proposed scheme acquires better performance gains and lower complexity.

Table 1 declares the specific value of the variable used in the simulation experiment. The rest of this paper is organized as follows. In Section 2, we introduce a multicarrier NOMA scenario and derive the combination optimization problem in the case of two users in a subchannel. In Section 3, we use DRL methods to solve the problem of user pairing. In Section 4, we give the simulation results. Finally, we draw conclusions in Section 5.

Table 1. Parameter setting.

Parameters	Values
Distance between user and BS	20–300 m
Distance between each user	<=10 m
Total bandwidth of BS	10 MHz
Total power of the BS	20 W
Path loss	2
Power spectral density	−174 dBm/Hz
QoS	2 bps/Hz
Number of hidden layers	2
Number of neurons in hidden layers	128

2. NOMA System and Optimization Problems

2.1. System Model

We consider a single-cell scenario of a downlink multicarrier NOMA system, as depicted in Figure 2. It is composed of a BS with total bandwidth B and M users distributed randomly around the BS. The BS is in the center of the cell. We assume that the user has a single antenna and the SIC is considered perfect. M_k denotes the number of UE on the k-th subchannel, and UE_m^k denotes the m-th UE on the k-th subchannel. At the BS, the superimposed signal to be sent on the k-th subchannel is expressed as

$$x_k = \sum_{m=1}^{M_k} \sqrt{P_m^k} s_m \tag{1}$$

where P_m^k denotes the power of the m-th UE on the k-th subchannel, and s_m is the signal that needs to be sent to the m-th UE. The superimposed signal received at the receiver is

$$y_m^k = \sqrt{P_m^k} h_m^k s_m + \sum_{n=1;n\neq m}^{M_k} \sqrt{P_n^k} h_m^k s_n + z_m^k \tag{2}$$

$h_m^k = g_m^k d_m^{-\alpha}$ is the channel status information (CSI) from the BS to m-th UE, which g_m^k is the Rayleigh channel coefficient, d is the distance between BS and the m-th UE, α is path loss exponent. In addition, the $z_m^k \sim \mathcal{CN}(0, \sigma_k^2)$ in (2) denotes the additive white Gaussian noise (AWGN). Let $T_m^k = \left|h_m^k\right|^2 / {\sigma_k}^2$ denote the channel noise ratio (CNR). Without losing generality, we assume that the CNR of each user on the k-th channel is ordered as $T_1^k > \cdots > T_m^k > \cdots > T_{M_k}^k$. The goal of the system is to maximize fairness, since the lower the CNR, the worse the channel condition, the more power should be allocated: $p_1^k < \cdots < p_m^k < \cdots < p_{M_k}^k$.

As shown in Figure 3, SIC is used at the receiver side for the NOMA system, for the m-th UE on the k-th channel, the signal of the user with more power is cancelled by SIC, and the signal with less power is considered to be noise. Its *SINR* can be denoted as

$$SINR_m^k = \frac{p_m^k T_m^k}{\sum_{n=1}^{m-1} p_n^k T_m^k + 1} \tag{3}$$

Consequently, the idealized reachable communication rate of the user on the k-th channel can be denoted as

$$R_m^k = B\log_2(1 + \frac{p_m^k T_m^k}{\sum_{n=1}^{m-1} p_n^k T_m^k + 1}) \tag{4}$$

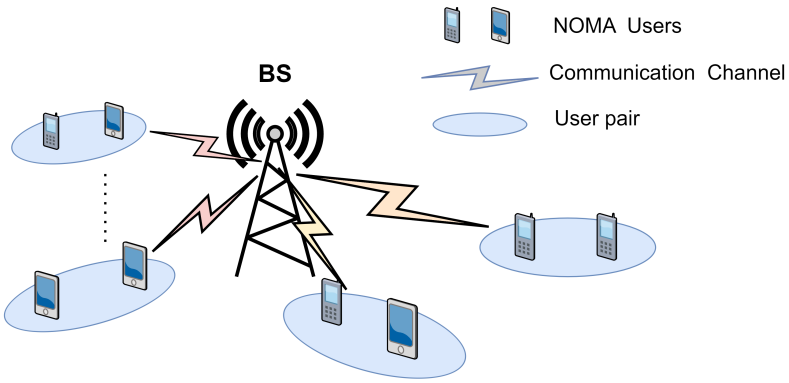

Figure 2. NOMA system model.

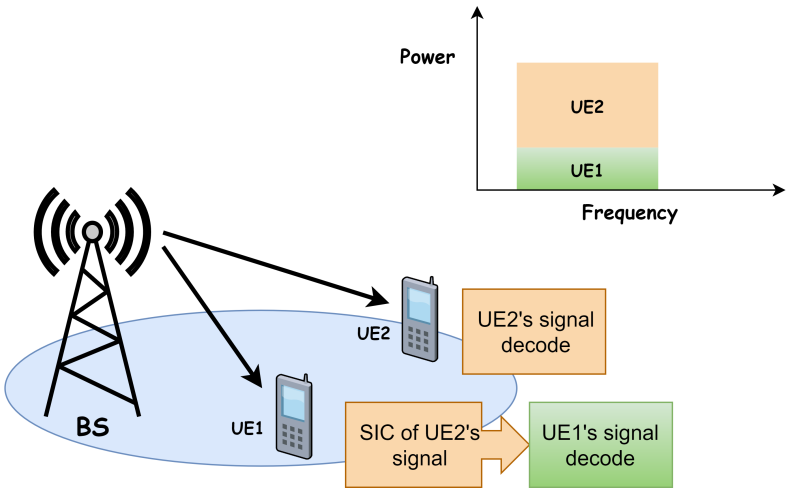

Figure 3. SIC for two users.

Apparently, the more users in the same channel, the higher correlation complexity and calculation latency. For more efficient research, we restrict the existence of only two users for each channel and $T_1^k > T_2^k$. Therefore, the rate for two users on the k-th channel are expressed, respectively, as

$$R_1^k = B\log_2(1 + p_1^k T_1^k) \tag{5}$$

$$R_2^k = B\log_2(1 + \frac{p_2^k T_2^k}{1 + p_1^k T_2^k}) \tag{6}$$

2.2. Optimization Problems

In this subsection, we will discuss the performance optimization problems that are relevant in the downlink multicarrier NOMA system. To enhance the performance of the NOMA system, we solve the problem of maximizing the total sum rate of all users while

satisfying the system constraints. The maximum sum rate (MSR) optimization could be formulated as

$$\max_{T_1^k, T_2^k, p_1^k, p_2^k} \quad \sum_{k=1}^{K} \left[R_1^k\left(T_1^k, p_1^k\right) + R_2^k\left(T_2^k, p_1^k, p_2^k\right) \right] \tag{7}$$

$$s.t. \quad R_m^k \geq \left(R_m^k\right)_{\min} \tag{8}$$

$$\sum_{k=1}^{K} \left(p_1^k + p_2^k\right) \leq P_K \tag{9}$$

$$0 \leq p_1^k \leq p_2^k \tag{10}$$

where $\left(R_m^k\right)_{\min} = B\log_2 A_m^k$, and P_K is the total power of BS. To facilitate the analysis, we first investigate the power allocation problem in a single-carrier NOMA scenario, i.e., a single-channel two-user scenario. This scenario can be represented as a sub-problem of (7) and can be formulated as

$$\max_{T_1^k, T_2^k, p_1^k, p_2^k} \quad \left[R_1^k\left(T_1^k, p_1^k\right) + R_2^k\left(T_2^k, p_1^k, p_2^k\right) \right] \tag{11}$$

$$s.t. \quad R_2^k \geq \left(R_2^k\right)_{\min} \tag{12}$$

$$p_1^k + p_2^k = P^k \tag{13}$$

$$0 \leq p_1^k \leq p_2^k \tag{14}$$

P^k denotes the total power on the k-th channel and is fully allocated to the two users. Let $R = R_1^k + R_2^k$, its derivative for p_1^k is easily verified as $R' = \frac{T_1^k - T_2^k}{\left(1+p_1^k T_1^k\right)\left(1+p_1^k T_2^k\right)} > 0$. As expected, the sum rate R increases with p_1^k. According to (6) and (12), in a realistic communication scenario, user rate also needs to satisfy QoS. Then, we obtain $p_1^k T_2^k + p_2^k T_2^k - p_1^k T_2^k A_2^k \geqslant A_2^k - 1$ and joint (13). Therefore, we obtain $p_1^k \leq \frac{P^k T_2^k - A_2^k + 1}{A_2^k T_2^k}$ based on (6) and constraint (12)–(14), obtain an upper bound on UE_1^k when $p_1^k = \frac{P^k T_2^k - A_2^k + 1}{A_2^k T_2^k}$ and $p_2^k = P^k - p_1^k$.

We have discussed the two-user case of downlink single-carrier NOMA, and, in the following, a multi-user case of multicarrier NOMA will be further studied. We adopt the closed-form solution for problem (11) and give the results here directly:

$$p_1^k = \frac{P^k T_2^k - A_2^k + 1}{A_2^k T_2^k} \tag{15}$$

$$p_2^k = P^k - p_1^k \tag{16}$$

The total power on the k-th channel P^k proposed by [18] is given in a waterfilling form as (for more details, in Appendix A)

$$P^k = \left[\frac{B}{\lambda} - \frac{A_2^k}{T_1^k} + \frac{A_2^k}{T_2^k} - \frac{1}{T_2^k} \right]_{\chi}^{\infty} \tag{17}$$

where $\chi = \frac{A_2^k\left(A_1^k - 1\right)}{T_1^k} + \frac{A_2^k - 1}{T_2^k}$, λ is chosen such that $\sum_{k=1}^{K} P^k = P_K$.

We have obtained the closed-form solution on each subchannel. It can be seen from (15)–(17) that the selection of two users in a single channel is the user-pairing problem, which directly affects (7). Therefore, based on the above results, we will research the pairing solution using the DRL method in the following.

3. Deep-Reinforcement Learning Method for User Pairing

In this scenario, the algorithm convergence time is very short; therefore, we assume that CSI is stable and constant in the process of algorithm training and user pairing. It is essential to maximize system sum rate by pairing the user with a proper scheme. For this exhaustive problem, we use the DRL method to solve it. In the following, we will profile the user-pairing problem based on a DRL and then use the AC algorithm to solve the user-pairing problem. AC applies a higher dimension and a simpler network, allowing the reachable rate of each pairing to be summed and converged towards the policy with the highest cumulative rate.

3.1. Actor–Critic Framework and Advantage Function

As shown in Figure 4, a combined approach, the AC algorithm is divided into two neural networks (NN), actor and critic. The actor is a policy-based function that is responsible for generating actions and interacting with the environment. Policy gradient (PG) as a classical policy-based algorithm, and its gradient can be expressed as (for more details, in Appendix B)

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\Gamma(\tau_n)\nabla\log p_\theta(a_t|s_t) \tag{18}$$

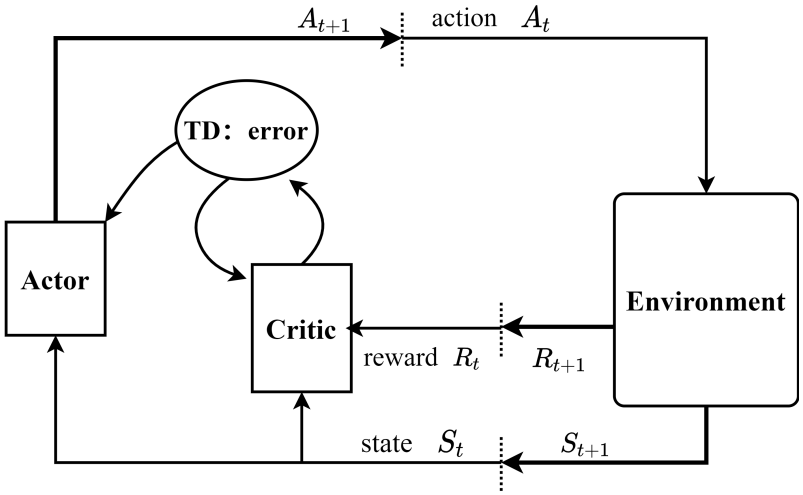

Figure 4. Actor–Critic framework.

It follows that the PG algorithm requires a complete trajectory to achieve the update of the NN parameters, which is based on the Monte Carlo method. The PG is updated in rounds and not in single steps, so the learning efficiency is small and low. Therefore, the AC framework adds another NN, critic, as a value-based function. The gradient in the AC framework is represented as

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\Xi(s_t,a_t)\nabla\log p_\theta(a_t|s_t) \tag{19}$$

where $\Xi(s_t,a_t)$ is generated by the critic, so we do not need to obtain the complete trajectory, but just pass a batch of actions and states into the critic, and the actor could update it according to the new gradient in real time. Meanwhile, $\Xi(s_t,a_t)$ is also the result of the critic's evaluation of this action performed by the actor. The cumulative reward is similar to an evaluator that evaluates the goodness or badness of the selected action.

However, AC has a certain shortcoming: the difficulty of convergence of the critic causes the whole algorithm to be extremely unstable. We introduce the concept of an advantage function, i.e., adding a baseline $V_\theta(s_n)$:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}(\Xi(s_n,a_n)-V_\theta(s_n))\nabla\log p_\theta(a_t|s_t) \tag{20}$$

However, the baseline requires another NN, and to make the critic simpler, we use temporal difference (TD) to estimate the advantage function:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\widehat{A}_\theta\nabla\log p_\theta(a_t|s_t) \tag{21}$$

where $\widehat{A}_\theta = r_t^n + \gamma V_\theta(s_{t+1}^n) - V_\theta(s_t^n)$. Therefore the AC algorithm using this gradient (21) is called Advantage Actor–Critic (A2C).

Before the beginning of training, we initialize the state space and environment. Then we use the configured A2C algorithm to train the NN. The actor is the policy-based function, and the neural network will choose the action according to the current state. The critic network will judge the value of the action so that the actor network will be updated, where the critic uses TD error for the actor value judgment. Then, the critic will update its own network parameters according to the mean square error of TD error.

3.2. Scene-Building

First, we elaborate on the relevant elements in DRL, as illustrated in Figure 5. The agent interacts with the environment as a learner and policymaker of the DRL. The function of the BS in the downlink NOMA system is to coordinate the resource allocation. Therefore, we choose the BS as the agent. The environment is the target of interaction with the agent, and the result of changes in the environment after each interaction is called the state. The agent can only take one action in each step when interacting with the environment. To make the problem fit into the DRL scenario, we will adopt a user-pairing matrix as the environment. All users are divided into two groups in the order of CNR from the smallest to the largest one. We use the first set of rows $\Omega' = \{UE_1, UE_2, \ldots, UE_{M/2}\}$, and the second set of columns $\Omega'' = \left\{UE_{(M/2)+1}, UE_{(M/2)+2}, \ldots, UE_M\right\}$ to denote a pairing matrix. The formed matrix Ω of $M/2 \times M/2$ records user pairing in time. The user pairing matrix is initialized to a zero matrix, and when all user pairings are complete, we call it a training epoch. Each step t in an epoch is a selection of an action, and after performing that action, the state changes from the previous state to the next state $(s_t \to s_{t+1})$, i.e., the state is the current user-pairing matrix. The amount of users is finite, and all the steps in each training epoch are complied with MDP, where the state at each step t can be denoted as $s_t \in \{\Omega_1, \Omega_2, \ldots, \Omega_t, \ldots, \Omega_{M/2}\}$.

The agent needs to choose the appropriate action in the current state, and different actions will have different effects on the environment. In addition, the selection of an action also needs to follow the prescribed policy. So in this DRL scenario, we define the action as $A_t \in \left\{a_t^{(M/2)+1}, a_t^{(M/2)+2}, a_t^{(M/2)+3}, \ldots, a_t^M\right\}$. At step $t(t \in \{1,2,3,\ldots,M/2\})$, if UE_t selects user $\zeta(\zeta \in \{(M/2)+1, (M/2)+2, (M/2)+3, \ldots, M\})$ for pairing, we will set $a_t^\zeta = 1$ otherwise $a_t^\zeta = 0$. After performing each action, the state is updated. Therefore, we use the reward as the performance of that action. In DRL, the purpose of the agent is to discover the strategy that maximizes the cumulative reward through learning. We define the reward as the sum rate of two users $r_t = R_t + R_\zeta$, and the cumulative reward as the total sum rate of system. If there are more than two UEs in a user pair, the reward is set to 0 and the training epoch stops. When all users have been paired, we obtain the cumulative total reward. The following algorithm gives the full process of solving our user-pairing problem by the DRL method.

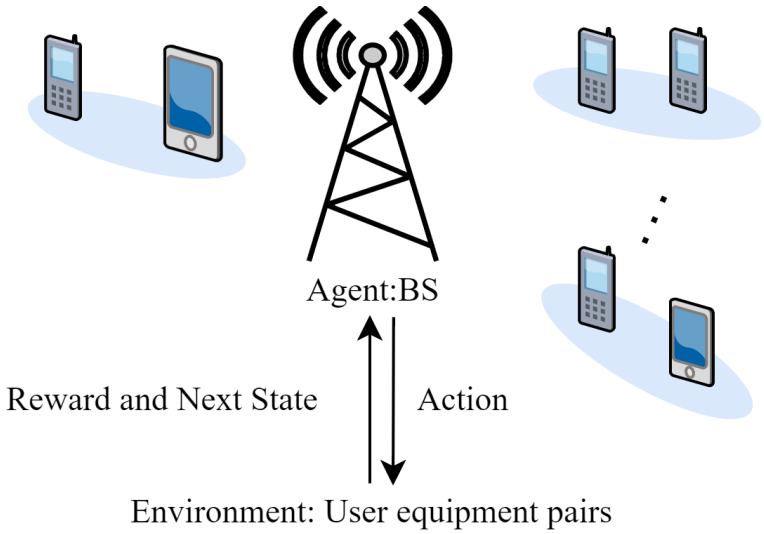

Figure 5. DRL scenario.

4. Simulation

In this section, simulation results are given to verify the feasibility of our proposed method. Since the DQN algorithm is an advanced version of the Q-learning form, which can handle high-dimensional action and state spaces such as A2C, for comparison we show the performance of DQN under the same conditions, and simulate the performance of the scheme for NOMA random pairing and OMA.

Algorithm 1 is implemented by Pytorch 1.9.1 based on Python 3.7, and users are randomly distributed in the area around the BS. The rest of parameters are detailed in Table 1.

Algorithm 1 User Pairing A2C Algorithm

Input: Initial matrix Ω^0; learning rate α
Output: Optimal pairing result Ω^1; The total sum rate R
Initialize: Actor network parameter θ; Critic network parameter λ; Max training epoch T; step $t = 0$
for $1, 2, \ldots T$ **do**
  for *step* $t = 1, 2, \ldots, M/2$ **do**
    Actor takes action $a_t \sim \pi_\theta(a_t|s_t)$ and obtain (s_t, a_t, s_{t+1}, r_t).
    Critic output $V_\lambda(s_t)$ and $V_\lambda(s_{t+1})$.
    Update V_λ by $((r_t + \gamma V_\lambda(s_{t+1})) - V_\lambda(s_t))^2$.
    Evaluate advantage function:
    $\widehat{A}_\theta(s_t, a_t) = r_t + \gamma V_\lambda(s_{t+1}) - V_\lambda(s_t)$.
    Loss: $\nabla_\theta J(\theta) = \widehat{A}_\theta(s_t, a_t) \nabla \log p_\theta(a_t|s_t)$.
    Update π_θ: $\theta^{t+1} = \theta^t + \alpha \nabla_\theta J(\theta)$.
  end
  Record $(s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_{M/2}, a_{M/2}, r_{M/2})$.
  Calculate the total sum rate of system $R = r_1 + r_2 + \cdots + r_{M/2}$.
  Output the optimal pairing matrix Ω^1.
end

When updating the actor parameter θ, we need to set an appropriate learning rate α. For a better increasing trend of the sum rate, we compare it by setting different learning rates. We add 10 users to the scenario, as illustrated in Figure 6. Obviously, when the learning rate is 0.001, the overall learning trend increases steadily with the number of training epochs and a better result is obtained. When the learning rate is set as 0.01, the performance begins to worsen until 1000 epochs, and the increasing trend tends to stop. When the learning rate is set as 0.005, the situation is almost similar to 0.01. Although the result is better than 0.01, there is no convergence trend. When the learning rate is set to 0.0001, the user sum rate has remained stable. On balance, we choose the learning rate of 0.001, and the amount of user equipment is set to 10 for training. As shown in Figure 7, it can be seen that the sum rate of the system reaches a maximum of more than 150 Mbps when the training epoch is around 13,000 times, and the algorithm output results will remain stable in more training epochs and the training effect is satisfactory. The ultimate goal of the algorithm is to find the user pairing that maximizes the sum rate. The sudden drop in the sum rate indicates that the algorithm is still trying other pairing schemes to prevent falling into the local optimal solution, but eventually converges to the strategy with the largest cumulative reward.

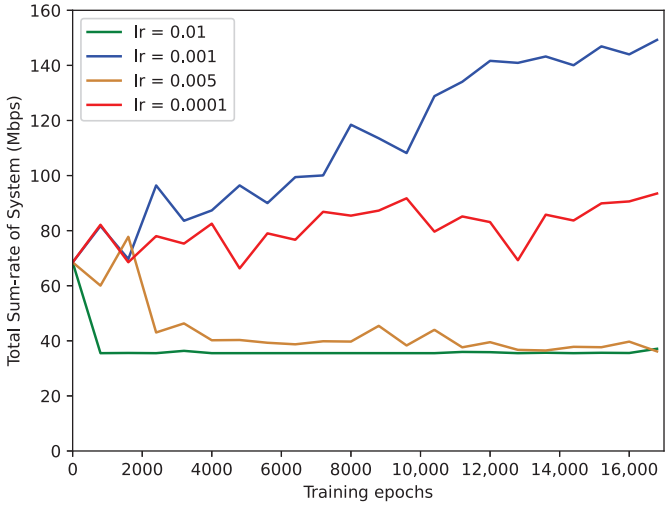

Figure 6. Training results at different learning rates.

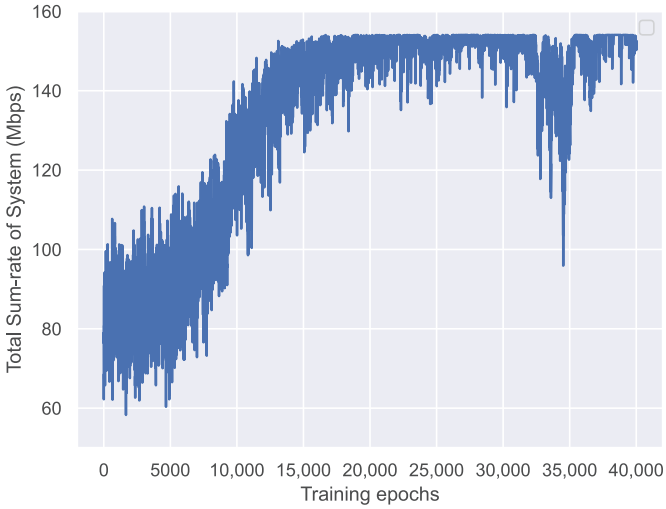

Figure 7. A2C Training result.

The performance of the four algorithms is compared in Figure 8 by increasing the amount of user equipment in the scenario, and it is clear that the NOMA user-pairing algorithm by

DRL outperforms the NOMA random-pairing algorithm and the OMA, and achieves higher performance than DQN with the A2C algorithm. The spectrum efficiency of the NOMA scheme is higher than that of OMA because only one user is assigned to one subchannel of the OMA. However, the output of the NOMA user-pairing algorithm by DRL, trained by the neural network, gradually approaches the optimal user-pairing scheme for that scenario, so the obtained results are better than the NOMA random-pairing algorithm in terms of algorithm efficiency and system performance. When the number of users is 10, the total sum rate of the system for A2C, DQN, NOMA random, OMA are 154.08 Mbps, 130.8 Mbps, 100.74 Mbps and 75.74 Mbps, respectively. The performance of A2C is improved by 17.8% compared to DQN and 52.9% compared to NOMA random.

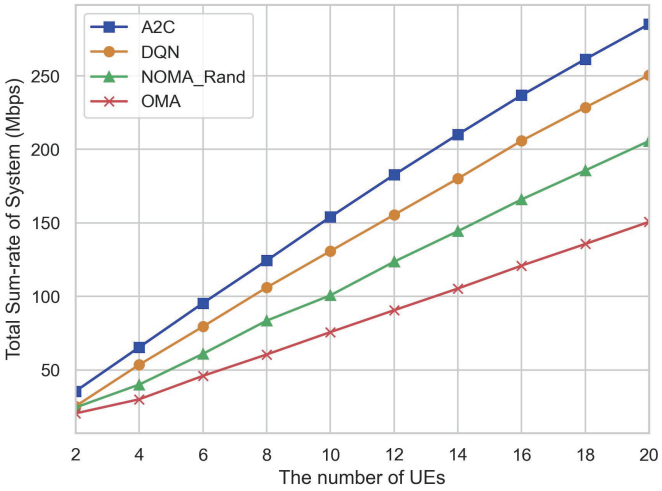

Figure 8. Algorithm performance for different number of users.

To further validate the algorithm performance, we also simulate the sum rate of the system with different algorithms under the scenario with 10 users and 10 Mbps bandwidth of BS. In Figure 9, the first subplot demonstrates the performance of the different algorithms with different BS power. The system sum rate of the four algorithms all increases gradually as the power of BS increments. The NOMA algorithm using DRL still achieves the largest sum rate. With the same settings as the first subplot and the fixed BS power (20 W), we examine the impact of the available bandwidth of BS on the total sum rate of the system. In the second subplot, the bandwidth of BS has a stronger impact than the BS power, and the NOMA algorithm using DRL is still optimal.

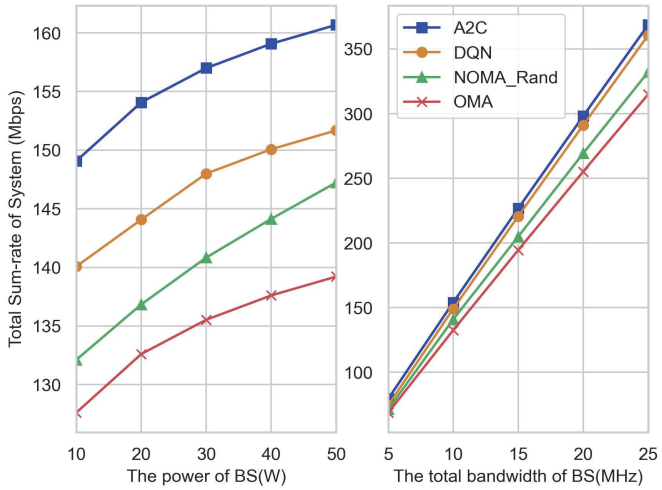

Figure 9. Different BS power and bandwidth.

In Figure 10, we depict the effect of noise on the algorithm and the system sum rate. It can be seen that a higher system sum rate is obtained with smaller noise power spectral density, and, as expected, the effect of SNR on the system performance is significant. The system sum rate that each algorithm can reach decreases as the noise power spectral density increases, but the performance of the NOMA algorithm using DRL is still optimal compared to the other two algorithms.

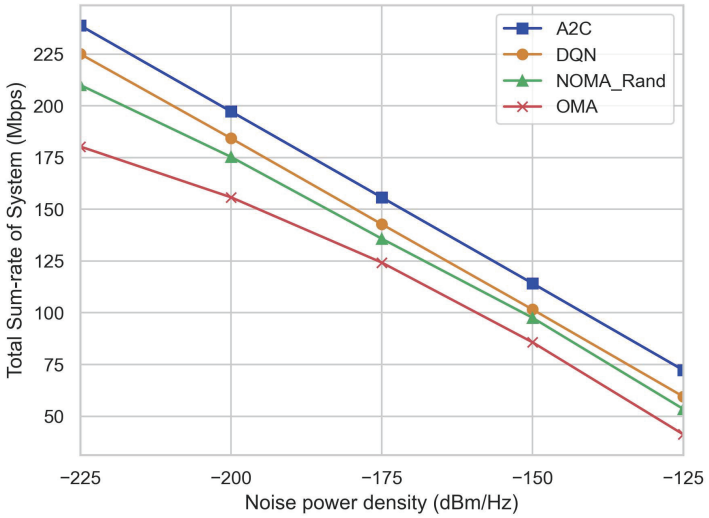

Figure 10. Performance of different algorithms at different noise power density.

5. Conclusions

In this article, we propose a dynamic user-pairing scheme in a multicarrier downlink NOMA system. Specifically, depending on the closed-form solution of the subchannel, we use the A2C algorithm to deal with the dynamic user-pairing problem. The work in this paper: 1. Deriving the combination optimization problem in the case of two users in a subchannel of multicarrier NOMA under the conditions of perfect CSI, and obtaining the closed-form solution of power allocation for subchannel by decomposing this problem. 2. The communication scenario is transformed into a DRL scenario, and the user-pairing problem is processed using the A2C algorithm to rapidly search for the optimal user-pairing scheme, which has lower complexity compared to the exhaustive search. The simulation results show that the A2C algorithm has significant advantages over the traditional NOMA random-pairing approach and OMA. For DQN, A2C uses a higher dimension, a simpler neural network, and is superior in terms of performance improvement.

Author Contributions: Conceptualization, X.W. (Xinshui Wang) and K.M.; methodology, X.W. (Xinshui Wang); software, K.M., X.W. (Xu Wang) and Z.L.; validation, X.W. (Xu Wang), Z.L. and Y.M.; writing—original draft preparation, X.W. (Xinshui Wang) and K.M.; writing—review and editing, X.W. (Xinshui Wang), K.M. and X.W. (Xu Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Natural Science Foundation of Shandong Province of China (Grant No. ZR2021MF013 and Grant No. ZR2021MF124).

Data Availability Statement: Not applicable.

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

Appendix A. The Derivation for the Power of the k-th Channel

When the rate of UE_1^k and UE_2^k are both minimum rates, $\left(R_1^k\right)_{\min} = B\log_2 A_1^k$ and $\left(R_2^k\right)_{\min} = B\log_2 A_2^k$, joint with (5) and (6) we can obtain the minimum power of the k-th

channel ${P^k}_{\min} = \frac{A_2^k\left(A_1^k-1\right)}{T_1^k} + \frac{A_2^k-1}{T_2^k}$. Therefore the power of the k-th channel optimization problem is as follows

$$\max_{P^k} \quad \sum_{k=1}^{K} R\left(P^k\right) = R_1^k\left(p_1^k\right) + \left(R_2^k\right)_{\min} \tag{A1}$$

$$s.t. \quad \sum_{k=1}^{K} P^k \leqslant P_K \tag{A2}$$

$$P^k \geqslant {P^k}_{\min} \tag{A3}$$

where $R_1^k\left(p_1^k\right) = B\log_2\left(\frac{A_2^kT_2^k - A_2^kT_1^k + T_1^kT_2^kP^k + T_1^k}{A_2^kT_2^k}\right)$. Obviously, $R\left(P^k\right)$ is a concave function and this optimization problem is a convex problem. Therefore, the Lagrangian function is expressed as

$$L\left(P^k\right) = R_1^k\left(p_1^k\right) + \left(R_2^k\right)_{\min} + \lambda\left({P^k}_{\min} - P^k\right) \tag{A4}$$

Then

$$\frac{\partial L}{\partial P^k} = \frac{BT_1^kT_2^k}{A_2^kT_2^k - A_2^kT_1^k + T_1^kT_2^kP^k + T_1^k} - \lambda = 0 \tag{A5}$$

We use the waterfilling form to show the result

$$P^k = \left[\frac{B}{\lambda} - \frac{A_2^k}{T_1^k} + \frac{A_2^k}{T_2^k} - \frac{1}{T_2^k}\right]_{\chi}^{\infty} \tag{A6}$$

where $\chi = {P^k}_{\min} = \frac{A_2^k\left(A_1^k-1\right)}{T_1^k} + \frac{A_2^k-1}{T_2^k}$. Equation (A6) represents the range of values of P^k, the specific value of P^k depends on λ, and λ is chosen such that $\sum_{k=1}^{K} P^k = \sum_{k=1}^{K}\left(p_1^k + p_2^k\right) = P_K$.

Appendix B. The Derivation for the Gradient of Policy Gradient

We give an actor $\pi_\theta(s)$ with neural network parameter θ, is interacting with the environment, and start with observation s_1. Actor decides to take a_1, obtains reward r_1. An episode is considered to be a trajectory $\tau = \{s_1, a_1, r_1, s_2, a_2, r_2 \cdots s_T, a_T, r_T\}$. So the total reward in this episode is $\Gamma(\tau) = \sum_{t=1}^{T} r_t$ and we define $\overline{\Gamma}_\theta$ as the expected value of Γ_θ which evaluates the goodness of an actor $\pi_\theta(s)$. Then the probability that we obtain the same trajectory in the following learning episode can be expressed as

$$\begin{aligned} P_\theta(\tau) &= p(s_1)p_\theta(a_1|s_1)p(r_1, s_2|s_1, a_1)p_\theta(a_2|s_2)\cdots \\ &= p(s_1)\prod_{t=1}^{T} p_\theta(a_t|s_t)p(r_t, s_{t+1}|s_t, a_t) \end{aligned} \tag{A7}$$

where P_θ is the probability of producing a certain action in NN θ and the transition probability follows Markov property, i.e., the state transition is only related to the last state. Assume that we produce a total of N trajectories $\{\tau_1, \tau_2, \tau_3, \cdots \tau_N\}$, so $\overline{\Gamma}_\theta$ can be expressed as

$$\overline{\Gamma}_\theta = \sum_{\tau}\Gamma(\tau)P_\theta(\tau) = \frac{1}{N}\sum_{n=1}^{N}\Gamma(\tau_n) \tag{A8}$$

Since our target is to maximize the expected reward value, taking a gradient ascend for θ, the problem is represented as

$$\theta^* = \arg\max_{\theta} \overline{\Gamma}_\theta \tag{A9}$$

$$\theta^{new} = \theta^{old} + \alpha \nabla \overline{\Gamma}_{\theta^{old}} \tag{A10}$$

where α is the learning rate and the gradient of $\overline{\Gamma}_\theta$:

$$\begin{aligned} \nabla \overline{\Gamma}_\theta &= \sum_{\tau} \Gamma(\tau) \nabla P_\theta(\tau) \\ &= \sum_{\tau} \Gamma(\tau) P_\theta(\tau) \frac{\nabla P_\theta(\tau)}{P_\theta(\tau)} \\ &= \sum_{\tau} \Gamma(\tau) P_\theta(\tau) \nabla \log P_\theta(\tau) \\ &= \frac{1}{N} \sum_{n=1}^{N} \Gamma(\tau_n) \nabla \log P_\theta(\tau_n) \\ &= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \Gamma(\tau_n) \nabla \log p_\theta(a_t | s_t) \end{aligned} \tag{A11}$$

References

1. David, K.; Berndt, H. 6g vision and requirements: Is there any need for beyond 5g? *IEEE Veh. Technol. Mag.* **2018**, *13*, 72–80. [CrossRef]
2. Saad, W.; Bennis, M.; Chen, M. A vision of 6g wireless systems: Applications, trends, technologies, and open research problems. *IEEE Netw.* **2020**, *34*, 134–142. [CrossRef]
3. Gong, Y.; Zhang, L.; Liu, R.; Yu, K.; Srivastava, G. Nonlinear mimo for industrial internet of things in cyber–physical systems. *IEEE Trans. Ind. Informatics* **2021**, *17*, 5533–5541. [CrossRef]
4. Jiang, T.; Cheng, H.V.; Yu, W. Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 1931–1945. [CrossRef]
5. Zeng, Y.; Zhang, R.; Lim, T.J. Wireless communications with unmanned aerial vehicles: Opportunities and challenges. *IEEE Commun.* **2016**, *54*, 36–42. [CrossRef]
6. Zhao, L.; Yang, K.; Tan, Z.; Song, H.; Al-Dubai, A.; Zomaya, A.Y.; Li, X. Vehicular computation offloading for industrial mobile edge computing. *IEEE Trans. Ind. Informatics* **2021**, *17*, 7871–7881. [CrossRef]
7. Zhao, L.; Yang, K.; Tan, Z.; Li, X.; Sharma, S.; Liu, Z. A novel cost optimization strategy for sdn-enabled uav-assisted vehicular computation offloading. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 3664–3674. [CrossRef]
8. Dai, L.; Wang, B.; Ding, Z.; Wang, Z.; Chen, S.; Hanzo, L. A survey of non-orthogonal multiple access for 5g. *IEEE Commun. Surv. Tutorials* **2018**, *20*, 2294–2323. [CrossRef]
9. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [CrossRef]
10. Benjebbour, A.; Saito, Y.; Kishiyama, Y.; Li, A.; Harada, A.; Nakamura, T. Concept and practical considerations of non-orthogonal multiple access (noma) for future radio access. In Proceedings of the 2013 International Symposium on Intelligent Signal Processing and Communication Systems, Naha, Japan, 12–15 November 2013; pp. 770–774.
11. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process.* **2017**, *34*, 26–38. [CrossRef]
12. Zhou, F.; Lu, G.; Wen, M.; Liang, Y.C.; Chu, Z.; Wang, Y. Dynamic spectrum management via machine learning: State of the art, taxonomy, challenges, and open research issues. *IEEE Netw.* **2019**, *33*, 54–62. [CrossRef]
13. Zamani, M.R.; Eslami, M.; Khorramizadeh, M. Optimal sum-rate maximization in a noma system with channel estimation error. In Proceedings of the Electrical Engineering (ICEE), Iranian Conference, Mashhad, Iran, 8–10 May 2018; pp. 720–724.
14. Zhu, L.; Zhang, J.; Xiao, Z.; Cao, X.; Wu, D.O. Optimal user pairing for downlink non-orthogonal multiple access (noma). *IEEE Wireless Commun. Lett.* **2019**, *8*, 328–331. [CrossRef]
15. Zhang, J.; Zhu, L.; Xiao, Z.; Cao, X.; Wu, D.O.; Xia, X. Optimal and sub-optimal uplink noma: Joint user grouping, decoding order, and power control. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 254–257. [CrossRef]
16. Mouni, N.S.; Kumar, A.; Upadhyay, P.K. Adaptive user pairing for noma systems with imperfect sic. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 1547–1551. [CrossRef]
17. Liang, G.; Zhu, Q.; Xin, J.; Feng, Y.; Zhang, T. Joint user-channel assignment and power allocation for non-orthogonal multiple access relaying networks. *IEEE Access* **2019**, *7*, 361–372. [CrossRef]
18. Zhu, J.; Wang, J.; Huang, Y.; He, S.; You, X.; Yang, L. On optimal power allocation for downlink non-orthogonal multiple access systems. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2744–2757. [CrossRef]

19. Xiao, L.; Li, Y.; Dai, C.; Dai, H.; Poor, H.V. Reinforcement learning-based noma power allocation in the presence of smart jamming. *IEEE Trans. Veh. Technol.* **2018**, *67*, 3377–3389. [CrossRef]
20. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Commun. Surv. Tutorials* **2019**, *21*, 3133–3174. [CrossRef]
21. Kibria, M.G.; Nguyen, K.; Villardi, G.P.; Zhao, O.; Ishizu, K.; Kojima, F. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *IEEE Access* **2018**, *6*, 328–338. [CrossRef]
22. Lin, T.; Zhu, Y. Beamforming design for large-scale antenna arrays using deep learning. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 103–107. [CrossRef]
23. Al-Eryani, Y.; Hossain, E. Self-organizing mmwave mimo cell-free networks with hybrid beamforming: A hierarchical drl-based design. *IEEE Trans. Commun.* **2022**, *70*, 3169–3185. [CrossRef]
24. Wang, X.; Zhang, Y.; Shen, R.; Xu, Y.; Zheng, F. Drl-based energy-efficient resource allocation frameworks for uplink noma systems. *IEEE Internet Things J.* **2020**, *7*, 7279–7294. [CrossRef]
25. Jiang, F.; Gu, Z.; Sun, C.; Ma, R. Dynamic user pairing and power allocation for noma with deep reinforcement learning. In Proceedings of the 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China, 29 March–1 April 2021.
26. Zhang, X.; Yu, P.; Feng, L.; Zhou, F.; Li, W. A drl-based resource allocation framework for multimedia multicast in 5g cellular networks. In Proceedings of the 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Jeju, Republic of Korea, 5–7 June 2019.
27. Ciftler, B.S.; Alwarafy, A.; Abdallah, M. Distributed drl-based downlink power allocation for hybrid rf/vlc networks. *IEEE Photon. J.* **2022**, *14*, 8632510. [CrossRef]
28. Lu, K.; Liu, X.; Ai, Z.; Liu, Z.; Tao, D.; Lou, S. A drl based real-time computing load scheduling method. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Penghu, Taiwan, 15–17 September 2021.