



Article Energy Efficient Power Allocation in Massive MIMO Based on Parameterized Deep DQN

Shruti Sharma * and Wonsik Yoon *

Department of Electrical and Computer Engineering, Ajou University, Suwon 16499, Republic of Korea * Correspondence: shruti@ajou.ac.kr (S.S.); wsyoon@ajou.ac.kr (W.Y.)

Abstract: Machine learning offers advanced tools for efficient management of radio resources in modern wireless networks. In this study, we leverage a multi-agent deep reinforcement learning (DRL) approach, specifically the Parameterized Deep Q-Network (DQN), to address the challenging problem of power allocation and user association in massive multiple-input multiple-output (M-MIMO) communication networks. Our approach tackles a multi-objective optimization problem aiming to maximize network utility while meeting stringent quality of service requirements in M-MIMO networks. To address the non-convex and nonlinear nature of this problem, we introduce a novel multi-agent DQN framework. This framework defines a large action space, state space, and reward functions, enabling us to learn a near-optimal policy. Simulation results demonstrate the superiority of our Parameterized Deep DQN (PD-DQN) approach when compared to traditional DQN and RL methods. Specifically, we show that our approach outperforms traditional DQN methods in terms of convergence speed and final performance. Additionally, our approach shows 72.2% and 108.5% improvement over DQN methods and the RL method, respectively, in handling large-scale multi-agent problems in M-MIMO networks.

Keywords: convergence; multi-agent; reinforcement learning; reward; user association

check for **updates**

Citation: Sharma, S.; Yoon, W. Energy Efficient Power Allocation in Massive MIMO Based on Parameterized Deep DQN. *Electronics* 2023, 12, 4517. https://doi.org/10.3390/ electronics12214517

Academic Editor: Andrea Asperti

Received: 29 September 2023 Revised: 30 October 2023 Accepted: 30 October 2023 Published: 2 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/).

1. Introduction

With the increasing demand for mobile communications and Internet of Things technologies, wireless networks are facing increased data traffic and resource management issues owing to the rapid growth of wireless applications. Fifth-generation cellular networks have gained considerable attention for achieving spectrum efficiency and storage capacity. Massive multiple-input multiple-output (M-MIMO) networks are reliable options to overcome data storage and capacity issues to satisfy diverse user requirements. The main concept in M-MIMO technology is to equip the base stations (BSs) with a large number (i.e., 100 or more) of wireless antennas to simultaneously serve numerous users, enabling significant improvement in spectrum efficiency [1,2].

The presence of a huge number of antennas in M-MIMO, data multiplexing, and management would make MIMO transceiver optimization more challenging compared to single-antenna networks. The multi-objective nature of M-MIMO transceivers has resulted in various optimization strategies being performed in the past, including user association [3], power allocation [4], and user scheduling [5]. A joint user association and resource allocation problem was investigated in [6–8]. Given the non-convex multiple objective function in the M-MIMO problem, achieving the Pareto optimal solution set in a multi-objective environment becomes more challenging. Recently, proposed methods to solve multi-objective problems include approaches based on linear programming [9], game-theory [10], and Markov approximation [11,12]. Success of these methods requires complete knowledge of the system, which is rarely available.

Thus, emerging machine learning (ML) is an efficient tool to solve such complex multiobjective problems. In this ML field, Reinforcement Learning (RL) is the most appropriate branch to solve a non-convex problem. In RL-based optimization methods, three major elements (i.e., agents, reward, and action) of the proposed solution enable the self-learning abilities from the environment.

The Q-learning algorithm is one of the widely used RL methods because it requires minimal computation. It can be expressed using single equations and does not need to know the state transition probability. The RL agents maximize the long-term rewards over the current optimal reward function [13,14] using a Q-learning algorithm [14–16]. The agents are free to change their actions independently in a single-agent RL method, leading to a fluctuation in the overall action space, as well as action and rewards of the different agents in the process [16]. Q-learning methods have been used for power and resource allocation in heterogeneous and cellular networks [17]. However, it may be considerably difficult to handle such large state and action spaces in M-MIMO systems using Q-learning methods. To handle these issues of RL methods, deep reinforcement learning (DRL) methods are coupled with deep learning and RL to enhance the performance of RL for large-scale scenario problems. Nowadays, DRL methods [18] are promising to handle these complicated objective functions. DRL methods have already been applied to several tasks, such as resource allocation, fog radio access networks, dynamic channels, access, and mobile computing [19–21].

In DRL, the deep Q-network (DQN) method is mostly employed to train the agents to achieve an optimal scheme from a large state and action space. In [22], Rahimi et al. gave an algorithm of DQN, which was based on deep neural networks and has been previously used in the literature. Zhao et al. used the DRL method for the efficient management of user association and resource allocation for maximizing the network utility and maintaining the quality of service (QoS) requirements [23]. In [24–27], the authors proposed a DQN algorithm to allocate power using a multi-agent DRL method. Recent advancements in DRL, particularly techniques like Deep Q-Networks (DQNs), have opened up new avenues for addressing resource allocation challenges in wireless networks. However, when it comes to applying DQNs to solve the combined problem of power allocation and user association, a critical step involves converting the continuous action space for power allocation into a discrete action space. This quantization process can potentially result in suboptimal power allocation decisions, limiting overall performance. Additionally, the complexity of a DQN grows significantly as the dimensions of the action space increase. This exponential complexity can lead to high power consumption and slow convergence rates, which are highly undesirable in practical applications. The motivation for our research stems from the evolving landscape of wireless communication networks, particularly resource allocation in massive MIMO systems. As the demand for high-speed, reliable, and energy-efficient wireless transmission increases, optimizing resource allocation becomes dominant. This challenge becomes even more noticeable when dealing with complex circumstances that involve both discrete and continuous decision-making. EE has appeared as a critical metric in modern wireless networks, as it directly effects sustainability and operational costs. Achieving high EE while ensuring seamless user association and power allocation is a complicated problem that necessitates advanced solutions. Our motivation, therefore, lies in addressing this pressing need for more efficient and sustainable wireless networks. We are driven by the potential to make a substantial impact on the way resources are managed, energy is conserved, and user satisfaction is enhanced. By introducing the PD-DQN algorithm and applying it to the joint user association and power allocation problem, we aim to contribute valuable insights and practical solutions to the field of wireless communication.

We are motivated by the vision of network operators and engineers with a versatile tool that can adapt to the workings of hybrid action spaces, ultimately leading to more high-performing wireless networks. To address these challenges, our paper introduces the use of Parameterized Deep Q-Network (PD-DQN) techniques, which deal with parameterized state spaces. However, it falls short in terms of estimation capabilities and tends to produce suboptimal policies due to its tendency to overestimate Q-values [28].

Our approach outperforms traditional DQN and RL methods with significant improvements in convergence speed and final performance. In [29], authors concentrate on hybrid beamforming in integrated satellite-terrestrial networks. Their research lies in achieving both secrecy and energy efficiency through carefully designed beamforming strategies. In this work [30], authors focus on the security aspects of resource allocation within cognitive satellite-terrestrial networks. They delve into the design of secure transmission techniques through strategic power and resource allocation. In [31], authors present a comprehensive examination of the issue of malicious reflector-assisted beamforming in IoT networks. Their primary focus is on devising power allocation strategies to counteract the impact of destructive beamforming. In [32], authors explore the advantages of incorporating reconfigurable intelligent surfaces (RISs) and their role in power allocation and system performance. PD-DQN is well suited for solving problems involving hybrid action spaces, making it a more efficient choice for the joint power allocation and user association problem, which uses discrete and continuous action space [28]. This hybrid approach is designed to address the challenges posed by a mixed discrete-continuous action space.

In this study, we introduced a novel approach with a PD-DQN algorithm [28]. The main contributions of this paper are as follows:

- Our work presents an advanced approach to the user association and power allocation problem, with the overarching objective of maximizing EE in the context of a massive MIMO network. By aiming for EE as the primary goal, we address a critical aspect of modern wireless communication networks, highlighting the practical relevance of our research.
- To effectively address power allocation challenges, we have carefully considered the design of the action space, the state space, and the reward function. Our approach leverages the model-free Deep Q-Network (DQN) framework and, notably, the PD-DQN framework, which empowers us to update policies within the hybrid discrete-continuous action space. This innovation is pivotal in enabling adaptive decision-making within complex real-world scenarios, underscoring the practicality of our methodology.
- The simulation results show that the proposed user association and power allocation method based on PD-DQN performs better than DRL and RL methods.

2. System Model

In this study, we considered a single-cell massive MIMO network, which consists of N remote radio heads (RRHs) and single antenna users. Here, RRHs are connected to a baseband unit via backhaul connections. Each RRH is equipped with M_{max} antennas. There are U single-antenna users served by N RRHs together operating in the same time frequency domain. It is assumed that $M_{max} > U$. In this network, we associate each user with a single RRH [7]. The set of users is denoted by U. Figure 1 shows the network architecture based on a DRL. These approaches utilize a credit assignment system to reward parameters based on their recent impact on the search process. They also employ an operator selection mechanism to determine which parameters to use, guided by the assigned credits.

It is assumed that the channel between the *u*th users and the *n*th RRH is given by

$$\mathbf{h}_{n,u} = \sqrt{\beta_{n,u} \, \mathbf{g}_{n,u}} \tag{1}$$

where $\beta_{n,u}$ signifies the large-scale fading coefficient, and $g_{n,u}$ signifies the small-scale fading coefficient. $g_{n,u}$ is also known as Rayleigh fading, and the elements are independent and identically distributed (i.i.d) random variables having zero mean and unit variance [8].

The received signal of the *u*th user on the *n*th RRH can be given by [7]

$$Y_{n,u} = \sqrt{p_u \mathbf{h}_{n,u} \mathbf{w}_{n,u}} s_{n,u} + \sum_{j=1, j \neq u}^{U} \sqrt{p_j \mathbf{h}_{n,u}} \mathbf{w}_{n,u} s_{n,j} + \mathbf{z}_{n,u}$$
(2)

where p_u is power transmitted through the *u*th user; $s_{n,u}$ is the $N_r \times 1$ data symbol of the *u*th user on the *n*th RRH; $\mathbf{w}_{n,u}$ is the $N_t \times N_r$ beamforming vector of the *u*th user on the *n*th RRH, which is given by $\mathbf{w}_{n,u} = \frac{\mathbf{h}_{n,u}}{\sqrt{\mathbb{E}\left\{\left\|\hat{\mathbf{h}}_{n,u}\right\|^2\right\}}}$; and $\mathbf{h}_{n,u}$ is the channel matrix of the

*u*th user on the *n*th RRH. $\mathbf{z}_{n,u}$ is the noise vector of independent and identically distributed (i.i.d.) additive complex Gaussian noise having zero mean and variance of σ . Without loss of generality, we set

$$E\left[s_{m,u^{s}m,u}^{H}\right] = \mathbf{I}, \ E\left[s_{m,u^{s}m,j}^{H}\right] = \mathbf{0}, \ (k \neq \mathbf{j}), \ E\left[s_{m,u^{n}m,u}^{H}\right] = \mathbf{0}$$



Figure 1. System model based on parameterized DRL.

Power Consumption

We considered the downlink phase for power consumption. The overall P_c is expressed as the sum of the transmit power and fixed power consumption of the RRHs and base units denoted as P_{FIX} and the power consumed by the components of the active antennas [5]. The total P_c can then be given by

$$P_c = P_{Fix} + P_a \sum_{n=1}^{N} M_n + \sum_{n=1}^{N} \sum_{u=1}^{U} \frac{1}{\nu} P_{n,u}$$
(3)

where P_a denotes power assumed for active antenna, and ν denotes power amplifier efficiency, $\nu \in (0, 1)$.

3. Problem Formulation

According to the system model, the ergodic achievable rate of the *u*th user is given by [10]

$$R_{n,u} = \log_2(1 + \Gamma_{u,n}) \tag{4}$$

where $\Gamma_{u,n}$ is signal-to-interference-plus-noise ratio of the uth and nth RRH and is given by

$$R_{n,u} = \frac{M_n p_n \Gamma_{n,u}}{\sum_{j=1}^N \sum_{q=1}^U p_{j,q} \beta_{j,u} + \sigma^2}$$
(5)

To deal with the above problem, the association between the *u*th users and nth RRH is given by

$$\mathbf{x}_{n,u} = \begin{cases} 1, u \in L_l \\ 0, u \in \frac{U}{L_l} l = 1, \dots L \end{cases}$$
(6)

The system energy efficiency (EE) can be expressed as

$$\max_{X,P} \eta = \frac{\sum_{u=1}^{U} R_u}{P_{FIX} + p_a \sum_{n=1}^{N} M_n + \sum_{n=1}^{N} \sum_{u=1}^{U} \frac{1}{v} p_{n,u}}$$
(7)

The optimization problem maximizing the system EE can be formulated as

...

P1:
$$\max_{X,P} \eta = \frac{\sum_{u=1}^{U} R_u(X,P)}{P_c(X,P)}$$
, s.t C1: $0 < P_{u,m} \le P_u^{\max}$
C2: $a_{u,m} \in \{0,1\}$
C3: $\sum_m a_{u,m} = 1$
C4: $R_u \ge R_{\min}$ (8)

In the above problem, C1 indicates that the transmitted power consumption is smaller than the transmit power limit of each RRH. Constraints C2 and C3 indicate that one user can only be associated with one RRH. C4 maintains the QoS requirement of each user and signifies the lower limit of the required transmit rate of users. Problem (8) is NP-hard and is usually difficult to find a feasible solution for [22]. Therefore, a multi-agent DRL approach was used to solve this problem, as described in the next section.

4. Multi-Agent DRL Optimization Scheme

The problem P1 is a non-convex problem where user association as well as power allocation approaches are involved. To solve this tractable problem, a multi-agent DQN-based RL technique was applied. The major component of the RL approach is based on the Markov decision-making process (MDP), which is a new proposed reward function, prior to the application of the multi-agent DQN approach.

4.1. Overview of RL Method

In this section, we present the overview of RL. In RL, the aim is to find optimal policy. The problem P1 is converted into a MDP (s, a, r, $P_{ss^{new}}$) similar to the existing work [26,27], where s, a, and r represent the set of state, set of action, and reward functions, respectively. $P_{ss^{new}}$ is the transition probability from state s to s^{new} with reward r. In the DRL, these state variables are defined as follows:

State space: In problem P1, the users as agents select the BSs for communication at time *t*. The network consists of *U* agents. The state space can be expressed as

$$s(t) = \{s_1(t), s_2(t), \dots, s_u(t), \dots, s_U(t)\}$$
(9)

Action space: At time t, the action of the agent is to control the transmit power level between the user association and BS. The action space consisting of each user can be defined as

$$a(t) = \{a_1(t), a_2(t), ..., a_u(t), ..., a_u(t)\}$$
(10)

Reward function: The energy efficiency of all users can be expressed as a system reward function

$$r(t) = \sum_{u}^{U} r_{u}(t) = \sum_{u=1}^{U} \eta_{u}(t)$$
(11)

where is r(t) the reward function, which is maximized to achieve the optimal policy with interaction with the outer environment.

Therefore, within the RL framework, the problem P1 can be transformed into problem P2, as follows:

$$P1: \max_{X,P,M} r_n \tag{12}$$

where X represents the user association matrix, and P denotes the power allocation vector. The agent identifies its state s(t) at time t and follows a policy π to perform an action a(t), that is, $a(t) = \pi(s(t))$. Following this, the users communicate with the BSs, and the reward function becomes r(t) = r(t|s = s(t), a = a(t)). Therefore, the future cumulative discounted reward at time t can be given by

$$R(t) = \sum_{\tau=t}^{T} \gamma^{\tau-t} r(\tau)$$
(13)

where $\gamma \in [0, 1]$ denotes the discount factor for the upcoming rewards. To solve the P2, a value function for policy π_u is defined as

$$V_u^{\pi_u} = E\left[\sum_{\tau=t}^T \gamma^{\tau-t} r_u(\tau) | s_u^{\tau}, a_u(\tau)\right]$$
(14)

where $E[\cdot]$ denotes the expectation operator. Using the Markov property, the value function is defined as

$$V_u^{\pi_u} = r(s_u^{\tau}, \pi_u) + \gamma \sum_{s^{new} \in S} P_{ss^{new}}(\pi_u) V_u^{\pi_u}(s_u^{\tau^{new}})$$
(15)

The Q-function when performing action $a_u(\tau)$ in state s_u^{τ} with policy π_u can be expressed as in [24], that is,

$$Q_{\pi}(s^{\tau}, a(\tau)) = E[R(\tau)|s^{\tau}, a(\tau)]$$
(16)

The optimal Q-value function satisfies the Bellman equation [33,34] derived as

$$Q_{\pi^*}(s^{\tau}, a(\tau)) = r(s^{\tau}, a(\tau)) + \gamma \sum_{s' \in S} P_{ss'}a(\tau) V^{\pi^*}(s^{\tau'})$$
(17)

Accordingly, the Bellman optimality Equation (17) [24], $V_u^{\pi_u^*}(s_u^{\tau^{new}})$ can be obtained as

$$V_{u}^{\pi_{u}^{*}}(s_{u}^{\tau^{new}}) = \max_{a'_{u}} Q_{\pi_{u}^{*}}\left(s_{u}^{\tau^{new}}, a_{u}^{\tau^{new}}(\tau)\right)$$
(18)

Adding Equations (17) and (18), we get

$$Q_{\pi^*}(s^{\tau}, a(\tau)) = r(s^{\tau}, a(\tau)) + \gamma \sum_{s' \in S} P_{ss'}a(\tau) \max_{a'_u} Q_{\pi^*_u} \left(s_u^{\tau^{new}}, a_u^{\tau^{new}}(\tau) \right)$$
(19)

The update of the Q-value function is given by [14] as

$$Q_{\pi^*}(s^{\tau}, a(\tau)) = (1 - \alpha)Q_{\pi}(s^{\tau}, a(\tau)) + \alpha[r(s^{\tau}, a(\tau))] + \gamma \max_{a'} Q_{\pi'}(s^{\tau}, a'(\tau'))$$
(20)

where α is learning rate scaled between 0 and 1 and updating speed of $Q_{\pi_u}(s_u^{\tau}, a_u(\tau))$. The RL algorithm shows good performance if size of states is small. In case of high dimensional state space, classical RL approaches fail to perform. Some states are not sampled because of the high dimensional state space, and they require several restrictions. First, the convergence rate might become slow, and storage of the lookup table becomes impractical. Thus, the use of the DRL method was explored to solve the problem with the large space.

4.2. Multi-Agent DQN Frameworks

In contrast to the classical Q-learning approach, the author in [23] proposed the DQN method, which was basically a DRL method. This DQN method relies on two components, e.g., replay memory and target network. The agent stores transitions $(s_u^{\tau}, a_u(\tau), r_u(\tau), s_u^{\tau^{new}})$

in a replay memory *D*. Then, it extracts this transition from memory *D* by using random sampling to compute the Q-value function. The agent uses the memory *D* in a part of mini-batch to train the Q-network, and then a gradient descent method is applied to update the weight parameter θ of the behavior network:

$$\pi_u = \max Q^*_{\pi_u}(s^{\tau}_u, a^{new}_u(\tau))$$

$$a'$$
(21)

In the DQN method, the types of networks are included; that is, DQN sets the θ_{target} target networks. The learning model calculates the target value y_j with a weight parameter θ_{target} for a certain time t, which can mitigate the volatility of the learning scheme. During the learning process, after several iterations of H the weight parameter θ is synchronized with the target network, $\theta \rightarrow \theta_{\text{target}}$. The agent utilizes a greedy random policy, which means that the agent randomly selects an action $a_u(\tau)$ parameter θ for the behavior network. Consequently, $a(\tau)$ value and θ are updated iteratively using the minimum loss function [35]:

$$L(\theta) = \sum [y_j - Q_{\pi_u}(s_u, a_u; \theta)], \qquad (22)$$

where $\boldsymbol{y}_j = r_j + \gamma \max_{a_u^{new}(j)} Q_{\pi_u}(s_u^{j^{new}}, a_u^{new}(j); \theta_{\text{target}}).$

The proposed DQN algorithm for user association and power allocation is shown in Algorithm 1.

Algorithm 1: DQN Based for User Association and Power Allocation Algorithm			
1.	Initialize $Q(s, a) = 0$; learning rate α , target network and replay memory D.		
2.	Set the weight, discount factor		
3.	for each training episode do		
4.	Initialize state s		
5.	Choose a random number		
6.	if $x < \varepsilon$ then		
7.	Choose action randomly;		
8.	else		
9.	Select action $a_u(\tau) = \max_{a^{new}}(s_u^{\tau}, a_u^{new}(\tau); \theta)$		
10.	end if		
11.	Execute action $a_u(\tau)$ and next state $s_u^{\tau^{new}}$		
12.	Calculate energy efficiency using Equation (13).		
13.	Store transition s_u^{τ} , $a_u(\tau)$, $r_u(\tau)$, $s_u^{\tau^{new}}$ in D		
14.	If the replay memory is full then		
15.	Random sampling a mini-batch from D		
16.	Perform gradient descent on $y_j - Q_{\pi_u}(s_u, a_u; heta)^2$ w.r.t. parameter $ heta$		
17	end if		
18	Update target network		
19	End for		

4.3. Parameterized Deep Q-Network Algorithm

The combined user association and power allocation procedure in a hybrid action space can be solved via parameterization, but it still has generalization problems. In order to solve this problem, we used the epsilon greedy exploration method, which enables the DQN to explore a wide variety of states and actions and improves generalization. In a hybrid action space, the Q-value is recast as $Q(s, a) = Q(a, \beta, \chi)$ where β denotes a discrete action, and χ denotes a continuous action. The reward is defined as parameterized double DQN with replay buffer EE. Whether UE is associated with BS or not, the user association will only have the two discrete values $\beta = 0$ and 1. On the other hand, χ is a matrix that represents various levels of power distribution. The PD-DQN for the user association and power allocation algorithm is presented in Algorithm 2.

1116			
1.	Initialize primary and target Deep Q-Networks (DQN) with random weights.		
2.	Set up a Mini-batch and a Replay Buffer for experience storage.		
3.	For each episode:		
Gen	erate an initial state (s_i) by selecting a random action.		
4.	Inside each episode loop:		
	While the episode is ongoing:		
	Choose an action based on an epsilon-greedy policy.		
	If a random number is less than epsilon:		
	Select a discrete action from a predefined set (β).		
	Otherwise:		
	Estimate Q-values for discrete actions using the primary Q-network and		
	choose the highest.		
	If a random number is less than epsilon:		
	Choose a continuous action from a predefined set (χ).		
	Otherwise:		
	Estimate Q-values for continuous actions using the primary Q-network and		
	choose the highest.		
	Execute the selected action in the environment following an epsilon-greedy policy		
5.	After each episode loop, sample a minibatch of experiences from the replay buffer.		
6.	For each experience in the minibatch:		
	Calculate the target Q-value using the target network.		
	If the episode is ongoing, compute the target Q-value for the next state.		
	If the episode is complete, compute the target Q-value with the reward.		
	Calculate the predicted Q-value for the current state.		
	Determine the difference between predicted and target Q-values and update		
	the primary Q-network accordingly.		
	Update the current state, target network weights.		
7.	Update the environment with user associations and power allocations.		
8.	Repeat the process for the next episode if needed.		
9.	End		

Algorithm 2: PD-DON for User Association and Power Allocation Algorithm

5. Simulation Results

In this section, the results of the simulation using the DRL algorithms are presented. We consider a distributed M-MIMO with three RRHs equipped with 300 antennas in the cell with diameter of 2000 m. We consider K = 20 randomly distributed users within the cell. The PC of each RRH is set to 10,000 mW.

The power consumed by the component of active antennas is 200 mW, and the power amplifier efficiency, ν , is 0.25. We assume a transmission bandwidth of 10 MHz [7]. The other parameters are given in Table 1. We consider the Hata-COST231 propagation model [8]. The large-scale-fading β in Equation (1) is borrowed from the literature [8] and is given by $10log_{10}(\beta_{n,u}) = PL_{n,u} + \Omega$ where *PL* is path loss, and Ω represents standard deviation. Here, $d_{n,u}$ is the distance between *u*th users and the nth RRH.

Figure 2 shows the energy efficiency of the proposed DRL algorithm and RL algorithm. Figure 2 demonstrates two observations that indicate that the EE achieved by the DRL method outperforms the RL methods. As the number of episodes increases up to 50, the system EE increases and tends to converge after 250 episodes for both schemes. Additionally, the learning speed of the Q-learning method is lower than that of the multi-agent DQN algorithm. For the Q-learning method, there is a slight improvement in the system EE at episode 120, whereas in the DRL approach, the system EE tends to be stable at episode 257. The EE is unstable at the beginning as seen in the DRL scheme, and the stability increases as the episodes increase, and thereafter increases slowly. This is because the agent selects actions in a random manner and stores the transition information in *D*.

Figure 3 shows the EE versus the number of users fixed at M = 20. From the figure we can see that, the EE generally first increases with K from 5 to 45 and then decreases flatten. This is due to the fact that when scheduling more users, more RRHs are activated to serve

users creating more interference noise. Furthermore, proposed DRL algorithm performs superior to QL.

 Table 1. Simulation parameters.

Parameter	Values
Standard deviation	8 dB
Path loss model PL	$PL = -140.6 - 35\log 10(d)$
Episodes	500
Steps T	500
Discount rate γ	0.9
Mini-batch size b	8
Learning rate	0.01
Replay memory size D	5000



Figure 2. Convergence of energy efficiency values.



Figure 3. EE versus number of users.

Figure 4 compares the EE performance at different discounted factors, $\gamma = \{0.1, 0.5, and 0.8\}$. It can be observed that a lower discount factor results in higher EE through different discount factors. Figure 4 illustrates that the PD-DQN method optimizes the user association and power allocation. Moreover, when the number of epochs increases, the EE performance of each user is better at a different discount factor.



Figure 4. EE versus number of Epochs.

From the Figure 5, we can observe that PDQN consistently outperforms both DQN and RL in terms of energy efficiency. The EE values achieved by PDQN show a steady increase, starting from 1.5 and reaching 5.25, indicating significant improvement. DQN and RL also exhibit improvements, but their EE values remain below that of PDQN. In terms of percentage improvement, PDQN surpasses DQN by an average of around 35%, while PDQN outperforms RL by approximately 40%. Additionally, DQN exhibits a slight advantage over RL, with an average improvement of about 5%.



Figure 5. EE versus number of antennas.

6. Conclusions

In this paper, we have studied the user association and power allocation problem in a massive MIMO based on a PD-DQN framework. The numerical results indicate that the PD-DQN approach performs better than the DQN and classical Q-learning schemes. The main motivation of this paper is to study the resource allocation scheme in M-MIMO. In addition, for the simulation results in this study, we considered the DQN approach to tackle the problem of user association and power allocation in M-MIMO. The aforementioned optimization problem was formulated to maximize the EE in the downlink network, and the convergence of the multi-agent DRL (DQN) algorithm was studied. Furthermore, convergence analyses confirmed that the proposed methods perform better in terms of EE than the RL method. Establishing the upper bound is important to set a benchmark for the achievable EE. In our study, this upper bound signifies the theoretical maximum EE that can be achieved in the given massive MIMO resource allocation problem under ideal conditions. Our study has showcased the potential of our PD-DQN framework to enhance energy efficiency in massive MIMO systems. The inclusion of the upper bound in our evaluation has not only validated the effectiveness of our approach but has also shed light on areas for further improvement. The convergence rate indicates that the proposed algorithm excels in terms of energy efficiency. Furthermore, additional simulation outcomes demonstrate superior energy efficiency across varying user counts, number of antennas, and diverse learning rates. The enhancement of the proposed PD-DQN on average may reach 72.2% and 108.5% over the traditional DQN and RL methods, respectively.

Author Contributions: Methodology, S.S.; Formal analysis, W.Y.; Investigation, S.S. and W.Y.; Writing—original draft, S.S.; Writing—review & editing, W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF), Ministry of Education, Science and Technology (Grant No. 2016R1A2B4012752).

Data Availability Statement: Data is confidential and can be made available on request from corresponding authors.

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

References

- Li, H.; Wang, Z.; Wang, H. An energy-efficient power allocation scheme for Massive MIMO systems with imperfect CSI. Digit. Signal Process. 2021, 112, 102964. [CrossRef]
- Rajoria, S.; Trivedi, A.; Godfrey, W.W.; Pawar, P. Resource Allocation and User Association in Massive MIMO Enabled Wireless Backhaul Network. In Proceedings of the IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 28 April–1 May 2019; pp. 1–6. [CrossRef]
- Ge, X.; Li, X.; Jin, H.; Cheng, J.; Leung, V.C.M. Joint user association and user scheduling for load balancing in heterogeneous networks. *IEEE Trans. Wireless Commun.* 2018, 17, 3211–3225. [CrossRef]
- Liang, L.; Kim, J.; Jha, S.C.; Sivanesan, K.; Li, G.Y. Spectrum and power allocation for vehicular communications with delayed CSI feedback. *IEEE Wirel. Commun. Lett.* 2017, 6, 458–461. [CrossRef]
- Bu, G.; Jiang, J. Reinforcement Learning-Based User Scheduling and Resource Allocation for Massive MU-MIMO System. In Proceedings of the 2019 IEEE/CIC International Conference on Communications in China (ICCC), Changchun, China, 11–13 August 2019; pp. 641–646. [CrossRef]
- Yang, K.; Wang, L.; Wang, S.; Zhang, X. Optimization of resource allocation and user association for energy efficiency in future wireless networks. *IEEE Access* 2017, 5, 16469–16477. [CrossRef]
- Dong, G.; Zhang, H.; Jin, S.; Yuan, D. Energy-Efficiency-Oriented Joint User Association and Power Allocation in Distributed Massive MIMO Systems. *IEEE Trans. Veh. Technol.* 2019, 68, 5794–5808. [CrossRef]
- Ngo, H.Q.; Ashikhmin, A.; Yang, H.; Larsson, E.G.; Marzetta, T.L. Cell-free massive MIMO versus small cells. *IEEE Trans.* Wirel. Commun. 2017, 16, 1834–1850. [CrossRef]
- Elsherif, A.R.; Chen, W.-P.; Ito, A.; Ding, Z. Resource Allocation And Inter-Cell Interference Management For Dual-Access Small Cells. *IEEE J. Sel. Areas Commun.* 2015, 33, 1082–1096. [CrossRef]
- Sheng, J.; Tang, Z.; Wu, C.; Ai, B.; Wang, Y. Game Theory-Based Multi-Objective Optimization Interference Alignment Algorithm for HSR 5G Heterogeneous Ultra-Dense Network. *IEEE Trans. Veh. Technol.* 2020, 69, 13371–13382. [CrossRef]
- 11. Zhang, X.; Sun, S. Dynamic scheduling for wireless multicast in massive MIMO HetNet. Phys. Commun. 2018, 27, 1–6. [CrossRef]

- 12. Nassar, A.; Yilmaz, Y. Reinforcement Learning for Adaptive Resource Allocation in Fog RAN for IoT with Heterogeneous Latency Requirements. *IEEE Access* 2019, *7*, 128014–128025. [CrossRef]
- Sun, Y.; Feng, G.; Qin, S.; Liang, Y.-C.; Yum, T.P. The Smart Handoff Policy for Millimeter Wave Heterogeneous Cellular Networks. IEEE Trans. Mob. Comput. 2018, 17, 1456–1468. [CrossRef]
- 14. Watkins, C.J.C.H.; Dayan, P. Q-Learning. Mach. Learn. 1992, 8, 279–292. [CrossRef]
- 15. Zhai, Q.; Bolić, M.; Li, Y.; Cheng, W.; Liu, C. A Q-Learning-Based Resource Allocation for Downlink Non-Orthogonal Multiple Access Systems Considering QoS. *IEEE Access* 2021, *9*, 72702–72711. [CrossRef]
- Amiri, R.; Mehrpouyan, H.; Fridman, L.; Mallik, R.K.; Nallanathan, A.; Matolak, D. A machine learning approach for power allocation in HetNets considering QoS. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–7. [CrossRef]
- Ghadimi, E.; Calabrese, F.D.; Peters, G.; Soldati, P. A reinforcement learning approach to power control and rate adaptation in cellular networks. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–7. [CrossRef]
- Meng, F.; Chen, P.; Wu, L. Power allocation in multi-user cellular networks with deep Q learning approach. In Proceedings of the ICC 2019—2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–7. [CrossRef]
- 19. Ye, H.; Li, G.Y.; Juang, B.F. Deep reinforcement learning based resource allocation for v2v communications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3163–3173. [CrossRef]
- 20. 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]
- Sun, Y.; Peng, M.; Mao, S. Deep reinforcement learning-based mode selection and resource management for green fog radio access networks. *IEEE Internet Things J.* 2019, 6, 960–1971. [CrossRef]
- Rahimi, A.; Ziaeddini, A.; Gonglee, S. A novel approach to efficient resource allocation in load-balanced cellular networks using hierarchical DRL. J. Ambient. Intell. Humaniz. Comput. 2021, 13, 2887–2901. [CrossRef]
- 23. Zhao, N.; Liang, Y.-C.; Niyato, D.; Pei, Y.; Wu, M.; Jiang, Y. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 5141–5152. [CrossRef]
- 24. Nasir, Y.S.; Guo, D. Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks. *IEEE J. Sel. Areas Commun.* 2019, 37, 2239–2250. [CrossRef]
- Xu, Y.; Yu, J.; Headley, W.C.; Buehrer, R.M. Deep Reinforcement Learning for Dynamic Spectrum Access in Wireless Networks. In Proceedings of the 2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, USA, 29–31 October 2018; pp. 207–212. [CrossRef]
- Li, M.; Zhao, X.; Liang, H.; Hu, F. Deep reinforcement learning optimal transmission policy for communication systems with energy harvesting and adaptive mqam. *IEEE Trans. Veh. Technol.* 2019, 68, 5782–5793. [CrossRef]
- Su, Y.; Lu, X.; Zhao, Y.; Huang, L.; Du, X. Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks. *IEEE Sens. J.* 2019, 19, 9561–9569. [CrossRef]
- Xiong, J.; Wang, Q.; Yang, Z.; Sun, P.; Han, L.; Zheng, Y.; Fu, H.; Zhang, T.; Liu, J.; Liu, H. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space. arXiv 2018, arXiv:1810.06394.
- Lin, Z.; Lin, M.; Champagne, B.; Zhu, W.-P.; Al-Dhahir, N. Secrecy-Energy Efficient Hybrid Beamforming for Satellite-Terrestrial Integrated Networks. *IEEE Trans. Commun.* 2021, 69, 6345–6360. [CrossRef]
- Lin, Z.; Niu, H.; An, K.; Hu, Y.; Li, D.; Wang, J.; Al-Dhahir, N. Pain without Gain: Destructive Beamforming from a Malicious RIS Perspective in IoT Networks. *IEEE Internet Things J.* 2023. [CrossRef]
- An, K.; Lin, M.; Ouyang, J.; Zhu, W.-P. Secure Transmission in Cognitive Satellite Terrestrial Networks. *IEEE J. Sel. Areas Commun.* 2016, 34, 3025–3037. [CrossRef]
- Lin, Z.; Niu, H.; An, K.; Wang, Y.; Zheng, G.; Chatzinotas, S.; Hu, Y. Refracting RIS-Aided Hybrid Satellite-Terrestrial Relay Networks: Joint Beamforming Design and Optimization. *IEEE Trans. Aerosp. Electron. Syst.* 2022, 58, 3717–3724. [CrossRef]
- Hsieh, C.-K.; Chan, K.-L.; Chien, F.-T. Energy-Efficient Power Allocation and User Association in Heterogeneous Networks with Deep Reinforcement Learning. *Appl. Sci.* 2021, 11, 4135. [CrossRef]
- 34. Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction; MIT Press: Cambridge, UK, 1998.
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* 2015, 518, 529–533. [CrossRef] [PubMed]

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