Energy Efficient Power Allocation in Massive MIMO based on Parameterized Deep DQN

Shruti Sharma * and Wonsik Yoon *

Posted Date: 3 October 2023

doi: 10.20944/preprints202310.0066.v1

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

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, South Korea
* Correspondence: shruti@ajou.ac.kr and wsyoon@ajou.ac.kr

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 RL method respectively in handling large-scale multi-agent problems in M-MIMO networks.

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

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 a reliable option 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 have 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 the resource allocation problem was investigated in [7, 8]. Given the non-convex multiple objective function in the M-MIMO problem, achieving the Pareto optimal solution set in 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 multi-objective 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.

 doi:10.20944/preprints202310.0066.v1

The Q-learning algorithm is one of the widely used RL methods because it requires a minimum computation. It can be expressed by 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 channel, access, and mobile computing [19-21].

In DRL, 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 [23], Rahimi et al. gave an algorithm of DQN which was based on deep neural networks and has been previously used in past literature data. In [23], Zhao et al. used the DRL method for the efficient management of user association and resource allocation for maximizing the network utility and maintains the quality of service (QoS) requirements. In [25-27], the authors proposed a DQN algorithm to allocate power using a multi-agent DRL method. Recent advancements DRL, particularly techniques like Deep Q-Networks (DQN), have opened up new avenues for addressing resource allocation challenges in wireless networks. However, when it comes to applying DQN 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 the overall performance. Additionally, the complexity of DQN grows significantly as the dimension of the action space increases. This exponential complexity can lead to high power consumption and slow convergence rates, which are highly undesirable in practical applications. 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 sub-optimal policies due to its tendency to overestimate Q-values [28]. 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 [25]. 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 PD-DQN algorithm [25]. The main contributions of this paper are listed in the following:

- This paper proposes a user association and power allocation problem with objective of maximizing EE in a massive MIMO network.
- To solve the power allocation problems, the action space, the state space, and the reward function have been considered. We apply the model-free DQN framework and PD-DQN to update policies in action space. We also employ the novel PD-DQN framework that is able to updating policies in a hybrid discrete-continuous action space.
- The simulation results show that the proposed user association and power allocation method based on PD-DQN perform better than DRL and RL method.

2. System Model

In this study, we considered 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

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.

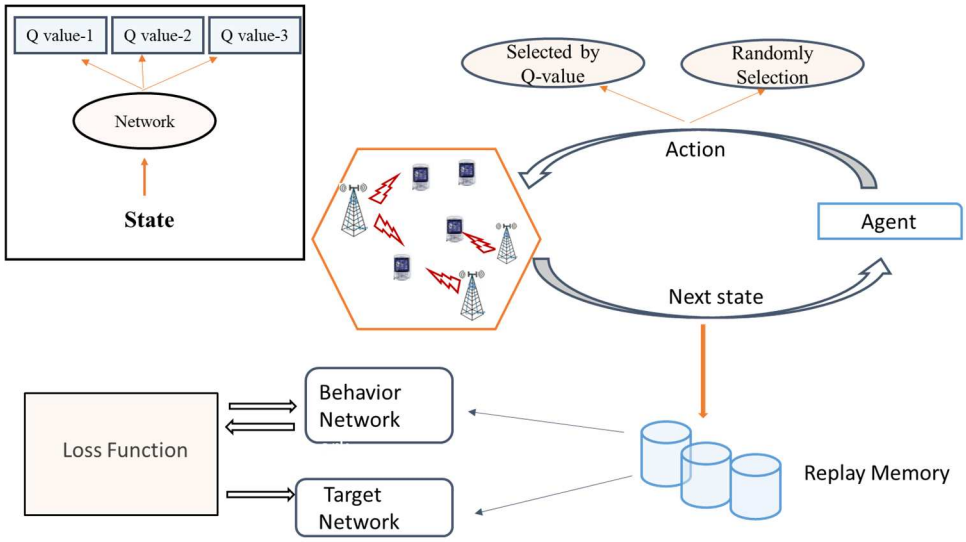

Figure 1. System Model based on DRL.

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

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

where $\boldsymbol{\beta_{n,u}}$ signifies the large-scale fading coefficient, and $\mathbf{g}_{n,u}$ signifies the small-scale fading coefficient. $\mathbf{g}_{n,u}$ 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]

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

where $\boldsymbol{p_u}$ is power transmitted through *u*th user, $\boldsymbol{s_{n,u}}$ is the $\boldsymbol{N_r} \times \mathbf{1}$ data symbol of the *u*th user on the *n*th RRH, $\mathbf{w}_{n,u}$ is the $\boldsymbol{N_t} \times \boldsymbol{N_r}$ beamforming vector of the *u*th user on the *n*th RRH which is given by $\mathbf{w}_{n,u} = \mathbf{h}_{n,u} \Big/ \sqrt{\mathbb{E}\left\{\left\|\hat{\mathbf{h}}_{n,u}\right\|^2\right\}}$ 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 identically distributed (i.i.d.) additive complex Gaussian noise having zero mean and variance of $\boldsymbol{\sigma}$.

Without loss of generality, we set

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

2.1. Power consumption

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

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

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

3. Problem formulation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

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

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

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

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

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

$$\boldsymbol{x_{n,u} = \begin{cases} 1, u \in L_l \\ 0, u \in \dfrac{U}{L_l} \end{cases} l = 1, \ldots L} \tag{6}$$

The system energy efficiency (EE) can be expressed as

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

The optimization problem maximizing the system EE can be formulated as

$$\text{P1:} \quad \max_{X,P} \boldsymbol{\eta = \frac{\sum_{u=1}^{U} R_u(X,P)}{P_c(X,P)}}, \tag{8}$$

$$\text{s.t} \quad \text{C1:} \quad \boldsymbol{0 < P_{u,m} \leq P_u^{\max}}$$

$$\text{C2:} \quad \boldsymbol{a_{u,m} \in \{0, 1\}}$$

$$\text{C3:} \quad \boldsymbol{\sum_m a_{u,m} = 1}$$

$$\text{C4:} \quad \boldsymbol{R_u \geq R_{\min}}$$

In the above problem, C1 denotes 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 [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, $\boldsymbol{P_{ss^{new}}}$) similar to the existing work [26-27], where s, a, and r represent the set of state, set of action, and reward functions, respectively. $\boldsymbol{P_{ss^{new}}}$ is the transition probability from state s to $\boldsymbol{s^{new}}$ 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

$$\boldsymbol{s(t) = \{s_1(t), s_2(t), \ldots s_u(t), \ldots, s_U(t)\}} \tag{9}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

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

$$\boldsymbol{a(t)} = \{\boldsymbol{a_1(t)}, \boldsymbol{a_2(t)}, .. \boldsymbol{as_u(t)}, \ldots, \boldsymbol{a_U(t)}\} \tag{10}$$

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

$$\boldsymbol{r(t)} = \sum_{u}^{U} \boldsymbol{r_u(t)} = \sum_{u=1}^{U} \boldsymbol{\eta_u(t)} \tag{11}$$

where is $\boldsymbol{r(t)}$ 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:

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

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 $\boldsymbol{\pi}$ to perform an action *a*(t) that is, $\boldsymbol{a(t)} = \boldsymbol{\pi(s(t))}$. Following this, the users communicate with the BSs and the reward function becomes $\boldsymbol{r(t)} = \boldsymbol{r(t|s = s(t), a = a(t))}$. Therefore, the future cumulative discounted reward at time t can be given by

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

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

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

where E[·] denotes the expectation operator. By Markov property, the value function is defined as

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

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

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

The optimal Q-value function satisfies the Bellman equation [29-30] derived as

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

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

$$\boldsymbol{V_u^{\pi_u^*}(s_u^{\tau^{new}})} = \boldsymbol{\max_{a'_u} Q_{\pi_u^*}\left(s_u^{\tau^{new}}, a_u^{\tau^{new}}(\tau)\right)} \tag{18}$$

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

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

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

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

where $\boldsymbol{\alpha}$ is learning rate scaled between 0 and 1 and updating speed of $\boldsymbol{Q_{\pi_u}(s_u^{\tau}, a_u(\tau))}$ 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 require several restrictions. First, the convergence rate might become slow and storage of lookup table becomes impractical. Thus, the use of the DRL method was explored to solve the problem with the large space.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

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 DRL method. This DQN method relies on two components, e.g., replay memory and target network. The agent stores transitions $\left(\boldsymbol{s_u^{\tau}}, \boldsymbol{a_u(\tau)}, \boldsymbol{r_u(\tau)}, \boldsymbol{s_u^{\tau^{new}}}\right)$in a replay memory *D*. Then extract this transition from memory *D* by using random sampling to compute 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 $\boldsymbol{\theta}$ of behavior network.

$$\boldsymbol{\pi_u} = \mathbf{max}_{\boldsymbol{a'}} \boldsymbol{Q^*_{\pi_u}}(\boldsymbol{s_u^{\tau}}, \boldsymbol{a_u^{new}(\tau)}) \tag{21}$$

In the DQN method, the types of networks are included, that is, DQN sets the $\boldsymbol{\theta}_{\text{target}}$ target networks. The learning model calculates the target value$\boldsymbol{y_j}$with a weight parameter$\boldsymbol{\theta}_{\text{target}}$for a certain time *t*, which can mitigate the volatility of the learning scheme. During the learning process, after several iterations *H* the weight parameter *θ* is synchronized with the target network $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_{\text{target}}$. The agent utilizes a greedy random policy means that the agent randomly selects an action $\boldsymbol{a_u(\tau)}$ parameter $\boldsymbol{\theta}$ for the behavior network. Consequently, $\boldsymbol{a(\tau)}$ value and $\boldsymbol{\theta}$ are updated iteratively using the minimum loss function [31]:

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

where $\boldsymbol{y_j} = \boldsymbol{r_j} + \boldsymbol{\gamma} \underset{\boldsymbol{a_u^{new}(j)}}{\boldsymbol{max}} \boldsymbol{Q_{\pi_u}}(\boldsymbol{s_u^{j^{new}}}, \boldsymbol{a_u^{new}(j)}; \boldsymbol{\theta}_{\text{target}})$

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 $\boldsymbol{x} < \boldsymbol{\varepsilon}$ then
7. Choose action randomly;
8. else
9. Select action $\boldsymbol{a_u(\tau)} = \underset{\boldsymbol{a^{new}}}{\boldsymbol{max}}(\boldsymbol{s_u^{\tau}}, \boldsymbol{a_u^{new}(\tau)}; \boldsymbol{\theta})$
10. end if
11. Execute action $\boldsymbol{a_u(\tau)}$ and next state $\boldsymbol{s_u^{\tau^{new}}}$
12. Calculate energy efficiency using Eq. (13).
13. Store transition $\boldsymbol{s_u^{\tau}}, \boldsymbol{a_u(\tau)}, \boldsymbol{r_u(\tau)}, \boldsymbol{s_u^{\tau^{new}}}$in *D*
14. If the replay memory is full then
15. Random sampling a mini-batch from *D*
16. Perform gradient descent on $\boldsymbol{y_j} - \boldsymbol{Q_{\pi_u}}(\boldsymbol{s_u}, \boldsymbol{a_u}; \boldsymbol{\theta})^2$ w.r.t. parameter $\boldsymbol{\theta}$
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 $\boldsymbol{Q(s, a)} = \boldsymbol{Q(a, \beta, \chi)}$ where $\boldsymbol{\beta}$ denotes a discrete action and $\boldsymbol{\chi}$ denotes a continuous action. The reward is defined as parameterized double DQN with replay buffer EE. Whether UE is associated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

with BS or not, the user association will only have the two discrete values $\boldsymbol{\beta}$ = 0 and 1. On the other hand, $\boldsymbol{\chi}$ is a matrix that represents various levels of power distribution. PD-DQN for user association and power allocation algorithm is presented in Algorithm 2.

Algorithm 2. PD-DQN for user association and power allocation algorithm

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:

Generate 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 ($\boldsymbol{\beta}$).
   - 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 ($\boldsymbol{\chi}$).
   - 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 done, 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

5. Simulation Result

In this section, the results of the simulation with the DRL algorithms are presented. We considered 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 10000 mW.

The power consumed by component of active antennas is 200 mW, and the power ampli fier efficiency, ν is 0.25. We assumed 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 eq (1) borrowed from literature [8] which is given by $\mathbf{10}\,\boldsymbol{log}_{\mathbf{10}}(\boldsymbol{\beta}_{\boldsymbol{n,u}}) = \boldsymbol{PL}_{\boldsymbol{n,u}} + \boldsymbol{\Omega}$ where PL is path loss and Ω represents standard deviation. Here, $d_{n,u}$ is the distance between uth users and nth RRH.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

Table 1. Simulation parameters.

Parameter	Values
Standard Deviation	8 dB
Path loss model PL	PL=−140.6−35log10(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 shows the energy efficiency of the proposed DRL algorithm and RL algorithm. Figure 2 gives two observations 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 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*.

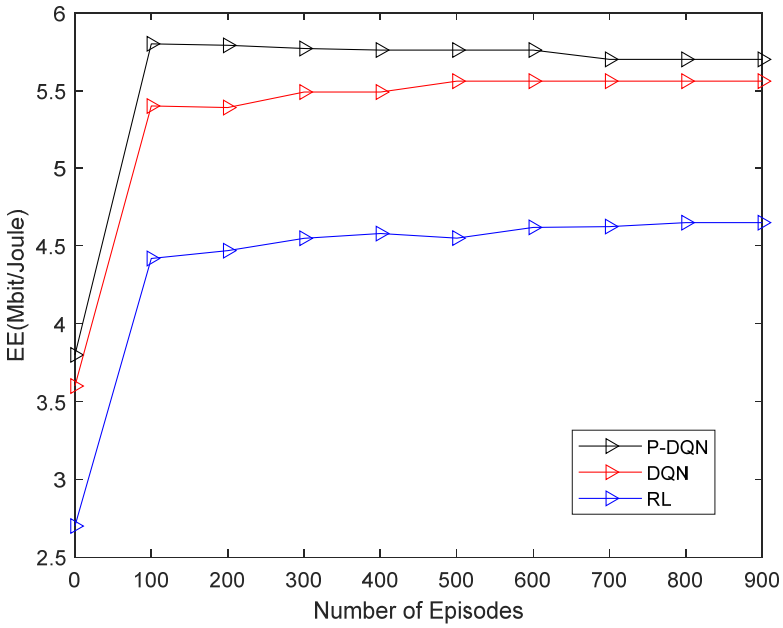

Figure 2. Convergence of energy efficiency values.

Figures 3 shows the EE versus the number of user fix *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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

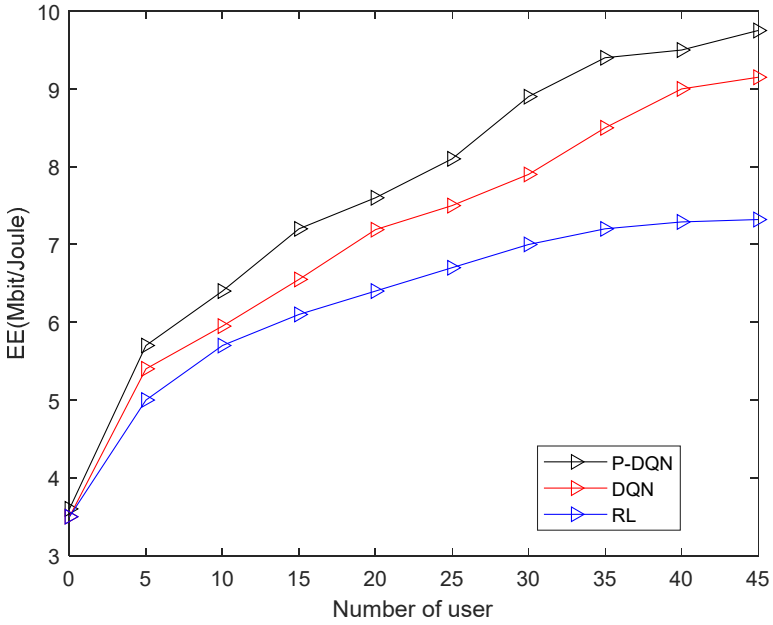

Figure 3. EE versus Number of user.

Figure 4 compares the EE performance at different discounted factors, γ = {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 methods optimize 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.

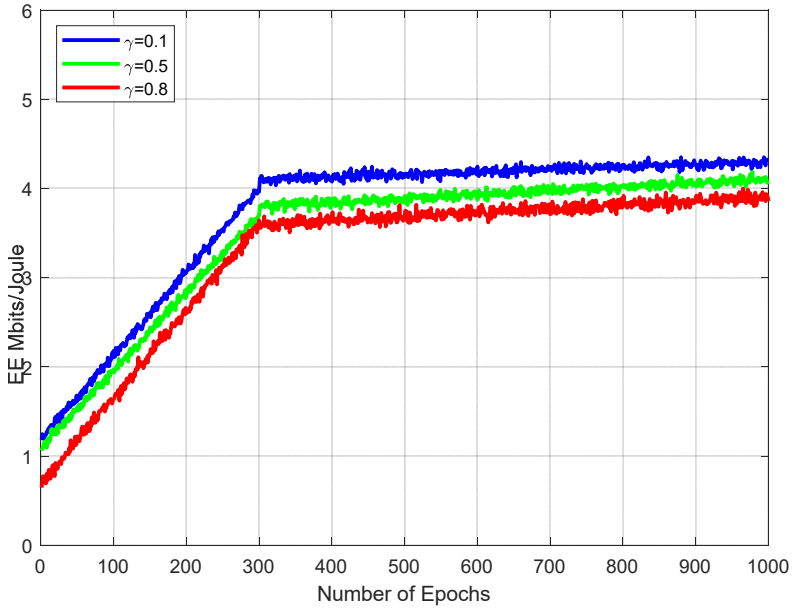

Figure 4. EE versus Number of users.

From the figure, 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%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0066.v1

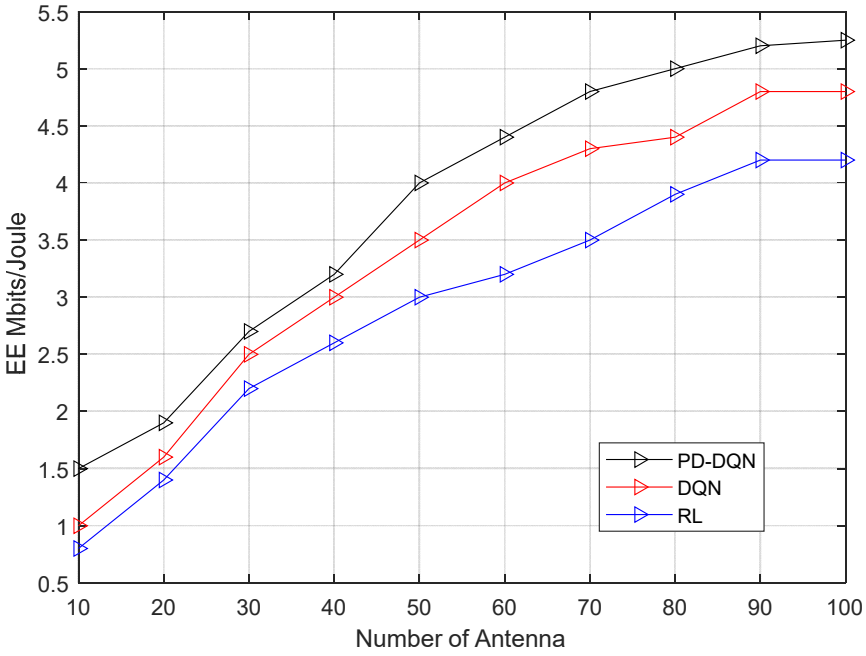

Figure 5. EE versus number of antennas.

5. Conclusion

In this paper, we have studied the user association and power allocation problem in a massive MIMO based on PD-DQN framework. The numerical results indicate that the PD-DQN approach performs better than the DQN and classical Q-learning scheme. The main motivation of this paper is to study resource allocation scheme in M-MIMO. In addition, for the simulation results in this study, we considered DQN approach to tackle the problem of user association and power allocation in M-MIMO. The aforedescribed 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. 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 respectively.

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

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

References

1. Hao L, Zhigang W, & Houjun W. (2021). An energy-efficient power allocation scheme for Massive MIMO systems with imperfect CSI, *Digital Signal Processing*, 112.doi.org/10.1016/j.dsp.2021.102964
2. Rajoria, S., Trivedi, A., Godfrey, W. W., & Pawar, P. (2019). Resource Allocation and User Association in Massive MIMO Enabled Wireless Backhaul Network. *IEEE 89th Vehicular Technology Conference (VTC2019-Spring),* (pp. 1-6). doi: 10.1109/VTCSpring.2019.8746401.
3. Ge, X., Li, X., Jin, H., Cheng, J., Leung, V.C.M., (2018). Joint user association and user scheduling for load balancing in heterogeneous networks, IEEE Trans. Wireless Commun. 17 (5), 3211–3225, doi:10.1109/TWC.2018.2808488.
4. Liang, L., and Kim, J., and Jha, S. C., and Sivanesan, K., and Li, G. Y. (2017). Spectrum and power allocation for vehicular communications with delayed CSI feedback, IEEE Wireless Communications Letters, 6, 458–461, doi: 10.1109/LWC.2017.2702747.

5. Bu, G., and Jiang, J. (2019). Reinforcement Learning-Based User Scheduling and Resource Allocation for Massive MU-MIMO System, 2019 IEEE/CIC International Conference on Communications in China (ICCC), Changchun, China, 2019, pp. 641-646, doi: 10.1109/ICCChina.2019.8855949
6. Yang, K., Wang, L., Wang, S, Zhang, X. (2017). Optimization of resource allocation and user association for energy efficiency in future wireless networks, IEEE Access., 5, 16469-16477, doi: 10.1109/ACCESS.2017.2722007.
7. Dong, G., Zhang, H., Jin, S., and Yuan, D. (2019). Energy-Efficiency-Oriented Joint User Association and Power Allocation in Distributed Massive MIMO Systems, in IEEE Transactions on Vehicular Technology, vol. 68 (6), 5794-5808. doi: 10.1109/TVT.2019.2912388.
8. Ngo, H. Q. et al. (2017). Cell-free massive MIMO versus small cells. IEEE Transaction on Wireless Communication, 16, 1834–1850, doi:10.1109/TWC.2017.2655515.
9. Elsherif, A. R, Chen, W.-P., Ito, A. and Ding, Z. (2015). Resource Allocation And Inter-Cell Interference Management For Dual-Access Small Cells. IEEE Journal of Selected Areas In Communication. 33 (6), 1082-1096, doi: 10.1109/JSAC.2015.2416990.
10. Sheng, J., Tang, Z., Wu, C., Ai, B. and Wang, Y. (2020). Game Theory-Based Multi-Objective Optimization Interference Alignment Algorithm for HSR 5G Heterogeneous Ultra-Dense Network. in IEEE Transactions on Vehicular Technology, 69(11), 13371-13382. doi: 10.1109/TVT.2020.3025778.
11. Zhang, X., Sun, S. (2018). Dynamic scheduling for wireless multicast in massive MIMO HetNet, Physical Communication, 27, 1-6. doi : 10.1016/j.phycom.2017.12.015.
12. Nassar, A., Yilmaz, Y. (2019). Reinforcement Learning for Adaptive Resource Allocation in Fog RAN for IoT with Heterogeneous Latency Requirements. in IEEE Access, 7, 128014-128025. doi :10.1109/ACCESS.2019.2939735 .
13. Sun, Y., Feng, G., Qin, S, Liang, Y.-C, and Yum. T. P. (2018). The Smart Handoff Policy For Millimeter Wave Heterogeneous Cellular Networks, IEEE Trans. Mobile Comput., 17 (6), 1456-1468, 2018, doi: 10.1109/TMC.2017.2762668.
14. Watkins, C. J., and Dayan, P. (1992). Q-Learning, Machine Learning, 8 (3-4), 279-292, doi: 10.1007/BF00992698.
15. Zhai, Q., Bolić, M., Li, Y., Cheng, W. and Liu, C. (2021). A Q-Learning-Based Resource Allocation for Downlink Non-Orthogonal Multiple Access Systems Considering QoS. in IEEE Access, 9, 72702-72711, doi: 10.1109/ACCESS.2021.3080283.
16. AMIRI, R., et al. (2018). A machine learning approach for power allocation in HetNets considering QoS. In The Proceedings of 2018 IEEE International Conference on Communications (ICC). Kansas City (MO, USA), 2018, p. 1–7. DOI: 10.1109/ICC.2018.8422864
17. Ghadimi, E., Calabrese, F. D. Peters, G. and Soldati, P. (2017). A reinforcement learning approach to power control and rate adaptation in cellular networks, in Proc. IEEE Int. Conf. Commun. *(ICC)*, 2017, pp. 1-7. doi: 10.1109/ICC.2017.7997440.
18. F. Meng, P. Chen, and L. Wu, Power allocation in multi-user cellular networks with deep Q learning approach, in Proc. IEEE Int. Conf. Commun (ICC), 2019, pp. 1–7, doi: 10.1109/ICC.2019.8761431.
19. Ye, H., Li, G.Y., Juang, B.F. (2019). Deep reinforcement learning based resource allocation for v2v communications, IEEE Trans. Veh. Technol. 68 (4), 3163–3173. doi: 10.1109/TVT.2019. 2897134.
20. Wei, Y., Yu, F.R., Song, M., Han, Z. (2019). Joint optimization of caching, computing, and radio resources for fog-enabled IOT using natural actor critic deep reinforcement learning. IEEE Internet Things J. 6 (22), 2061–2073, doi: 10.1109/JIOT. 2018.2878435.
21. Sun, Y., Peng, M., Mao,S. (2019). Deep reinforcement learning-based mode selection and resource management for green fog radio access networks. IEEE Internet Things J., 6 (2), 960–1971, doi: 10.1109/JIOT.2018.2871020.
22. Rahimi, A., Ziaeddini, A. & Gonglee, S. (2021) A novel approach to efficient resource allocation in load-balanced cellular networks using hierarchical DRL. J Ambient Intell Human Comput. doi: 1007/s12652-021-03174-0.
23. Zhao, N., Liang, Y.-C., Niyato, D., Pei, Y., Wu, M. and Jiang, Y(2018). Deep reinforcement learning for user association and resource allocation in heterogeneous networks. in IEEE Globecom, Abu Dhabi, UAE, Dec. 2018, pp. 1–6, doi: 10.1109/TWC.2019.2933417.

24. Nasi, Y. S. and Guo, D. (2019), Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks, in IEEE J. Sel. Areas in Commun, 37 (10), 2239-2250, doi: 10.1109/JSAC.2019.2933973.
25. Xu, Y., Yu, J., and William C. H. and Buehrer, R (2018). Deep Reinforcement Learning for Dynamic Spectrum Access in Wireless Networks, 2018 IEEE Military Communications Conference (MILCOM), pp. 207-212, doi: 10.1109/MILCOM.2018.8599723.
26. Li, M., Zhao, X., Liang, H., Hu. F., (2019). Deep reinforcement learning optimal transmis- sion policy for communication systems with energy harvesting and adaptive mqam, IEEE Trans. Veh. Technol. 68 (6), 5782–5793. doi: 10.1109/TVT. 2019.2911544.
27. Su, Y. Lu, X. Zhao, Y. Huang, L., Du, X. (2019). Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks. IEEE Sensors Journal, 19(20), 9561-9569. doi: 10.1109/JSEN.2019.2925719.
28. 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.
29. Hsieh C-K, Chan K-L, Chien F-T. Energy-Efficient Power Allocation and User Association in Heterogeneous Networks with Deep Reinforcement Learning. *Applied Sciences*. 2021; 11(9):4135. https://doi.org/10.3390/app11094135.
30. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press Cambridge, 1998.
31. Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. Nature, 518 (7540), 529–533, doi:10.1038/nature1423