applied sciences

MDPI

Article

Hybrid NOMA/OMA-Based Dynamic Power Allocation Scheme Using Deep Reinforcement Learning in 5G Networks

Hoang Thi Huong Giang [1], Tran Nhut Khai Hoan [2], Pham Duy Thanh [1] and Insoo Koo [1,*]

[1] School of Electrical Engineering, University of Ulsan, Ulsan 44610, Korea; huonggiangtdt@gmail.com (H.T.H.G.); duythanhtdtk12@gmail.com (P.D.T.)
[2] College of Engineering Technology, Can Tho University, Can Tho 94000, Vietnam; tnkhoan@ctu.edu.vn
* Correspondence: iskoo@ulsan.ac.kr; Tel.: +82-52-259-1249

Received: 3 June 2020; Accepted: 17 June 2020; Published: 20 June 2020

Abstract: Non-orthogonal multiple access (NOMA) is considered a potential technique in fifth-generation (5G). Nevertheless, it is relatively complex when applying NOMA to a massive access scenario. Thus, in this paper, a hybrid NOMA/OMA scheme is considered for uplink wireless transmission systems where multiple cognitive users (CUs) can simultaneously transmit their data to a cognitive base station (CBS). We adopt a user-pairing algorithm in which the CUs are grouped into multiple pairs, and each group is assigned to an orthogonal sub-channel such that each user in a pair applies NOMA to transmit data to the CBS without causing interference with other groups. Subsequently, the signal transmitted by the CUs of each NOMA group can be independently retrieved by using successive interference cancellation (SIC). The CUs are assumed to harvest solar energy to maintain operations. Moreover, joint power and bandwidth allocation is taken into account at the CBS to optimize energy and spectrum efficiency in order to obtain the maximum long-term data rate for the system. To this end, we propose a deep actor-critic reinforcement learning (DACRL) algorithm to respectively model the policy function and value function for the actor and critic of the agent (i.e., the CBS), in which the actor can learn about system dynamics by interacting with the environment. Meanwhile, the critic can evaluate the action taken such that the CBS can optimally assign power and bandwidth to the CUs when the training phase finishes. Numerical results validate the superior performance of the proposed scheme, compared with other conventional schemes.

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

1. Introduction

Recently, fourth-generation (4G) systems reached maturity, and will evolve into fifth-generation (5G) systems where limited amounts of new spectrum can be utilized to meet the stringent demands of users. However, critical challenges will come from explosive growth in devices and data volumes, which require more efficient exploitation of valuable spectrum. Therefore, non-orthogonal multiple access (NOMA) is one of the potential candidates for 5G and upcoming cellular network generations [1–3].

According to NOMA principles, multiple users are allowed to share time and spectrum resources in the same spatial layer via power-domain multiplexing, in contrast to conventional orthogonal multiple access (OMA) techniques consisting of frequency-division multiple access (FDMA) and time division multiple access (TDMA) [4]. Interuser interference can be alleviated by performing successive interference cancellation (SIC) on the receiver side. There has been a lot of research aimed at sum

rate maximization, and the results showed that higher spectral efficiency (SE) can be obtained by using NOMA, compared to baseline OMA schemes [5–8]. Zeng et al. [5] investigated a multiple-user scenario in which users are clustered and share the same transmit beamforming vector. Di et al. [6] proposed a joint sub-channel assignment and power allocation scheme to maximize the weighted total sum rate of the system while adhering to a user fairness constraint. Timotheou et al. [7] studied a decoupled problem of user clustering and power allocation in NOMA systems in which the proposed user clustering approach is based on exhaustive search with a high required complexity. Liang et al. [8] studied solutions for user pairing, and investigated the power allocation problem by using NOMA in cognitive radio (CR) networks.

Nowadays, energy consumption for wireless communications is becoming a major social and economic issue, especially with the explosive amounts of data traffic. However, limited efforts have been devoted to the energy-efficient resource allocation problem in NOMA-enabled systems [9–11]. The authors in [9] maximized energy efficiency subject to a minimum required data rate for each user, which leads to a nonconvex fractional programming problem. Meanwhile, a power allocation solution aiming to maximize the energy efficiency under users' quality of service requirements was investigated [10]. Fang et al. [11] proposed a gradient-based binary search power allocation approach for downlink NOMA systems, but it requires high complexity. NOMA was also applied to future machine-to-machine (M2M) communications in [12], and it was shown that the outage probability of the system can be improved when compared with OMA. Additionally, by jointly studying beamforming, user scheduling, and power allocation, the system performance of millimeter wave (mmWave) networks was studied [13].

On the other hand, CR (one of the promising techniques to improve SE), has been extensively investigated for decades. In it, cognitive users (CUs) can utilize the licensed spectrum bands of the primary users (PUs) as long as the interference caused by the CUs is tolerable [14–16]. Goldsmith et al. in [17] proposed three operation models (opportunistic spectrum access, spectrum sharing, and sensing-based enhanced spectrum sharing) to exploit the CR technique in practice. It is conceivable that the combination of CR with NOMA technologies is capable of further boosting the SE in wireless communication systems. Therefore, many studies on the performance of spectrum-sharing CR combined with NOMA have been analyzed [18,19].

Along with the rapid proliferation of wireless communication applications, most battery-limited devices become useless if their battery power is depleted. As one of the remedies, energy harvesting (EH) exploits ambient energy resources to replenish batteries, such as solar energy [20], radio frequency (RF) signals [21], and both non-RF and RF energy harvesting [22], etc. Among various kinds of renewable energy resources, solar power has been considered one of the most effective energy sources for wireless devices. However, solar power density highly depends on the environment conditions, and may vary over time. Thus, it is critical to establish proper approaches to efficiently utilize harvested energy for wireless communication systems.

Many early studies regarding NOMA applications have mainly focused on the downlink scenario. However, there are fewer contributions investigating uplink NOMA, where an evolved NodeB (eNB) has to receive different levels of transmitted power from all user devices using NOMA. Zhang et al. in [23] proposed a novel power control scheme, and the outage probability of the system was derived. Besides, the user-pairing approach was studied in many predefined power allocation schemes in NOMA communication systems [24] in which internet of things (IoT) devices first harvest energy from BS transmissions in the harvesting phase, and they then utilize the harvested energy to perform data transmissions using the NOMA technique during the transmission phase. The pricing and bandwidth allocation problem in terms of energy efficiency in heterogeneous networks was investigated in [25]. In addition, the authors in [20] proposed joint resource allocation and transmission mode selection to maximize the secrecy rate in cognitive radio networks. Nevertheless, most of the existing work on resource allocation assumes that the amount of harvested energy is known, or that traffic loads are predictable, which is hard to obtain in practical wireless networks.

Since the information regarding network dynamics (e.g., harvested energy distribution, primary user's behavior) is sometimes unavailable in the cognitive radio system, researchers usually formulate optimization problems as the framework of a Markov decision process (MDP) [20,22,26,27]. Reinforcement learning is one of the potential approaches to obtaining the optimal solution for an MDP problem by interacting with the environment without having prior information about the network dynamics or without any supervision [28–30]. However, it is a big issue for reinforcement learning to have to deal with large-state-space optimization problems. For this reason, deep reinforcement learning (DRL) is being investigated extensively these days in wireless communication systems where deep neural networks (DNNs) work as function approximators and are utilized to learn the optimal policy [31–33]. Meng et al. proposed a deep reinforcement learning method for a joint spectrum sensing and power control problem in a cognitive small cell [31]. In addition, deep Q-learning was studied for a wireless gateway that is able to derive the optimal policy to maximize throughput in cognitive radio networks [32]. Zhang et al. [33] proposed an asynchronous advantage, deep actor-critic-based scheme to optimize spectrum sharing efficiency and guarantee the QoS requirements of PUs and CUs.

To the best of our knowledge, there has been little research into resource allocation using deep reinforcement learning under a non-RF energy-harvesting scenario in uplink cognitive radio networks. Thus, we propose a deep actor-critic reinforcement learning framework for efficient joint power and bandwidth allocation by adopting hybrid NOMA/OMA in uplink cognitive radio networks (CRNs). In them, solar energy-powered CUs are assigned the proper transmission power and bandwidth to transmit data to a cognitive base station in every time slot in order to maximize the long-term data transmission rate of the system. Specifically, the main contributions of this paper are as follows.

- We study a model of a hybrid NOMA/OMA uplink cognitive radio network adopting energy harvesting at the CUs, where solar energy-powered CUs opportunistically use the licensed channel of the primary network to transmit data to a cognitive base station using NOMA/OMA techniques. Beside that, a user-pairing algorithm is adopted such that we can assign orthogonal frequency bands to each NOMA group after pairing. We take power and bandwidth allocation into account such that the transmission power and bandwidth are optimally utilized by each CU under energy constraints and environmental uncertainty. The system is assumed to work on a time-slotted basis.
- We formulate the problem of long-term data transmission rate maximization as the framework of a Markov decision process (MDP), and we obtain the optimal policy by adopting a deep actor-critic reinforcement learning (DACRL) framework under a trial-and-error learning algorithm. More specifically, we use DNNs to approximate the policy function and the value function for the actor and critic components, respectively. As a result, the cognitive base station can allocate the appropriate transmission power and bandwidth to the CUs by directly interacting with the environment, such that the system reward can be maximized in the long run by using the proposed algorithm.
- Lastly, extensive numerical results are provided to assess the proposed algorithm performance through diverse network parameters. The simulation results of the proposed scheme are shown to be superior to conventional schemes where decisions on transmission power and bandwidth allocation are taken without long-term considerations.

The rest of this paper is structured as follows. The system model is presented in Section 2. We introduce the problem formulation in Section 3, and we describe the deep actor-critic reinforcement learning scheme for resource allocation in Section 4. The simulation results and discussions are in Section 5. Finally, we conclude the paper in Section 6.

2. System Model

We consider an uplink CRN that consists of a cognitive base station (CBS), a primary base station (PBS), multiple primary users, and $2M$ cognitive users as illustrated in Figure 1. Each CU is outfitted

with a single antenna to transmit data to the CBS, and each is equipped with an energy-harvesting component (i.e., solar panels). The PBS and PUs have the license to use the primary channel at will. However, they do not always have data to transmit on the primary channel. Meanwhile, the CBS and the CUs can opportunistically utilize the primary channel by adopting a hybrid NOMA/OMA technique when the channel is sensed as free. To this end, the CBS divides the set of CUs into pairs according to Algorithm 1 where the CU having the highest channel gain will be coupled with the CU having the lowest channel gain, and one of available channels will be assigned to these pairs. More specifically, the CUs are arranged into M NOMA groups, and the primary channel is divided into multiple subchannels to apply hybrid NOMA/OMA for the transmissions between the CUs and the CBS, with $\mathcal{G} = \{G_1, G_2, G_3, ..., G_M\}$ denoting the set of NOMA groups. Additionally, M NOMA groups are assigned to M orthogonal subchannels, $\mathcal{SC} = \{SC_1, SC_2, SC_3, ..., SC_M\}$, of the primary channel such that the CUs in each NOMA group can transmit on the same subchannel and will not interfere with the other groups. In this paper, successive interference cancellation (SIC) [34] is applied at the CBS for decoding received signals, which are transmitted from the CUs. Moreover, we assume that the CUs always have data to transmit, and the CBS has complete channel state information (CSI) of all the CUs.

Algorithm 1 User-pairing Algorithm

1: **Input:** channel gain, number of groups, M, number of CUs, $2M$.
2: Sort the channel gain of all CUs in decending order: $g_1 \geq g_2 \geq ... \geq g_{2M}$
3: Define set of channel gains $U = \{g_1, g_2, ..., g_{2M}\}$
4: **for** $j = 1 : M$
5: $\quad G_j = \{\emptyset\}$
6: $\quad G_{\max} = \max\{U\}, G_{\min} = \min\{U\}$
7: $\quad G_j = G_j \cup G_{\max} \cup G_{\min}$
8: $\quad U = U \backslash G_{\max} \backslash G_{\min}$
9: **end for**
10: **Output:** Set of CU pairs.

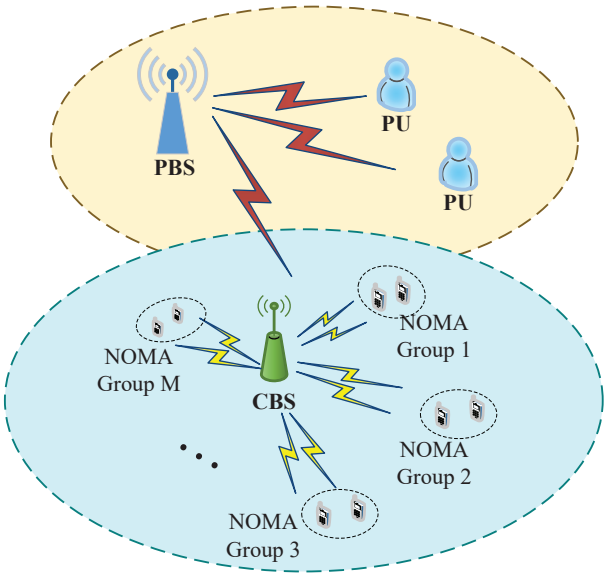

Figure 1. System model of the proposed scheme.

The network system operation is illustrated in Figure 2. In particular, at the beginning of a time slot, with duration τ_{ss}, all CUs concurrently perform spectrum sensing and report their local results to the CBS. Based on these sensing results, the CBS first decides the global sensing result as to whether

the primary channel is busy or not following the combination rule [35,36], and then allocates power and bandwidth to all CUs for uplink data transmission. As a consequence, according to the allocated power and bandwidth of the NOMA groups, the CUs in each NOMA group can transmit their data to the CBS through the same subchannel without causing interference with other groups within duration $\tau_{Tr} = T_{tot} - \tau_{ss}$, where T_{tot} is the total time slot duration. Information regarding the remaining energy in all the CUs is updated to the CBS at the end of each time slot. Each data transmission session of the CUs may take place in more than one time slot until all their data have been transmitted successfully.

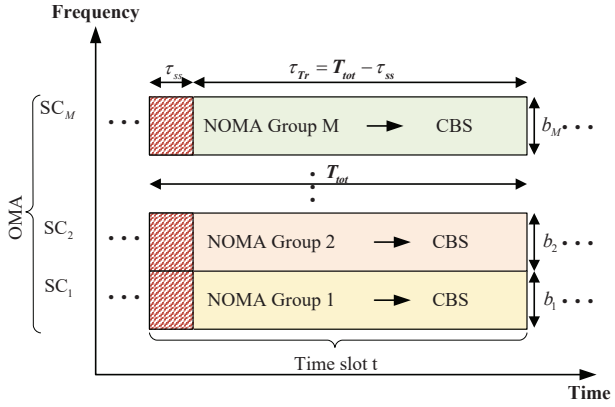

Figure 2. Time frame of the cognitive users' operations.

During data transmission, the received composite signal at the CBS on subchannel SC_m is given by

$$y_m(t) = \sqrt{P_{1m}(t)}x_{1m}(t)h_{1m} + \sqrt{P_{2m}(t)}x_{2m}(t)h_{2m} + \omega_m, \quad (1)$$

where $P_{im}(t) = e_{im}^{tr}(t)/\tau_{Tr} \,|\, i \in \{1,2\}, m \in \{1,2,...,M\}$ is the transmission power of CU_i in NOMA group G_m, in which $e_{im}^{tr}(t)$ is the transmission energy assigned for CU_{im} in time slot t; $x_{im}(t)$ denotes the transmit signal of CU_{im} in time slot t, $\left(\mathbb{E}\{|x_{im}(t)|^2\} = 1\right)$; ω_m is the additive white Gaussian noise (AWGN) at the CBS on subchannel SC_m with zero mean and variance σ^2; and h_{im} is the channel coefficient between CU_{im} and the CBS. The overall received signal at the CBS in time slot t is given by

$$y(t) = \sum_{m=1}^{M} y_m(t). \quad (2)$$

The received signals at the CBS on different sub-channels are independently retrieved from composite signal $y_m(t)$ using the SIC technique. In particular, the CU's signal with the highest channel gain is firstly decoded, and then it will be removed from composite signal at the CBS, in a successive manner. The CU's signal with the lower channel gain in the sub-channel is treated as noise of the CU with the higher channel gain. We assume perfect SIC implementation at the CBS. The achievable transmission rate for the CUs in NOMA group G_m are

$$\begin{aligned} R_{1m}(t) &= \frac{\tau_{Tr}}{T_{tot}} \times b_m(t) \times \log_2\left[1 + \frac{P_{1m}(t)g_{1m}}{P_{2m}(t)g_{2m}+\sigma^2}\right] \\ R_{2m}(t) &= \frac{\tau_{Tr}}{T_{tot}} \times b_m(t) \times \log_2\left[1 + \frac{P_{2m}(t)g_{2m}}{\sigma^2}\right], \end{aligned} \quad (3)$$

where $b_m(t)$ is the amount of bandwidth allocated to subchannel SC_m in time slot t, $g_{im} = |h_{im}|^2$ is the channel gain of CU_{im} on subchannel m, and $g_{1m} \geq g_{2m}$. Since the channel gain of CU_{1m}, g_{1m}, is higher, CU_{1m} has a higher priority for decoding. Consequently, the signal of CU_{1m} is decoded first by treating the signal of CU_{2m} as interference. Next, user CU_{1m} is removed from signal $y_m(t)$, and the signal of user CU_{2m} is decoded as interference-free. The sum achievable transmission rate of NOMA group G_m can be calculated as:

$$R_m(t) = R_{1m}(t) + R_{2m}(t). \tag{4}$$

The sum achievable transmission rate at the CBS can be given as follows:

$$R(t) = \sum_{m=1}^{M} R_m(t). \tag{5}$$

Energy Arrival and Primary User Models

In this paper, the CUs have a finite capacity battery, E_{bat}, which can be constantly recharged by the solar energy harvesters. Therefore, the CUs can perform their other operations and harvest solar energy simultaneously. For many reasons (such as the weather, the season, different times of the day), the harvested energy from solar resources may vary in practice. Herein, we take into account a practical case, where the harvested energy of CU_i in NOMA group G_m (denoted as e_{im}^h) follows a Poisson distribution with mean value ξ_{avg}, as studied in [37]. The arrival energy amount that CU_{im} can harvest during time slot t can be given as $e_{im}^h(t) \in \left\{e_1^h, e_2^h, ..., e_v^h\right\}$ where $0 < e_1^h < e_2^h < ... < e_v^h < E_{bat}$. The cumulative distribution function can be given as follows:

$$F\left(e_{im}^h(t); \xi_{avg}\right) = \sum_{k=0}^{e_{im}^h(t)} e^{-\xi_{avg}} \frac{\left(\xi_{avg}\right)^k}{k!}. \tag{6}$$

Herein, we use a two-state Markov discrete-time process to model the state of the primary channel, as depicted in Figure 3. We assume that the state of the primary channel does not change during the time slot duration, T_{tot}, and the primary channel can switch states between two adjacent time slots. The state transition probabilities between two time slots are denoted as $P_{ij}\,|i, j \in \{F, B\}$, in which F stands for the *free* state, and B stands for the *busy* state. In this paper, we consider cooperative spectrum sensing, in which all CUs collaboratively detect spectrum holes based on an energy detection method, and they send these local sensing results to the CBS. Subsequently, the final decision on the primary users' activities is attained by combining the local sensing data at the CBS [36]. The performance of the cooperative sensing scheme can be evaluated based on probability of detection P_d and probability of false alarm P_f. P_d is denoted as the probability that the PU's presence is correctly detected (i.e., the primary channel is actually used by the PUs). Meanwhile, P_f is denoted as the probability that the PU's is detected to be active, but it is actually inactive (i.e., the sensing result of the primary channel is busy, but the primary channel is actually free) .

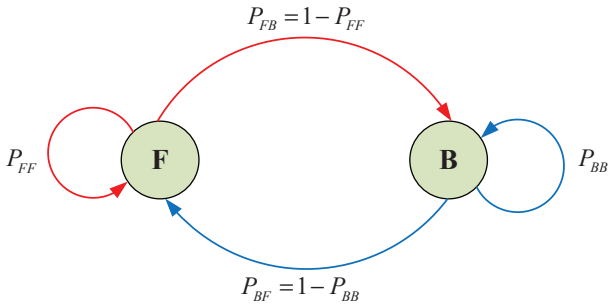

Figure 3. Markov chain model of the primary channel.

3. Long-Term Transmission Rate Maximization Problem Formulation

In this section, we aim at maximizing the long-term data transmission rate for uplink NOMA/OMA. The $2M$ users in the CRN can be decoupled into pairs according to their channel

gain, as described in Algorithm 1. After user pairing, the joint power allocation and bandwidth allocation problem can be formulated as follows:

$$\begin{aligned} \boldsymbol{a}^{*}(t) = \arg\max_{\boldsymbol{a}(t)} & \sum_{k=t}^{\infty} \sum_{m=1}^{M} R_m(k) \\ s.t. & 0 \le e_{im}^{tr} \le e_{max}^{tr} \\ & \sum_{m=1}^{M} \boldsymbol{b}_m(t) = W \end{aligned}, \tag{7}$$

where $\boldsymbol{a}(t) = \{\boldsymbol{b}(t), \boldsymbol{\varepsilon}(t)\}$ represents the action that the CBS assigns to the CUs in time slot t; $\boldsymbol{b}(t)$ indicates a vector of the allocated bandwidth portions assigned to the corresponding sub-channel, where $\boldsymbol{b}(t) = \{b_1(t), b_2(t), ..., b_M(t)\} \Big| \sum_{m=1}^{M} b_m(t) = W$ is the assigned bandwidth amount for m^{th} sub-channel; $\boldsymbol{\varepsilon}(t) = \left[e_{11}^{tr}(t), e_{21}^{tr}(t), , e_{12}^{tr}(t), e_{22}^{tr}(t), ..., e_{1\mathrm{M}}^{tr}(t), e_{2\mathrm{M}}^{tr}(t)\right]$ refers to a transmission energy vector of the CUs, where $e_{im}^{tr}(t) \in \left\{0, e_1^{tr}, e_2^{tr}, ..., e_{\max}^{tr}\right\}$ is the transmission energy value for CU_{im}, and e_{max}^{tr} represents the upper-bounded value of transmission energy for each CU in time slot t.

4. Deep Reinforcement Learning-Based Resource Allocation Policy

In this section, we first reformulate the joint power and bandwidth allocation problem, which is aimed at maximizing the long-term data transmission rate of the system as the framework of an MDP. Then, we apply a DRL approach to solve the problem, in which the agent (i.e., the CBS) learns to create the optimal resource policy via trial-and-error interactions with the environment. One of the disadvantages of reinforcement learning is that the high computational costs can be imposed due to the long time learning process of a system with high state space and action space. However, the proposed scheme requires less computation overhead by adopting deep neural networks, as compared to other algorithms such as value iteration-based dynamic programming in partially observable Markov decision process (POMDP) framework [20] in which the transition probability of the energy arrival is required for obtaining the solution. Thus, the complex in formulation and computation can be relieved regardless of the dynamic properties of the environment by using the proposed scheme, as compared to POMDP scheme.

Furthermore, the advantage of a deep reinforcement learning scheme as compared with the POMDP scheme is that the unknown harvested energy distribution can be estimated to create the optimal policy by interacting with the environment over the time horizon. In addition, the proposed scheme can work effectively in a large-state-and-space system by adopting deep neural networks. However, other reinforcement learning schemes such as Q-learning [38], actor-critic learning [39] might not be appropriate to large-state-and-space systems. In the proposed scheme, a deep neural network was trained to obtain the optimal policy where the reward of the system converges to optimal value. Then, the system can choose an optimal action at every state according to that policy learned from the training phase without re-training. Thus, deep actor-critic reinforcement learning can be more applicable to the wireless communication system.

4.1. Markov Decision Process

Generally, the purpose of reinforcement learning is for the agent to learn how to map each system state to an optimal action through a trial-and-error learning process. In this way, the accumulated sum of rewards can be maximized after a number of training time slots. Figure 4 illustrates the traditional reinforcement learning via agent–environment interaction. In particular, the agent observers the system state and then chooses an action at the beginning of a time slot. After that, the system receives the corresponding reward at the end of the time slot, and transfers to the next state based on the performed action. The system will be updated and will then go into the next interaction between agent and environment.

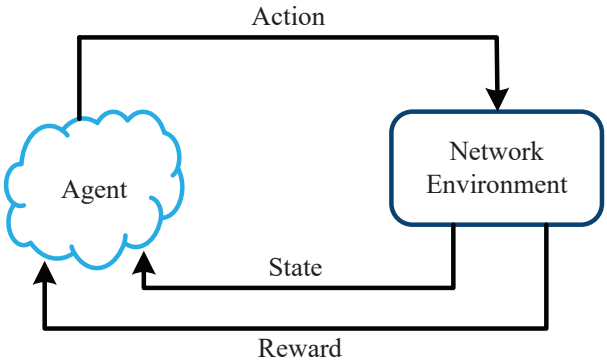

Figure 4. The agent–environment interaction process.

We denote the state space and action space of the system in this paper as $\mathbb{S}$ and $\mathbb{A}$, respectively; $s(t) = \{\mu(t), \boldsymbol{e}^{re}(t)\} \in \mathbb{S}$ represents the state of the network in time slot t, where $\mu(t)$ is the probability (*belief*) that the primary channel is free in that time slot, and $\boldsymbol{e}^{re}(t) = \left[e_{11}^{re}(t), e_{21}^{re}(t), e_{12}^{re}(t), e_{22}^{re}(t), e_{1M}^{re}(t), e_{2M}^{re}(t)\right]$ denotes a vector of remaining energy of CUs, where $0 \leq e_{im}^{re} \leq E_{bat}$ represents the remaining energy value of CU_{im}. The action at the CBS is denoted as $\boldsymbol{a}(t) = \{\boldsymbol{b}(t), \boldsymbol{\varepsilon}(t)\} \in \mathbb{A}$. In this paper, we define the reward as the sum data rate of the system, as presented in Equation (5).

The decision-making process can be expressed as follows. At the beginning of time slot t, the agent observes the state, $s(t) \in \mathbb{S}$, from information about the environment, and then chooses action $\boldsymbol{a}(t) \in \mathbb{A}$ following a stochastic policy, $\pi(a|s) = \Pr(a(t) = a|s(t) = s)$, which is mapped from the environment state to the probability of taking an action. In this work, the network agent (i.e, the CBS) determines the transmission power for each CU and decides whether to allocate the bandwidth portion to the NOMA groups in each time slot. Then, the CUs perform their operations (transmit data or stay silent) according to the assigned action from the CBS. Afterward, the instant reward, $R(t)$, which is defined in Equation (5), is fed back to the agent, and the environment transforms to the next state, $s(t+1)$. At the end of the time slot, the CUs report information about the current remaining energy level in each CU to the CBS for network management. In the following, we describe the way to update information about the belief and the remaining energy based on the assigned actions at the CBS.

4.1.1. Silent Mode

The global sensing decision shows that the primary channel is busy in the current time slot, and thus, the CBS trusts this result and has all CUs stay silent. As a consequence, there is no reward in this time slot, i.e., $R(t) = 0$. The belief in current time slot t can be calculated according to Bayes' rule [40], as follows:

$$\mu(t) = \frac{\mu(t) P_f}{\mu(t) P_f + (1 - \mu(t)) P_d}. \tag{8}$$

Belief $\mu(t+1)$ for the next time slot is updated as follows:

$$\mu(t+1) = \mu(t) P_{FF} + (1 - \mu(t)) P_{BF}. \tag{9}$$

The remaining energy of CU_{im} for the next time slot is updated as

$$e_{im}^{re}(t+1) = \min\left(e_{im}^{re}(t) + e_{im}^{h}(t) - e_{ss}, E_{bat}\right), \tag{10}$$

where e_{ss} is the consumed energy from the spectrum sensing process.

4.1.2. Transmission Mode

The global sensing decision indicates that the primary channel is free in the current time slot, and then, the CBS assigns transmission power levels to the CUs for transmitting their data to the CBS. We assume that the data of the CUs will be successfully decoded if the primary channel is actually free; otherwise, no data can be retrieved due to collisions between the signals of the PUs and CUs. In this case, there are two possible observations, as follows.

Observation 1 (Φ_1)*:* All data are successfully received and decoded at the CBS at the end of the time slot. This result means the primary channel was actually free during this time slot, and the global sensing result was correct. The total reward for the network is calculated as

$$R\left(s\left(t\right)|\Phi_1\right) = \sum_{m=1}^{M} R_m\left(t\right), \tag{11}$$

where the immediate data transmission rate of NOMA group G_m, $R_m\left(t\right)$, can be computed with Equation (4). Belief $\mu\left(t+1\right)$ for the next time slot is updated as

$$\mu\left(t+1\right) = P_{FF}. \tag{12}$$

The remaining energy in CU_{im} for the next time slot will be

$$e_{im}^{re}\left(t+1\right) = \min\left(e_{im}^{re}\left(t\right) + e_{im}^{h}\left(t\right) - e_{ss} - e_{im}^{tr}\left(t\right), E_{bat}\right), \tag{13}$$

where $e_{im}^{tr}\left(t\right)$ denotes the transmission energy assigned to CU_{im} in time slot t.

Observation 2 (Φ_2)*:* The CBS can not successfully decode the data from the CUs at the end of time slot t due to collisions between the signals of the CUs and the PUs. It implies that the primary channel was occupied, and misdetection happened. In this case, no reward is achieved, i.e., $R\left(s\left(t\right)|\Phi_2\right) = 0$. Belief $\mu\left(t+1\right)$ for the next time slot can be updated as

$$\mu\left(t+1\right) = P_{BF}. \tag{14}$$

The remaining energy in CU_{im} for the next time slot is updated by

$$e_{im}^{re}\left(t+1\right) = \min\left(e_{im}^{re}\left(t\right) + e_{im}^{h}\left(t\right) - e_{ss} - e_{im}^{tr}\left(t\right), E_{bat}\right). \tag{15}$$

In reinforcement learning, the agent is capable of improving the policy based on the recursive lookup table of state-value functions. The state-value function, $V^{\pi}\left(s\right)$, is defined as the maximum expected value of the accumulated reward starting from current state s with the given policy, which is written as [28]:

$$V^{\pi}\left(s\right) = E\left\{\sum_{t=1}^{\infty} \gamma^{t} R\left(t\right)|s\left(t\right) = s, \pi\right\}, \tag{16}$$

where $E\left\{.\right\}$ denotes the expectation, in which $\gamma \in \left(0,1\right)$ is the discount factor, which can affect the agent's decisions on myopic or foresighted operations; π is the stochastic policy, which maps environment state space $\mathbb{S}$ to action space $\mathbb{A}$, $\pi\left(a|s\right) = \Pr\left(\boldsymbol{a}\left(t\right) = a|s\left(t\right) = s\right)$. The objective of the resource allocation problem is to find optimal policy π^{*} that provides the maximum discounted value function in the long run, which can satisfy the Bellman equation as follows [41]:

$$\pi^{*}\left(a|s\right) = \arg\max_{\pi} V^{\pi}\left(s\right). \tag{17}$$

The policy can be explored by using an $\epsilon - greedy$ policy in which a random action is chosen with probability ϵ, or an action can be selected based on the current policy with probability $(1-\epsilon)$ during the training process [42]. As a result, the problem of joint power and bandwidth allocation in Equation (7) can be rewritten as Equation (17), and the solution to deep actor-critic reinforcement learning will be presented in the following section.

4.2. Deep Actor-Critic Reinforcement Learning Algorithm

The maximization problem in Equation (17) can be solved by using the actor-critic method, which is derived by combining the value-based method [43] and the policy-based method [44]. The actor-critic structure involves two neural networks (actor and critic) and an environment, as shown in Figure 5. The actor can determine the action according to the policy, and the critic evaluates the selected actions based on value functions and instant rewards that are fed back from the environment. The input of the actor is the state of the network, and the output is the policy, which directly affect how the agent chooses the optimal action. The output of the critic is a state-value function $V^{\pi}(s)$, which is used to calculate the temporal difference (TD) error. Thereafter, the TD error is used to update the actor and the critic.

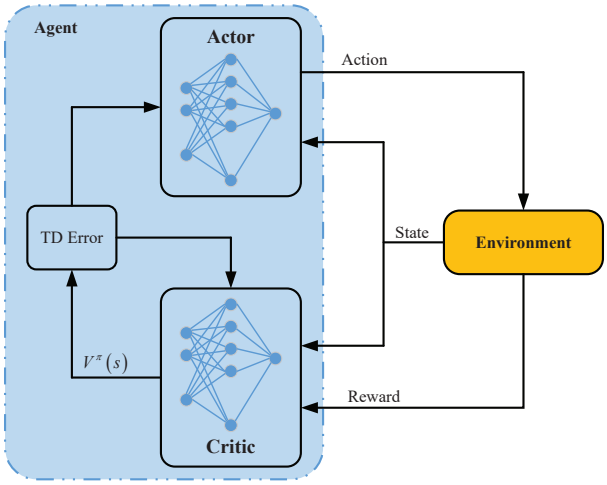

Figure 5. The structure of deep actor-critic reinforcement learning.

Herein, both the policy function in the actor and the value function in the critic are approximated with parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$, respectively, by two sequential models of a deep neural network. Both value function parameter $\boldsymbol{\omega}$ and policy parameter $\boldsymbol{\theta}$ are stochastically initialized and updated constantly by the critic and the actor, respectively, during the training process.

4.2.1. The Critic with a DNN

Figure 6 depicts the DNN at the critic, which is composed of an input layer, two hidden layers, and an output layer. The critic network is a feed-forward neural network that evaluates the action taken by the actor. Then, the evaluation of the critic is used by the actor to update its control policy. The input layer of the critic is an environment state, which contains $(2M+1)$ elements. Each hidden layer is a fully connected layer, which involves H_C neurons and uses a rectified linear unit (ReLU) activation function [45,46] as follows:

$$f_{ReLU}(z) = \max(0, z), \tag{18}$$

where $z = \sum_{i=1}^{2M+1} \boldsymbol{\omega}_i s_i(t)$ is the estimated output of the layer before applying the activation function, in which $s_i(t)$ indicates the ith element of the input state, $s(t)$, and $\boldsymbol{\omega}_i$ is the weight for the ith input.

The output layer of the DNN at the critic contains one neuron and uses the linear activation function to estimate the state-value function, $V^{\pi}(s)$. In this paper, the value function parameter is optimized by adopting stochastic gradient descent with a back-propagation algorithm to minimize the loss function, defined as the mean squared error, which is computed by

$$\mathcal{L}_{\boldsymbol{\omega}} = \delta^2(t), \tag{19}$$

where $\delta(t)$ is the TD error between the target value and the estimated value, which is given by

$$\delta(t) = E[R(t) + \gamma V_{\boldsymbol{\omega}}(s(t+1)) - V_{\boldsymbol{\omega}}(s(t))], \tag{20}$$

and it is utilized to evaluate selected action $\boldsymbol{a}(t)$ of the actor. If the value of $\delta(t)$ is positive, the tendency to choose action $\boldsymbol{a}(t)$ in the future, when the system is in the same state, will be strengthened; otherwise, it will be weakened. The critic parameter can be updated in the direction of the gradient, as follows:

$$\Delta\boldsymbol{\omega} = \alpha_c \delta(t) \nabla_{\boldsymbol{\omega}} V_{\boldsymbol{\omega}}^{\pi}(s(t)), \tag{21}$$

where α_c is the learning rate of the critic.

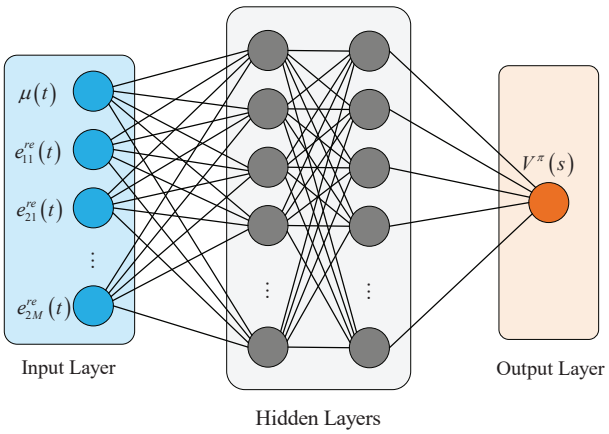

Figure 6. The deep neural network in the critic.

4.2.2. The Actor with a DNN

The DNN in the actor is shown in Figure 7, which includes an input layer, two hidden layers, and an output layer. The input layer of the actor is the current state of the environment. There are two hidden layers in the actor, where each hidden layer is comprised of H_A neurons. The output layer of the actor provides the probabilities of selecting actions in a given state. Furthermore, the output layer utilizes the soft-max activation function [28] to compute the policy of each action in the action space, which is given as:

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{e^{z_a}}{\sum_{\boldsymbol{a}' \in \mathbb{A}} e^{z_{\boldsymbol{a}'}}}, \tag{22}$$

where z_a is the estimated value for the preference of choosing action $\boldsymbol{a}$. In the actor, the policy can be enhanced by optimizing the state-value function as follows:

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) &= E[V^{\pi}(s)] \\ &= \sum_{s \in \mathbb{S}} d^{\pi}(s) V^{\pi}(s), \end{aligned} \tag{23}$$

where $d^{\pi}(s)$ is the state distribution. Policy parameter $\boldsymbol{\theta}$ can be updated toward the gradient ascending to maximize the objective function [39], as follows:

$$\Delta\boldsymbol{\theta} = \alpha_a \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}), \tag{24}$$

where α_a denotes the learning rate of actor network, and policy gradient $\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}})$ can be computed by using the TD error [47]:

$$\nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = E\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(s, \boldsymbol{a})\, \delta(t)\right]. \tag{25}$$

It is worth noting that TD error $\delta(t)$ is supplied by the critic. The training procedure of the proposed DACRL approach is summarized in Algorithm 2. In the algorithm, the agent interacts with the environment and learns to select optimal action in each state. The convergence of the proposed algorithm depends on number of steps per episode, the number of training episodes and the learning rate, which is discussed in the following section.

Algorithm 2 The training procedure of the deep actor-critic reinforcement learning algorithm

Input: $\mathbb{S}$, $\mathbb{A}$, γ, α_a, α_c, $\boldsymbol{e^{re}}(t)$, $\mu(t)$, E_{ca}, ξ_{avg}, T, W, ϵ_{min}, ϵ_{max}, ϵ_d.
Initialize network parameters of the actor and the critic: $\boldsymbol{\theta}, \boldsymbol{\omega}$.
Initialize $\epsilon = \epsilon_{max}$.
for each episode $e = 1, 2, 3, ..., L$:
    Sample an initial state $s \in \mathbb{S}$.
    for each time step $t = 0, 1, 2, 3..., T-1$:
        Observe current state $s(t)$, and estimate state value $V_{\omega}^{\pi}(s(t))$.
        Choose an action:

$$\boldsymbol{a}(t) = \begin{cases} \arg\max \pi_{\boldsymbol{\theta}}(a(t)|s(t)) & \text{w.p } 1 - \text{ffl} \\ \text{random action } a(t) \in \mathbb{A} & \text{otherwise} \end{cases}$$

        Execute action $\boldsymbol{a}(t)$, observe current reward $R(t)$.
        Update epsilon rate $\epsilon = \max(\epsilon.\epsilon_d, \epsilon_{\min})$
        Update next state $s(t+1)$
        Critic Process:
        Estimate next state value $V_{\omega}^{\pi}(s(t+1))$.
        Critic calculates TD error $\delta(t)$
        if episode is end at time slot t:
            $\delta(t) = R(t) - V_{\omega}(s(t))$.
        else
            $\delta(t) = R(t) + \gamma V_{\omega}(s(t+1)) - V_{\omega}(s(t))$.
        end if
        Update parameter of critic network $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} + \Delta\boldsymbol{\omega}$
        Actor Process:
        Update parameter of actor network $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$
    end for
end for
Output: Final policy $\pi_t^*(\boldsymbol{a}(t)|s(t))$.

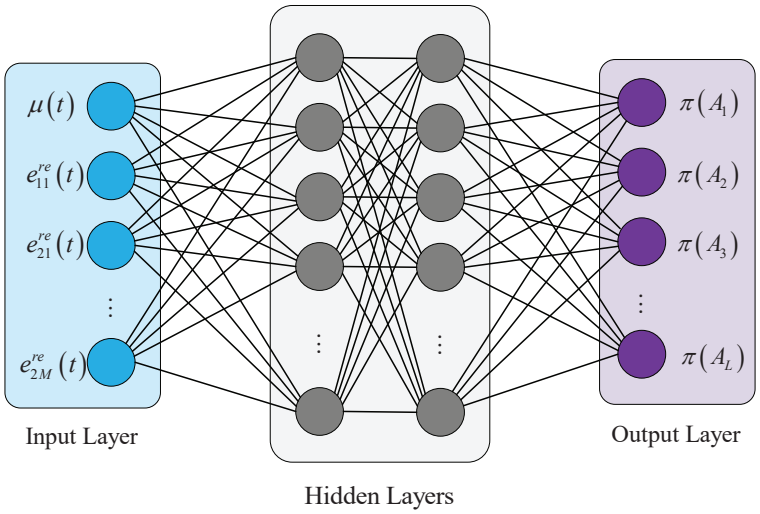

Figure 7. The deep neural network in the actor.

5. Simulation Results

In this section, we investigate the performance of uplink NOMA systems using our proposed scheme. The simulation results are compared with other myopic schemes [48] (Myopic-UP, Myopic-Random, and Myopic-OMA) in terms of average data transmission rate and energy efficiency. In the myopic schemes, the system only maximizes the reward in the current time slot, and the system bandwidth is allocated to the group only if it has at least one active CU in the current time slot. In particular, with the Myopic-UP scheme, the CBS arranges the CUs into different pairs based on Algorthim 1. In the Myopic-Random scheme, the CBS randomly decouples the CUs into pairs. In the Myopic-OMA scheme, the total system bandwidth is divided equally into sub-channels in order to assign them to each active CU without applying user pairing. In the following, we analyze the influence of the network parameters on the schemes through the numerical results.

In this paper, we used Python 3.7 with the TensorFlow deep learning library to implement the DACRL algorithm. Herein, we consider a network based on different channel gain values between the CUs and the CBS, such as $h_1 = -20$ dB, $h_2 = -25$ dB, $h_3 = -30$ dB, $h_4 = -35$ dB, $h_5 = -40$ dB, $h_6 = -45$ dB, where $h_1, h_2, h_3, h_4, h_5, h_6$ are the channel gains between $CU_1, CU_2, CU_3, CU_4, CU_5, CU_6$ and the CBS, respectively. Two sequential DNNs are utilized to model the value function and the policy function in the proposed algorithm. Each DNN is designed with an input layer, two hidden layers and an output layer as described in Section 4. The number of neurons in each hidden layer of the value function DNN in the critic, and the policy function in the actor, are set at $H_C = 24$ and $H_A = 24$, respectively. For the training process, value function parameter $\boldsymbol{\omega}$ and the policy parameter $\boldsymbol{\theta}$ are stochastically initialized by using uniform Xavier initialization [49]. The other simulation parameters for the system are shown in Table 1.

We first examine the average transmission rates of the the DACRL scheme under different training iterations, T, while the number of episodes, L, increases from 1 to 400. We achieved the results by calculating the average transmission rate after separately running the simulation 20 times, as shown in Figure 8. The curves sharply increase in the first 50 training episodes, and then gradually converge to the optimal value. We can see that the agent needs more than 350 episodes to learn the optimal policy at $T = 1000$ iterations per episode. However, with the increment in T, the algorithm begins to converge faster. For instance, the proposed scheme learns the optimal policy in less than 300 episodes when $T = 2000$. Nevertheless, it might take a very long time for the training process if each episode uses too many iterations, and the algorithm evenly converges to a locally optimal policy. As a result, the number of training iterations per episode and the number of training episodes should not be too large or too small. In the rest of the simulations, we set training episodes at 300 and training iterations at 2000.

Table 1. Simulation Parameters.

Parameter	Description	Value
M	Number of groups	3
T_{tot}	Time slot duration	200 ms
τ_{ss}	Sensing duration	2 ms
W	Total system bandwidth	1 Hz
E_{bat}	Battery capacity	30 µJ
e_{ss}	Sensing cost	1 µJ
e^{tr}	Transmission energy	0, 10, 20 µJ
ξ_{avg}	Mean value of harvested energy	5 µJ
μ	Initial belief that the primary channel is free	0.5
P_{FF}	Transition probability of the primary channel from state F to itself	0.8
P_{BF}	Transition probability of the primary channel from state B to state F	0.2
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
σ^2	Noise variance	−80 dB
γ	Discount factor	0.9
α_a	Learning rate of the actor	0.001
α_c	Learning rate of the critic	0.005
ϵ	Epsilon rate	$1 \rightarrow 0.01$
ϵ_d	Epsilon decay	0.9999
L	Number of episodes	300
T	Number of iterations per episode	2000

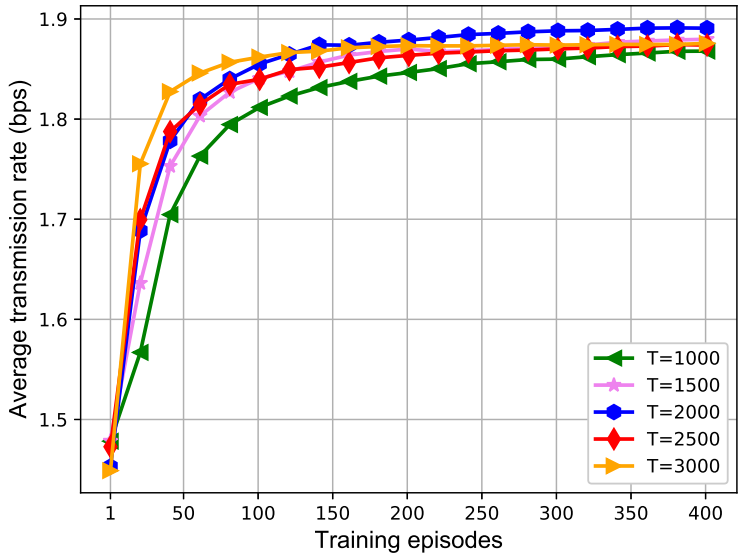

Figure 8. The convergence rate of the proposed actor-critic deep reinforcement learning with different training steps in each episode.

Figure 9 shows the convergence rate of the proposed scheme according to various values of actor learning rate α_a and critic learning rate α_c. The figure shows that the reward converges faster with increments in the learning rates. In addition, we can observe that the proposed scheme with actor learning rate $\alpha_a = 0.001$ and critic learning rate $\alpha_c = 0.005$ provides the best performance after 300 episodes. When the learning rates of the actor and the critic increase to $\alpha_a = 0.01$ and $\alpha_c = 0.005$, respectively, the algorithm converges very fast, but does not bring a good reward due to underfitting. Therefore, we set the actor and critic learning rates at $\alpha_a = 0.001$ and $\alpha_c = 0.005$, respectively, for the rest of the simulations.

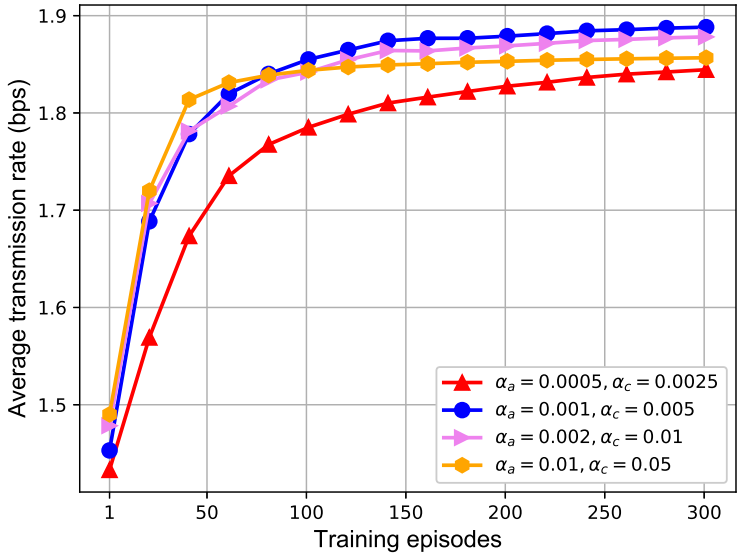

Figure 9. The convergence rate of the proposed actor-critic deep reinforcement learning according to different learning rate values.

Figure 10 illustrates the average transmission rates under the influence of mean harvested energy. We can see that the average transmission rate of the system increases when the mean value of harvested energy grows. The reason is that with an increase in ξ_{avg}, the CUs can harvest more solar energy, and thus, the CUs have a greater chance to transmit data to the CBS. In addition, the average transmission rate of the proposed scheme dominates the conventional schemes because the conventional schemes focus on maximizing the current reward, and they ignore the impact of the current decision on the future reward. Thus, whenever the primary channel is free, these conventional schemes allow all CUs to transmit their data by consuming most of the energy in the battery in order to maximize the instant reward. This makes the CUs stay silent in the future due to energy shortages. Although the Myopic-Random scheme had lower performance than the Myopic-UP scheme, it still had greater rewards than Myopic-OMA. This outcome demonstrates the efficiency of the hybrid NOMA/OMA approach, compared with the OMA approach, in terms of average transmission rate.

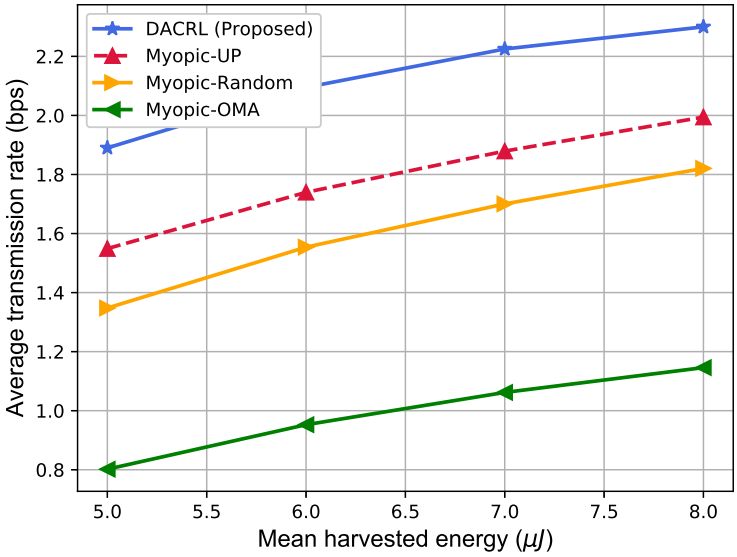

Figure 10. Average transmission rate according to different values for mean harvested energy.

In Figure 11, the energy efficiency of the schemes was compared with respect to the mean value of the harvested energy. In this paper, we define energy efficiency as the transmission data rate obtained at the CBS over the total energy consumption of the CUs during the operations. We can see that the energy efficiency declines as ξ_{avg} rises. The reason is that when the harvested energy goes up, the CUs can gather more energy for their operations; however, the amount of energy overflowing the CUs' batteries also increases. The curves show that the performance of the proposed scheme outperforms the other conventional schemes because the DACRL agent can learn about the dynamic arrival of harvested energy from the environment. Thus, the proposed scheme can make proper decision in each time slot.

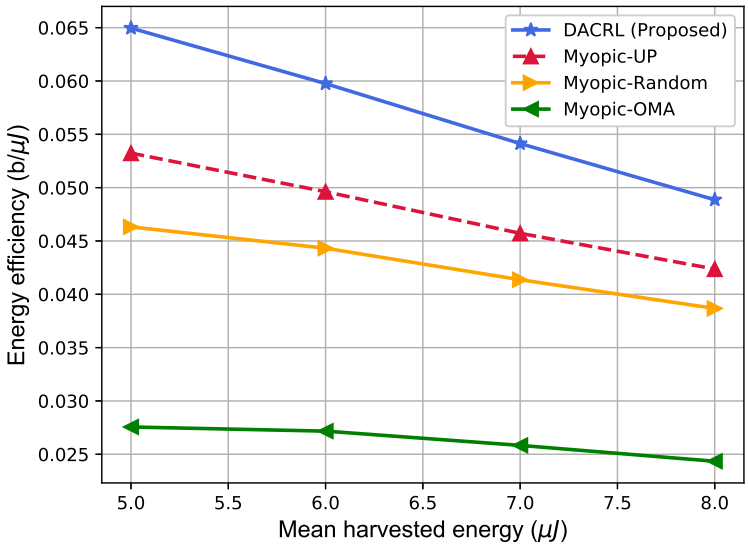

Figure 11. Energy efficiency according to different values of harvested mean energy.

In Figures 12 and 13, we plot the average transmission rate and the energy efficiency, respectively, based on differing noise variance at the CBS. The curves show that system performance notably degrades when noise variance increases. To explain this, noise variance will degrade the data transmission rate, as shown in Equation (3). As a consequence, energy efficiency also decreases with an increment in noise variance. Based on noise variance at the CBS, the figures verify that the proposed scheme dominates the myopic schemes.

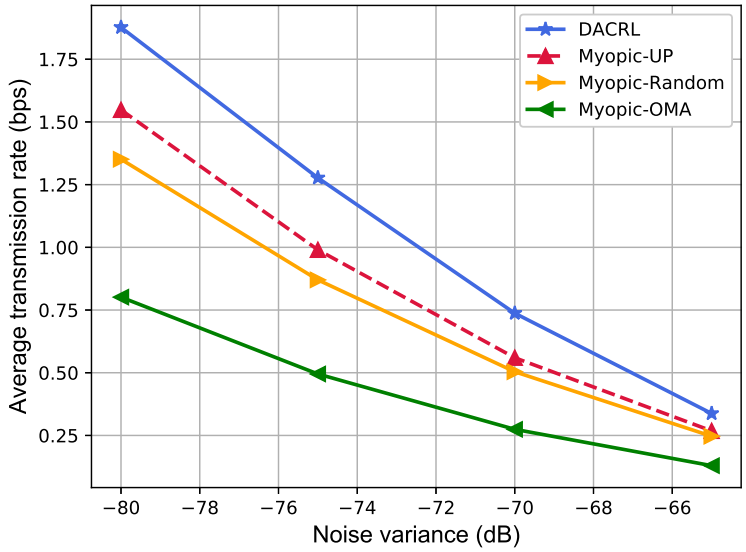

Figure 12. Average transmission rate according to noise variance.

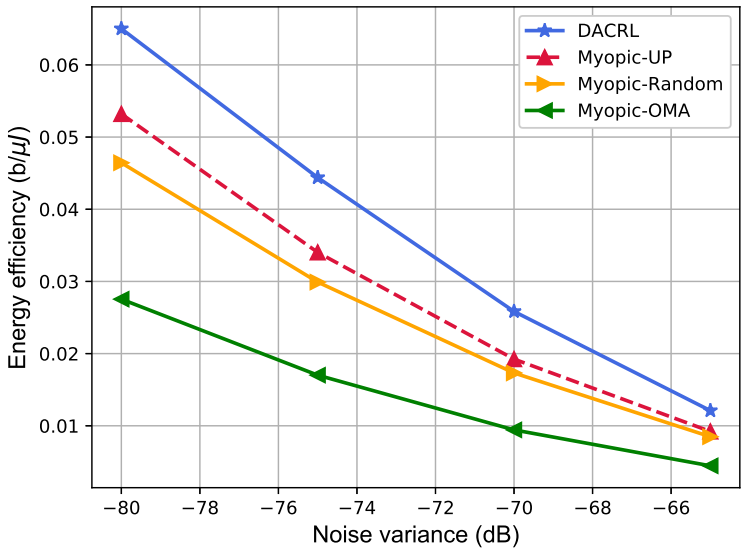

Figure 13. Energy efficiency according to noise variance.

6. Conclusions

In this paper, we investigated a deep reinforcement learning framework for joint power and bandwidth allocation by adopting both hybrid NOMA/OMA and user pairing in uplink CRNs. The DACRL algorithm was employed to maximize the long-term transmission rate under the energy constraint in the CUs. A DNN was applied to approximate the policy function and the value function such that the algorithm can work in the system with large state and action spaces. The agent of the DACRL can explore the optimal policy by interacting with the environment. As a consequence, the CBS can effectively allocate bandwidth and power to the CUs based on the current network state in each timeslot. The simulation results verified the advantages of the proposed scheme in improving network performance under various network conditions in the long run, compared to the conventional schemes.

Author Contributions: All authors conceived and proposed the research idea. H.T.H.G. made the formulation and performed the simulations under the supervision of T.N.K.H. and P.D.T.; I.K. analyzed the simulation results; H.T.H.G. wrote the draft paper; T.N.K.H., P.D.T. and I.K. reviewed and edited the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant through the Korean Government (MSIT) under Grant NRF-2018R1A2B6001714.

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

References

1. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chih-Lin, I.; Poor, H.V. Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun. Mag.* **2017**, *55*, 185–191. [CrossRef]
2. Dai, L.; Wang, B.; Yuan, Y.; Han, S.; Chih-Lin, I.; Wang, Z. Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.* **2015**, *53*, 74–81. [CrossRef]
3. Islam, S.M.R.; Zeng, M.; Dobre, O.A.; Kwak, K.-S. Resource allocation for downlink NOMA systems: Key techniques and open issues. *IEEE Wirel. Commun.* **2018**, *25*, 40–47. [CrossRef]
4. Yu, W.; Musavian, L.; Ni, Q. Link-layer capacity of NOMA under statistical delay QoS guarantees. *IEEE Trans. Commun.* **2018**, *66*, 4907–4922. [CrossRef]
5. Zeng, M.; Yadav, A.; Dobre, O.A.; Tsiropoulos, G.I.; Poor, H.V. On the sum rate of MIMO-NOMA and MIMO-OMA systems. *IEEE Wirel. Commun. Lett.* **2017**, *6*, 534–537. [CrossRef]
6. Di, B.; Song, L.; Li, Y. Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 7686–7698. [CrossRef]
7. Timotheou, S.; Krikidis, I. Fairness for non-orthogonal multiple access in 5G systems. *IEEE Signal Process. Lett.* **2015**, *22*, 1647–1651. [CrossRef]

8. Liang, W.; Ding, Z.; Li, Y.; Song, L. User pairing for downlink nonorthogonal multiple access networks using matching algorithm. *IEEE Trans. Commun.* **2017**, *65*, 5319–5332. [CrossRef]
9. Zhang, Y.; Wang, H.-M.; Zheng, T.-X.; Yang, Q. Energy-efficient transmission design in non-orthogonal multiple access. *IEEE Trans. Veh. Technol.* **2017**, *66*, 2852–2857. [CrossRef]
10. Hao, W.; Zeng, M.; Chu, Z.; Yang, S. Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access. *IEEE Wireless Commun. Lett.* **2017**, *6*, 782–785. [CrossRef]
11. Fang, F.; Zhang, H.; Cheng, J.; Leung, V.C.M. Energy-efficient resource allocation for downlink non-orthogonal multiple access network. *IEEE Trans. Commun.* **2016**, *64*, 3722–3732. [CrossRef]
12. Lv, T.; Ma, Y.; Zeng, J.; Mathiopoulos, P.T. Millimeter-wave NOMA transmission in cellular M2M communications for Internet of Things. *IEEE Internet Things J.* **2018**, *5*, 1989–2000. [CrossRef]
13. Cui, J.; Liu, Y.; Ding, Z.; Fan, P.; Nallanathan, A. Optimal user scheduling and power allocation for millimeter wave NOMA systems. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 1502–1517. [CrossRef]
14. Ahmad, W.S.H.M.W.; Radzi, N.A.M.; Samidi, F.S.; Ismail, A.; Abdullah, F.; Jamaludin, M.Z.; Zakaria, M.N. 5G Technology: Towards Dynamic Spectrum Sharing Using Cognitive Radio Networks. *IEEE Access* **2020**, *8*, 14460–14488. [CrossRef]
15. Amjad, M.; Musavian, L.; Rehmani, M.H. Effective Capacity in Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3007–3038. [CrossRef]
16. Zhou, F.; Beaulieu, N.C.; Li, Z.; Si, J.; Qi, P. Energy-efficient optimal power allocation for fading cognitive radio channels: Ergodic capacity, outage capacity, and minimum-rate capacity. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 2741–2755. [CrossRef]
17. Goldsmith, A.; Jafar, S.A.; Maric, I.; Srinivasa, S. Breaking spectrum gridlock with cognitive radios: An information theoretic perspective. *Proc. IEEE* 2009, *97*, 894–914. [CrossRef]
18. Lv, L.; Chen, J.; Li, Q.; Ding, Z. Design of cooperative nonorthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis. *IEEE Trans. Commun.* **2017**, *65*, 2641–2656. [CrossRef]
19. Lv, L.; Ni, Q.; Ding, Z.; Chen, J. Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over nakagamim fading channels. *IEEE Trans. Veh. Technol.* **2017**, *66*, 5506–5511. [CrossRef]
20. Thanh, P.D.; Hoan, T.N.K.; Koo, I. Joint Resource Allocation and Transmission Mode Selection Using a POMDP-Based Hybrid Half-Duplex/Full-Duplex Scheme for Secrecy Rate Maximization in Multi-Channel Cognitive Radio Networks. *IEEE Sens. J.* **2020**, *20*, 3930–3945. [CrossRef]
21. Lu, X.; Wang, P.; Niyato, D.; Kim, D.I.; Han, Z. Wireless networks with RF energy harvesting: A contemporary survey. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 757–789. [CrossRef]
22. Giang, H.T.H.; Hoan, T.N.K.; Thanh, P.D.; Koo, I. A POMDP-based long-term transmission rate maximization for cognitive radio networks with wireless-powered ambient backscatter. *Int. J. Commun. Syst.* **2019**, *32*, e3993. [CrossRef]
23. Zhang, N.; Wang, J.; Kang, G.; Liu, Y. Uplink nonorthogonal multiple access in 5g systems. *IEEE Commun. Lett.* **2016**, *20*, 458–461. [CrossRef]
24. Ni, Z.; Chen, Z.; Zhang, Q.; Zhou, C. Analysis of RF Energy Harvesting in Uplink-NOMA IoT-Based Network. In Proceedings of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019; pp. 1–5.
25. Gussen, C.M.G.; Belmega, E.V.; Debbah, M. Pricing and bandwidth allocation problems in wireless multi-tier networks. In Proceedings of the 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA, 6–9 November 2011; pp. 1633–1637.
26. Arunthavanathan, S.; Kandeepan, S.; Evans, R.J. A Markov Decision Process-Based Opportunistic Spectral Access. *IEEE Wirel. Commun. Lett.* **2016**, *5*, 544–547. [CrossRef]
27. Xiao, H.; Yang, K.; Wang, X.; Shao, H. A robust MDP approach to secure power control in cognitive radio networks. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, USA, 10–15 June 2012; pp. 4642–4647.
28. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; A Bradford Book; MIT Press: London, UK, 2018.
29. Li, R.; Zhao, Z.; Chen, X.; Palicot, J.; Zhang, H. TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 2000–2011. [CrossRef]

30. Puspita, R.H.; Shah, S.D.A.; Lee, G.; Roh, B.; Oh, J.; Kang, S. Reinforcement Learning Based 5G Enabled Cognitive Radio Networks. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 16–18 October 2019; pp. 555–558.
31. Meng, X.; Inaltekin, H.; Krongold, B. Deep Reinforcement Learning-Based Power Control in Full-Duplex Cognitive Radio Networks. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, UAE, 9–13 December 2018; pp. 1–7.
32. Ong, K.S.H.; Zhang, Y.; Niyato, D. Cognitive Radio Network Throughput Maximization with Deep Reinforcement Learning. In Proceedings of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019; pp. 1–5.
33. Zhang, H.; Yang, N.; Huangfu, W.; Long, K.; Leung, V.C.M. Power Control Based on Deep Reinforcement Learning for Spectrum Sharing. *IEEE Trans. Wirel. Commun.* **2020**. . [CrossRef]
34. Ding, Z.; Yang, Z.; Fan, P.; Poor, H.V. On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users. *IEEE Signal Process. Lett.* **2014**, *21*, 1501–1505. [CrossRef]
35. Han, W.; Li, J.; Li, Z.; Si, J.; Zhang, Y. Efficient Soft Decision Fusion Rule in Cooperative Spectrum Sensing. *IEEE Trans. Signal Process.* **2013**, *61*, 1931–1943. [CrossRef]
36. Ma, J.; Zhao, G.; Li, Y. Soft Combination and Detection for Cooperative Spectrum Sensing in Cognitive Radio Networks. *IEEE Trans. Wirel. Commun.* **2008**, *7*, 4502–4507.
37. Lee, P.; Eu, Z.A.; Han, M.; Tan, H. Empirical modeling of a solar-powered energy harvesting wireless sensor node for time-slotted operation. In Proceedings of the 2011 IEEE Wireless Communications and Networking Conference, Cancun, Mexico, 28–31 March 2011; pp. 179–184.
38. Kawamoto, Y.; Takagi, H.; Nishiyama, H.; Kato, N. Efficient Resource Allocation Utilizing Q-Learning in Multiple UA Communications. *IEEE Trans. Netw. Sci. Eng.* **2019**, *6*, 293–302. [CrossRef]
39. Wei, Y.; Yu, F.R.; Song, M.; Han, Z. User Scheduling and Resource Allocation in HetNets With Hybrid Energy Supply: An Actor-Critic Reinforcement Learning Approach. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 680–692. [CrossRef]
40. Stone, J.V. *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*; Sebtel Press: Sheffield, UK, 2013; p. 174.
41. Grondman, I.; Busoniu, L.; Lopes, G.A.D.; Babuska, R. A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. *IEEE Trans. Syst. Man Cybern. Part C* **2012**, *42*, 1291–1307. [CrossRef]
42. Wei, Y.; Yu, F.R.; Song, M.; Han, Z. Joint Optimization of Caching, Computing, and Radio Resources for Fog-Enabled IoT Using Natural Actor–Critic Deep Reinforcement Learning. *IEEE Internet Things J.* **2019**, *6*, 2061–2073. [CrossRef]
43. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double Q-Learning. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2094–2100.
44. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016; pp. 1–14, .
45. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
46. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
47. Zhong, C.; Lu, Z.; Gursoy, M.C.; Velipasalar, S. A Deep Actor-Critic Reinforcement Learning Framework for Dynamic Multichannel Access. *IEEE Trans. Cognit. Commun. Netw.* **2019**, *5*, 1125–1139. [CrossRef]
48. Wang, K.; Chen, L.; Liu, Q. On Optimality of Myopic Policy for Opportunistic Access With Nonidentical Channels and Imperfect Sensing. *IEEE Trans. Veh. Technol.* 2014, 63, 2478–2483. [CrossRef]
49. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *The Thirteenth International Conference on Artificial Intelligence and Statistics*, 2nd ed.; Athena Scientic: Belmont, MA, USA, 2001; Volume 1–2.