applied sciences

MDPI

Article

Energy-Efficient Power Allocation and User Association in Heterogeneous Networks with Deep Reinforcement Learning

Chi-Kai Hsieh, Kun-Lin Chan and Feng-Tsun Chien *

Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan; mee29616243.ee04@g2.nctu.edu.tw (C.-K.H.); jaja567404.ee04@nctu.edu.tw (K.-L.C.)
* Correspondence: ftchien@mail.nctu.edu.tw

Abstract: This paper studies the problem of joint power allocation and user association in wireless heterogeneous networks (HetNets) with a deep reinforcement learning (DRL)-based approach. This is a challenging problem since the action space is hybrid, consisting of continuous actions (power allocation) and discrete actions (device association). Instead of quantizing the continuous space (i.e., possible values of powers) into a set of discrete alternatives and applying traditional deep reinforcement approaches such as deep Q learning, we propose working on the hybrid space directly by using the novel parameterized deep Q-network (P-DQN) to update the learning policy and maximize the average cumulative reward. Furthermore, we incorporate the constraints of limited wireless backhaul capacity and the quality-of-service (QoS) of each user equipment (UE) into the learning process. Simulation results show that the proposed P-DQN outperforms the traditional approaches, such as the DQN and distance-based association, in terms of energy efficiency while satisfying the QoS and backhaul capacity constraints. The improvement in the energy efficiency of the proposed P-DQN on average may reach 77.6% and 140.6% over the traditional DQN and distance-based association approaches, respectively, in a HetNet with three SBS and five UEs.

Keywords: energy efficiency; power allocation; user clustering; reinforcement learning

Citation: 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. https://doi.org/10.3390/app11094135

Academic Editor: Francesco Guidi

Received: 10 April 2021
Accepted: 29 April 2021
Published: 30 April 2021

1. Introduction

With the exponential growth of wireless Internet-of-Things (IoT) sensors and ultra-reliable requirement in the next-generation cellular networks, the global mobile data traffic is expected to reach about 1 zettabyte by 2022 according to Cisco's forecast [1]. To meet the demands of higher data traffic in wireless links either in fixed sensors for IoT networks or mobile devices in cellular networks, the network infrastructure inevitably will need to expand dramatically, which will result in tremendous escalation of energy consumption and backhaul traffic. Energy efficiency and spectral efficiency are therefore critical issues when designing next-generation wireless communication systems with enabling quality-of-service (QoS) guarantees for radio devices with considerations of efficient power consumption [2,3].

Enhancing the cell density is one of the approaches to meet the need of high date rate under limited bandwidth in centralized networks. Heterogeneous networks (HetNets) have therefore emerged as a standard part of future mobile networks to improve the system capacity and energy efficiency through more flexible design of transmission power allocation and smaller coverage sizes by densely deployed small base stations (SBSs) [3–5]. However, the interference problem caused from various SBSs is the primary challenge for effective system capacity improvement.

1.1. Motivation

Cell densification of HetNet enhances the spectrum efficiency in an energy-efficient way. On the other hand, cell densification potentially increases the inter-cell interference, especially at the cell edges, which deteriorates the QoS of user equipments (UEs).

Interference coordination schemes by means of radio resource allocation and power control can be implemented to achieve higher energy efficiency and spectral efficiency [2]. However, the growing complexity of wireless networks due to increased links and heterogeneous network structure create tremendous challenges for system designs, thus calling for more intelligent techniques for effective yet efficient resource management strategies. In this perspective, data-driven machine learning techniques have been regarded as viable new approaches to dealing with complex network dynamics [5–10]. Compared with traditional model-based algorithms [11–14], deep reinforcement learning (DRL), leveraging recent advances in deep neural networks with reinforcement learning [15–17], can autonomously extract features from the raw data with different formats and complex correlations experienced by the mobile environments. This potentially reduces the cost of data pre-processing [11,18]. In view of this, in this paper, we provide a reinforcement-learning-based solution for power allocation and radio device association with the objective of maximizing energy efficiency while satisfying the required QoS and wireless backhaul capacity constraints.

1.2. Prior Work

In order to intelligently manage the interwoven dynamics underlying the wireless sensor or mobile networks in which a variety of network parameters are generally unknown, deep reinforcement learning (DRL)-based approaches have been applied to tackle the challenges of radio resource management in wireless networks, e.g., [11,16,18–27], due to DRL's ability to extract features from raw data, learn complex correlations generated by the mobile environment, and make sequential decisions through interactions with the environment without knowledge of complete environment information. In the applications of power allocation, the objectives can often be categorized into three types: capacity maximization in [11,19,20], energy saving in [16], and maximization of capacity for consumed energy (which is defined as the energy efficiency) in [21–24].

Meng et al. [11] propose several DL-based algorithms to handle the power allocation with the aim to maximize the sum rate in multiuser wireless cellular networks, in which the DL-based data-driven approaches are demonstrated to outperform the traditional model-based power allocation method. Nasir and Guo [16] utilize multi-agent deep RL to adaptively control the discrete power level (i.e., the range of possible power values is quantized into a number of discrete levels) for each user where the policy can be made without requiring to know the instantaneous CSI. Park and Lim [18] tackle the problem of discrete power allocation and mode selection in device-to-device (D2D) communication networks using DQN with energy efficiency as the reward in the learning process. Amiri et al. [19] apply the cooperative Q-learning for power control at discrete levels, but the effect of channel variations is not considered. Ahmed and Hossain [20] employ deep Q-learning, which was originally proposed in [28], to update the transmission power allocated for each user at the small cell base station in HetNet, in which the power is quantized into discrete levels. Xu et al. [21] present a novel DRL-based technique for resource allocation by considering power efficiency in cloud radio access networks (RANs) and ensuring QoS guarantee. Lu et al. [22] propose a DRL-DPT framework, in which the agent learns directly from the expected objective instead of critic value, for energy efficiency maximization without explicit simulation results for QoS guarantee. Wei et al. determine the number of users and subchannels with corresponding power allocation in HetNets using a policy-gradient based actor–critic learning algorithm [23]. Instead of quantizing the power into discrete levels, Meng et al. [11] model the power as a continuous action and adaptively update the continuous power using the deep deterministic policy gradient-based (DDPG-based) reinforcement learning [29,30]. While the above-mentioned research successfully apply the DRL-based techniques to power allocation in heterogenous networks, the problem of user association is not jointly considered, and various practical constraints in the network are not accounted for, such as the limited backhaul capacity [13] in each small cell base station. A novel energy-efficient joint power allocation and user

association using deep reinforcement learning is studied by Li et al. in [24], where the power is considered as belonging to a discrete set and the learning process is not bounded by any constraints. A summary of the related work is provided in Table 1.

Table 1. Summary of Related Work in Power Allocation and User Association. A check mark "✓" means the issue in the column is considered in the corresponding reference, whereas a cross mark "×" means otherwise.

Ref.	Objective	QoS Constraint	Backhaul Constraint	User Association	Power Allocation	Method
[11]	Data Rate	×	×	×	Continuous	DDPG
[16]	Data Rate	×	×	×	Discrete	Distributed DQN
[18]	Power Efficiency	×	×	×	Discrete	DQN
[19]	Data Rate	✓	×	×	Discrete	DQN
[20]	Data Rate	×	×	×	Discrete	DQN
[21]	Power Efficiency	✓	×	×	×	DQN
[22]	Energy Efficiency	✓	×	×	Continuous	DRL-DPT
[23]	Energy Efficiency	×	×	Not Exactly	Continuous	Actor-Critic
[24]	Energy Efficiency	×	×	✓	Discrete	DQN
This work	Energy Efficiency	✓	✓	✓	Continuous	P-DQN

1.3. Contributions of the Research

In contrast with existing studies, which quantized the continuous set into discrete space [16,19,20], we propose utilizing the parameterized deep Q-network (P-DQN) to handle the problem with a *hybrid* action space composed of *discrete* user association and *continuous* power allocation [31]. This overcomes the difficulty of traditional DQN which can cope with RL problems having discrete action spaces, either with intrinsically discrete actions or with discrete actions quantized from continuous action space. In this work, we provide a joint solution for power allocation and user association with the objective of maximizing downlink energy efficiency under backhaul link constraint and QoS guarantee using P-DQN. A flexible reward function is devised to meet each user equipment's QoS demands in different traffic scenarios, and a penalty mechanism is introduced when the backhaul link constraint is violated. Simulation results demonstrate that the P-DQN outperforms other approaches in terms of overall energy efficiency while satisfying QoS requirements and backhaul constraints. The main contributions of this paper are summarized as follows:

- We provide a joint solution for the power allocation and user association with the objective of maximizing downlink energy efficiency under backhaul link constraint and QoS guarantee. We employ the novel model-free parameterized deep Q-network (P-DQN) framework that is capable of updating policies in a hybrid discrete-continuous action space (i.e., discrete BS-UE association and continuous power allocation).
- To the best of our knowledge, most DRL-based research about power allocation do not consider the wireless backhaul capacity constraint and user QoS. We design the flexible reward function to meet the QoS demands at different traffic scenarios and introduce a penalty mechanism when the backhaul link constraint is violated. We verify by simulations that the proposed P-DQN framework outperforms other proposed

approaches in terms of overall energy efficiency while satisfying QoS requirements and backhaul constraints.

1.4. Organization

The rest of the paper is organized as follows. The system model is described in Section 2. We next present the joint energy-efficiency maximization problem of clustering decision and power allocation in Section 3. Simulation results are given in Section 4. Finally, we provide concluding remarks in Section 5.

2. System Model

2.1. Heterogeneous Network

Consider the downlink of a two-tier HetNet composed of a macro base station (MBS), J small BSs (SBS), and K UEs, with $\mathcal{J} = \{1, 2, \ldots, J\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$ being the sets of SBSs and UEs, respectively. The system network is depicted in Figure 1. In this paper, we assume there is no cross-tier interference in the network, which can be achieved by using different frequency bands for transmissions in the two tiers (e.g., sub-6 GHz in tier 1 and millimeter wave bands in tier 2).

The MBS is equipped with an antenna array of size N_T, which is assumed to be larger than the the number of SBS J, i.e., and $N_T > J$. Orthogonal frequency division multiple access (OFDMA) is utilized for the downlink communication between SBSs and UEs, with a total number of subchannels N_{sub}.

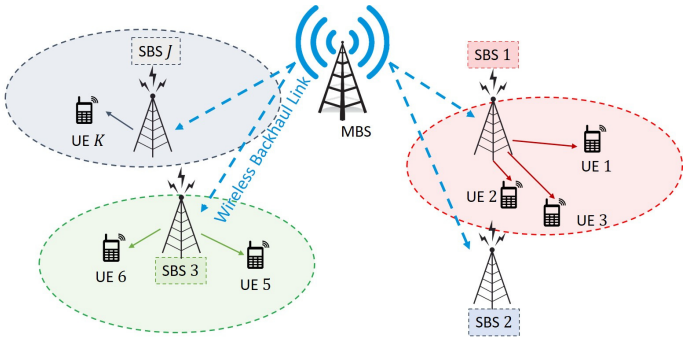


Figure 1. An illustration of the considered wireless network. Each cluster consists of an SBS and its serving UEs.

2.2. User Association

Each UE is assumed to be associated with only one SBS, but each SBS can serve multiple UEs using OFDMA. The UEs served by the same SBS constitute a *cluster*. Let $\mathcal{F}_k$ denote the set of subchannels allocated to the kth UE and $c_{j,k} \in \{0, 1\}$ represent the status of user association, i.e., $c_{j,k} = 1$ if the kth UE is associated with the jth SBS and $c_{j,k} = 0$ otherwise. Then, the set of UEs in the cluster j is given by $\mathcal{C}_j = \{k : c_{j,k} = 1, k \in \mathcal{K}\}$, with $|\mathcal{C}_j|$ being the number of UEs in $\mathcal{C}_j$. The SBS serving the kth UE can be represented by $\mathcal{S}_k = \{j : c_{j,k} = 1, j \in \mathcal{J}\}$. Note that, since each user is assumed to be associated with only one SBS in this paper, $|\mathcal{S}_k|$ equals one. The set of active SBSs is $\mathcal{J}^{active} = \{j \mid |\mathcal{C}_j| > 0\}$.

The spectral efficiency of the kth UE is given by

$$\rho_k = \sum_{f \in \mathcal{F}_k} \log_2\Big(1 + SINR_{k,f}\Big), \tag{1}$$

where the signal-to-noise-plus-noise ratio (SINR) is

$$SINR_{k,f} = \frac{\sum_{j=1}^{J} c_{j,k} g_{j,k,f} P_{j,k,f}}{\sigma^2 + I_{k,f}}$$

with $g_{j,k,f}$ being the channel gain between the jth SBS and the kth UE in subchannel f, $I_{k,f}$ the interference observed by the kth UE, and σ^2 the noise power. Specifically, the channel gain is defined as $g_{j,k,f} = |h_{j,k,f}|^2$, where $h_{j,k,f}$ is the corresponding channel coefficient. The transmit power $P_{j,k,f}$ from SBS j to UE k in subchannel f needs to satisfy the power constraint $0 \leq \sum_{k \in \mathcal{C}_j} \sum_{f \in \mathcal{F}_k} P_{j,k,f} \leq P_{SBS_j,max}$, where $P_{SBS_j,max}$ is the maximum transmit power of the jth SBS. The user sum-rate for the jth cluster is given by

$$\rho_j^{SBS} = \sum_{k \in \mathcal{C}_j} \rho_k = \sum_{k=1}^{K} c_{j,k} \rho_k. \tag{2}$$

We consider the scenario that each SBS allocates orthogonal subchannels to different UEs within its serving coverage, so there is no intra-cluster interference in each cluster. Each UE can acquire at least one subchannel for data transmission if the cluster size (i.e., the number of served UEs) is not larger than the number of subchannels. Without intra-cluster interference, the interference term I_k is composed only by the *inter-cluster* interference and can be expressed by

$$I_{k,f} = \sum_{u \notin \mathcal{C}_{\mathcal{S}_k}} \sum_{f \in \mathcal{F}_k \cap \mathcal{F}_u} g_{j,k,f} P_{j,u,f}. \tag{3}$$

Detailed notation descriptions are summarized in Table 2.

Table 2. Notation summary.

Notation	Definition
$\mathcal{J}, \mathcal{K}$	set of SBSs and set of UEs
$\mathcal{F}_k$	set of subchannels allocated to the kth UE
$\mathcal{S}_k$	the SBS serving the kth UE
B_{sub}	subchannel bandwidth
N_T	MBS antenna array size
P_T	total power consumption at all the SBSs.
$g_{j,k,f}$	channel gain between SBS j and UE k in the fth subchannel
$h_{j,k,f}$	channel coefficient between SBS j and UE k in the fth subchannel
$P_{SBS_j,max}$	maximum power available at the jth SBS
$P_{j,k,f}$	* transmit power from SBS j to UE k in the fth subchannel
p_k^{UE}	** transmit power from the associated SBS to UE k
$\lvert\mathcal{J}^{active}\rvert$	number of active SBSs
σ^2	noise power
$I_{k,f}$	interference experienced by UE k in subchannel f
$\mathcal{C}_j$	the set of UEs in cluster j
$c_{j,k}$	link indicator between SBS j and UE k
c^{UE}	user association
$SINR_{k,f}$	SINR for UE k in the fth subchannel
ρ_k	capacity for UE k
v_k	capacity threshold for UE k
ρ_j^{SBS}	user sum-rate for SBS j
R_j^{SBS}	maximum downlink data rate for SBS j

* For the optimization approach. ** For the DRL-based approach.

2.3. Power Consumption

The system power consumption includes the operational power, which is the minimum amount of power to keep the SBS active, and data transmission power. Operational powers for SBSs and MBS are expressed as $P_{o,SBS}$ and $P_{o,MBS}$, respectively. The total power consumption of all SBSs is

$$P_T = |\mathcal{J}^{active}| \cdot P_{o,SBS} + \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{C}_j} \sum_{f \in \mathcal{F}_k} P_{j,k,f},$$

where $\mid \mathcal{J}^{active} \mid$ is the number of active SBSs.

3. Problem Formulation

3.1. Optimization Problem

We aim for a joint solution optimizing for the user association and transmit power allocation with the objective of maximizing energy efficiency, which is defined as the achievable sum rate per consumed power, in the downlink of the two-tier HetNet while considering QoS guarantee and wireless backhaul link capacity constraints. The problem can be formulated as

$$\max_{\{c_{j,k}\},\{P_{j,k,f}\}} \quad \frac{1}{P_T}\sum_{k=1}^{K}\rho_k \tag{4a}$$

$$\text{subject to} \quad C_1 : \sum_{j} c_{j,k} = 1,\ c_{j,k} \in \{0,1\},\ \forall j \in \mathcal{J},\ k \in \mathcal{K} \tag{4b}$$

$$C_2 : 0 \leqslant \sum_{k\in\mathcal{C}_j}\sum_{f\in\mathcal{F}_k} P_{j,k,f} \leqslant P_{SBS_j,max},\ \forall j \in \mathcal{J},\ k \in \mathcal{K} \tag{4c}$$

$$C_3 : \rho_k \geqslant \nu_k,\ \forall k \in \mathcal{K} \tag{4d}$$

$$C_4 : \mid \mathcal{C}_j \mid \leqslant |\mathcal{C}_j|_{max},\ \forall j \in \mathcal{J} \tag{4e}$$

$$C_5 : \rho_j^{SBS} \leqslant R_j^{SBS},\ \forall j \in \mathcal{J} \tag{4f}$$

C_1 in (4b) assumes that each UE is served by only one SBS, and C_2 in (4c) refers to transmit power limit at the jth SBS with $P_{SBS_j,max}$ the maximum power available at the jth SBS. C_3 in (4d) indicates the QoS requirement for each UE, where ν_k is the capacity threshold for UE k. C_4 in (4e) is the cluster size constraint with $|\mathcal{C}_j|_{max}$ the maximum allowable number of users in $\mathcal{C}_j$. This ensures that UEs in the same cluster are assigned different subchannels to avoid intra-cluster interference. C_5 in (4f) indicates the backhaul link capacity constraint, where R_j^{SBS} is the maximum achievable downlink data rate for SBS j. Note that the subchannel assignment is assumed to be known and is not considered in this work.

The strategy in (4a) attempts to maximize the energy efficiency by finding the optimal user association and power allocation, which is generally a challenging problem with various unknowns and hybrid unknown spaces (continuous power and discrete clustering) in the system. Furthermore, the optimization problem in (4a) deals with a *one-shot* scenario at a certain time instant which needs to be re-evaluated when the network evolves to the next time instant. To tackle the challenges, we are therefore motivated to resort to the techniques of reinforcement learning (RL).

3.2. Reinforcement Learning

RL as one kind of machine learning is well known for its capability of making decisions sequentially in dynamic environments, where the decision-making agent interacts with the environments by an appropriately chosen *action* which is based on its past experiences learned through a *reward* function and on its current environment *state* the agent is experiencing [15]. These constitute the three fundamental elements in an RL: state (s_t), action (a_t), and reward ($r(s_t, a_t)$). Typically, an RL formulates the environment dynamics as a Markov decision process (MDP), and the primary objective is to determine the *action* contingent upon a certain state at each time step such that the expected discount cumulated reward is maximized. More specifically, traditional Q-learning aims at finding the *action* that maximizes the action-value function $Q(s, a)$ defined by

$$Q(s,a) \triangleq E\left[\sum_{k=0}^{\infty}\gamma^k r(s_{t+k}, a_{t+k})\Big| s_t = s, a_t = a\right], \tag{5}$$

where γ is the discount factor.

It is a challenging task to evaluate the Q-function in (5) in many applications, since the statistical properties between any two states are often not easy to obtain. Fortunately, thanks to the advancement in deep neural networks (DNNs), evaluations of the Q-function in (5) can be well approximated by properly designed DNNs [28]. Specifically, during the training phase, the weights in the deep Q learning (DQN) network are trained to extract features from raw data with corresponding target values obtained by the recursive Bellman equation developed from (5). The success of DQN has led to an explosive subsequent development in the area of deep reinforcement learning (DRL), such as the more stabilized version of the DQN (double DQN in [32]), the extension to continuous actions using the deep deterministic policy gradient (DDPG) in [29], and the TD3 [33].

3.3. State, Action, and Reward Function

In this research, the state, action, and reward function associated with the penalty mechanism for the wireless backhaul capacity constraint in the considered RL-based joint power allocation and user association are defined as follows:

- **State**: The state at the t_{th} time slot is defined as the user data rate in that time slot

$$\boldsymbol{s}_t = [\rho_1(t), \ldots, \rho_K(t)]. \tag{6}$$

- **Action**: The action in the t_{th} time slot is defined as

$$\boldsymbol{a}_t = \left[\boldsymbol{c}^{UE}(t),\, \boldsymbol{p}^{UE}(t)\right], \tag{7}$$

  where $\boldsymbol{c}^{UE} = [c_{j,k}]_{j=1:J,k=1:K}$, with $c_{j,k} \in \{0,1\}, j \in \mathcal{J}, k \in \mathcal{K}$, and $\boldsymbol{p}^{\boldsymbol{UE}}(t)$ indicating the sets of user associations and power allocations, respectively. More specifically, the power allocation set is given by $\boldsymbol{p}^{\boldsymbol{UE}}(t) = [p_1^{UE}(n), \ldots, p_K^{UE}(n)]$, where $p_k^{UE}(t) = [P_{\mathcal{S}_k,k,f}(t)]_{f:f\in\mathcal{F}_k}$ is the vector of allocated power for data transmission in all subchannels assigned to UE k from its associated SBS $\mathcal{S}_k$.
- **Reward**: We aim to maximize the overall energy efficiency as in (4a) while maintaining QoS for each US and satisfying the backhaul link capacity constraint for each SBS. Hence, the reward r_t at the t_{th} time slot is defined as

$$r(s_t, a_t) = \begin{cases} r^{'}(s_t, a_t), & \text{if } \rho_j^{SBS} \le R_j^{SBS}, \forall j \in \mathcal{J}, \\ r^{'}(s_t, a_t) - r_{th}, & \text{if } \rho_j^{SBS} > R_j^{SBS}, \text{ for some } j \in \mathcal{J}, \end{cases} \tag{8}$$

  where

$$r^{'}(s_t, a_t) \triangleq \lambda_1 Z_{\kappa_1(t)} - \lambda_2 Z_{\kappa_2(t)},$$

  with $\kappa_1(t) = \frac{1}{P_T}\sum_{k=1}^{K}\rho_k(t)$ being the system energy efficiency and

$$\kappa_2(t) = \sum_{k=1}^{K}(\rho_k(t) - \nu_k)^2$$

  being the penalty term which discourages the agent from taking the actions such that the capacity of each user deviates too much from the QoS threshold, and $Z_{\kappa_1(t)}$ and $Z_{\kappa_2(t)}$ are the Z-scores (i.e., standardized results) of $\kappa_1(t)$ and $\kappa_2(t)$, respectively. r_{th} is a threshold used to reduce the likelihood of violating the backhaul capacity constraint.

One of the challenges when transforming a traditional optimization problem into a DRL problem is to devise proper handling of the constraints in the original optimization problem. In this paper, the penalty term in $\kappa_2(t)$ is designed to improve the QoS satisfaction in constraint C_3 through reducing the number of UEs whose achievable rates are much higher or lower than the capacity thresholds. The weights $\lambda_1, \lambda_2 \in [0,1]$ control the significance of the corresponding term. Operators can tune the weights λ_1 and λ_2 according

to their needs, e.g., setting $\lambda_1 > \lambda_2$ at off-peak traffic periods and $\lambda_1 < \lambda_2$ at peak traffic hours to enhance each user's service experiences. Note that the reward function defined in (8) may not be feasible in practice, and judicious setting for the associate weights λ_1 and λ_2 are needed based on trial-and-error efforts. Furthermore, in order to guide the agent to follow the backhaul link constraint, a penalty mechanism is introduced here for the agent to adjust corresponding actions. If one of the SBSs experiences a sum rate that violates the backhaul capacity constraint, the agent receives a penalty and restarts a new episode in the learning process. On the other hand, in this paper, the cluster size constraint is dealt with by including only the legitimate discrete actions, each of which allows no more than N_{sub} users in each cluster, in the *constrained* discrete action space for the entire learning process. This guarantees the output of the discrete action satisfies the cluster size constraint. Finally, in order to accommodate the power constraint in C_2, techniques mentioned in the above can also be utilized. Alternatively, in this work, modifications have been made to the power constraint such that each user's allocated power is restricted by $\sum_{f \in \mathcal{F}_k} P_{j,k,f} \leq P_{k,max}$, where the per-user maximum power $P_{k,max}$ is assumed to be $P_{k,max} = \frac{1}{|\mathcal{C}_j|} P_{SBS_j,max}$ for $k \in \mathcal{C}_j$. In this case, the total power constraint in C_2 can be satisfied, though in a suboptimum fashion. This per user power constraint can be facilitated by the actor-parameter network in the proposed P-DQN in a much easier way.

3.4. Parameterized Deep Q Network

Recent progress in deep RL (DRL) approaches has made the DRL, such as DQN, a viable technique to tackle various resource allocation problems in wireless networks. However, in order for DQN to be able to solve the joint power allocation and user association problem considered in this work, the continuous action space in the power allocation has to be quantized into discrete action space first. Quantization of the continuous action space may round off potentially optimal power allocations. Moreover, the complexity of the DQN increases exponentially with the dimension of the action space, leading to undesirable huge consumption of power and slowdown of convergence speed. To overcome this difficulty, in this paper, we propose employing the P-DQN [31,34] for the joint power allocation and user association because of its capability of solving problems with hybrid action space.

The parameterized action space is denoted by by $\mathcal{A}^{PA} = \{(c, x_c) | x_c \in \mathcal{A}_c, \text{for all } c \in \mathcal{A}_d\}$, where $\mathcal{A}_c$ and $\mathcal{A}_d$ are the continuous and discrete action spaces, respectively. When the discrete action c takes all possible combinations into consideration without constraints, the discrete action space $\mathcal{A}_d = \{[c_{k,j}] : c_{k,j} \in \{0,1\}, k \in \mathcal{K}, j \in \mathcal{J}\}$. Each discrete action c has a corresponding continuous parameter $x_c \in \mathcal{X}_c$, where $\mathcal{X}_c$ is the set of all users' power allocations $\boldsymbol{p}^{UE}(t)$ in this work, for a discrete action c. The primary network of the P-DQN (without the stabilizing target networks) is presented in Figure 2. The primary network in Figure 2 consists of an actor-parameter network $x_c(s;\theta)$ with weights θ, which maps the state c and each discrete action to its corresponding continuous parameter, and an actor network $Q(s, c, x_c; \omega)$ with weights ω, which evaluates the action-value Q-function, i.e., the long term expected cumulative reward $Q(s, a)$ defined in (5), and the action can be explicitly represented by the 2-tuple $a = (c, x_c)$ to emphasize the hybrid nature in the action. Typically, the weights θ in the actor-parameter network can be determined by maximizing the expected action-value function $E[Q(s, c, x_c(c;\theta);\omega)]$. Furthermore, the weights ω in the actor network can be updated by minimizing the mean-squared error $E[(y_t - Q(s_t, a_t;\omega))^2]$, where y_n is the target value in the network and $a_t = (c(t), x_c(t))$ [28,31].

In order to stabilize the P-DQN, an additional *target network*, combining the original primary network shown in Figure 2, is built to produce the target value y_t needed in the actor network [28]. Furthermore, to expedite the training process, an experience replay buffer $\mathcal{D}$ is implemented to provide random samples for evaluating the means appearing in the loss functions of both the actor network and actor-parameter network. With the

replay buffer, the loss functions for the actor-parameter $x_c(s;\theta)$ and the actor $Q(s,c,x_c;\omega)$ can be obtained by the following sample means:

$$L^x(\theta) = -\frac{1}{N}\sum_{i=1}^{N} Q(s_i, c_i, x_{c_i}(s_i;\theta);\omega)$$

$$L^Q(\omega) = \frac{1}{N}\sum_{i=1}^{N} (y_i - Q(s_i, c_i, x_{c_i}(s_i;\theta);\omega))^2,$$

where $y_i = r_i + \gamma \max_{c'} Q\left(s_{i+1}, c', x_{c'}(s_{i+1};\theta^-);\omega^-\right)$ is evaluated by the *target network* for stability with weights θ^- and ω^-, $(s_i, c_i, x_{c_i}, r_i, s_{i+1}) \in \mathcal{D}$ is sampled from the replay buffer, and N is the size of the mini-batch (i.e., sample size). The weights θ and ω are updated according to

$$\theta \leftarrow \theta - \alpha_{a,p}\nabla_\theta L^x(\theta) \tag{9}$$

$$\omega \leftarrow \omega - \alpha_a \nabla_\omega L^Q(\omega), \tag{10}$$

where $\alpha_{a,p}$ and α_a are the learning rate for the weights in the actor-parameter and actor network, respectively.

At a given state s, the actor-parameter network produces the continuous parameters, which maximize the average $Q(s,c,x_c;\omega)$ for each discrete action c. Then, the actor network determines the action $(c, x_c) = \arg\max_{(c,x_c)} Q(s,c,x_c;\omega)$ after the action-value Q-function has been evaluated with the aid of the target network. In the training phase, the off-policy scheme is implemented where the agent selects the action based on ϵ-greedy policy and generates the Q-target using the greedy policy for exploration. The algorithm of the proposed P-DQN with the target networks is summarized in Algorithm 1.

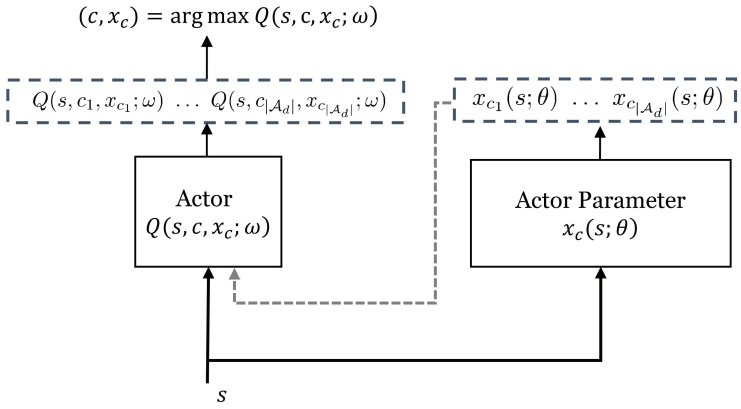


Figure 2. Illustration of the primary network in P-DQN. At each time slot, the actor-parameter $x_c(s;\theta)$ decides the continuous parameter $\boldsymbol{x}(s;\theta) = \left[x_{c_1}(s;\theta), \ldots, x_{c_{|\mathcal{A}_d|}}(s;\theta)\right]^T$ based on the current state s. Then actor $Q(s,c,x_c;\omega)$ takes the action according to the current state s and the continuous parameter $\boldsymbol{x}(s;\theta)$.

Algorithm 1 Parameterized Deep Q-Network (P-DQN) Algorithm with the quasi-static target networks.

Input: Learning rates $\{\alpha_a, \alpha_{a,p}\}$, exploration parameter ϵ, mini-batch size B, a probability distribution ξ.

Initialize network weights ω, ω^-, θ, and θ^-.

for $t = 1, 2, \ldots, T$ **do**

Determine the action parameters $x_c(s_t; \theta_t)$ by the actor-parameter network.

Select action $a_t = (c_t, x_{c_t})$ according to the ϵ-greedy policy:

$$a_t = \begin{cases} \text{a sample from the distribution } \xi, & \text{with probability } \epsilon \\ (c_t, x_{c_t}) : \ c_t = \arg\max\limits_{c \in \mathcal{A}_d} Q(s_t, c, x_c; \omega) \text{by the actor network}, & \text{with probability } 1-\epsilon. \end{cases}$$

Take action a_t , observe reward r_t and the next state s_{t+1}.

Store the experience (s_t, a_t, r_t, s_{t+1}) into $\mathcal{D}$.

Draw N samples of experience (s_i, a_i, r_i, s_{i+1}) randomly from $\mathcal{D}$.

Define the target y_i by

$$y_i = \begin{cases} r_i, & \text{if } s_{i+1} \text{ is the terminal state,} \\ r_i + \gamma \max\limits_{c' \in \mathcal{A}_d} Q\left(s_{i+1}, c', x_{c'}(s_{i+1}; \theta^-); \omega^-\right), & \text{otherwise.} \end{cases}$$

Use (y_i, s_i, a_i) to compute the gradient $\nabla_\omega L^Q(\omega)$ and $\nabla_\theta L^x(\theta)$.

Update the parameters ω, ω^-, θ, θ^-.

end for

4. Simulation Results

4.1. Simulation Setup

In the simulation, a HetNet with three SBSs and five UEs uniformly located in a macrocell with radius 500 m is considered. Backhaul transmission model considered in [35] is adopted in the simulations. The MBS is equipped with 100 antennas and has 20 beamforming groups [35]. Slow Rayleigh fading channels are adopted for simulations where the channel remains unchanged throughout each episode. The Rayleigh channel coefficient is modeled as $h \sim \mathcal{CN}(0,1)$. We also adopt the non-line-of-sight path-loss model for urban MBSs and SBSs [36]. Each subchannel is randomly allocated to a user, and the subchannel allocation is assumed known for the agent. The other settings of the simulation are summarized in Table 3. The Adam optimizer is employed for all DNNs that are embedded in P-DQN. The ϵ-greedy algorithm and Ornstein-Uhlenbeck noise is used for explorations of discrete actions and continuous parameters, respectively. We set the threshold $r_{th} = 0.1$, discount factor $\gamma = 0.95$, batch size $N = 128$, the maximum number of episodes as 2000, and the maximum steps per episode as 100. Other parameter settings used in the P-DQN are given in Table 4. The simulation codes used in this research can be found in [37].

Table 3. Simulation Parameters.

Parameter	Value
Carrier frequency	$f_c = 2$ GHz
Subchannel bandwidth	$B_{sub} = 15$ kHz
Number of subchannels	$N_{sub} = 3$
Number of subchannels per user	$\|\mathcal{F}_k\| = 1$
MBS antenna array size	$N_T = 100$
MBS beamforming group size	$N_g = 20$
The radius of the entire network	500 m
Number of SBS	$J = 3$
Number of UE	$K = 5$
SINR threshold of UE	$\nu_k = 1$ for each UE
Path loss model (a. and b. indicate the model for UE and SBS, respectively)	a. $30.53 + 36.71 \times \log_{10} d_{j,k}$ in dB, $d_{j,k}$ in km b. $19.77 + 3.91 \times \log_{10} d_j$ in dB, d_j in km
Rayleigh channel coefficient	$h \sim \mathcal{CN}(0,1)$
Noise power spectral density	$N_0 = -174$ dBm/Hz
Maximum transmit power of SBS	$P_{SBS_j,max} = 24$ dBm
Maximum cluster size	$\|\mathcal{C}_j\|_{max} = N_{sub} = 3$
Transmit power of MBS	$P_{MBS} = 43$ dBm
Operational power of SBS	$P_{o,SBS} = 0.5$ W
Operational power of MBS	$P_{o,MBS} = 130$ W

Table 4. The settings for the deep neural networks used in the P-DQN.

	Actor	Actor-Parameter
Learning rate	$\alpha_a = 10^{-5}$	$\alpha_{a,p} = 10^{-5}$
Exploration	ϵ-greedy	Ornstein-Uhlenbeck noise
Number of Outputs	$\|\mathcal{A}_d\|$	$K \cdot \|\mathcal{A}_d\|$
Hidden layer	ReLu, 16 ReLu, 128 ReLu, 512	ReLu, 256
Number of Inputs	$K + K \cdot \|\mathcal{A}_d\|$	K

4.2. Performance Analysis

In the simulations, we compare the proposed P-DQN for joint user clustering and power allocation with the following approaches:

- **Nearest SBS + Random Power**: Each UE is associated with the nearest SBS. Random power means that each SBS serves the UEs in its cluster with random powers in a way that the resulting sum rate within the cluster cannot exceed the power and backhaul capacity limit.
- **Best Channel + Random Power**: Each UE is associated with the SBS with the best received signal power, which depends on the UE-SBS distance as well as the small-scale fading effect. Furthermore, each SBS serves the UEs in its cluster with random power allocations under the power and backhaul capacity constraint.

- **P-DQN Association + Random Power**: The user association policy is accomplished by the proposed P-DQN, whereas each SBS serves the UEs in its cluster with random powers under the total power and backhaul capacity constraint.
- **Random Association + P-DQN Power**: Each SBS allocates the power to its serving UEs based on the policy determined by the P-DQN. Each UE is randomly associated with one SBS in such a way that the random association policy obeys the backhaul link constraint.
- **DQN with Five Discrete Power Levels**: The continuous power space is quantized into L non-uniform power intervals, with L discrete power levels $\frac{P_{SBS_j,max}}{10^{\mathcal{L}}}$ with $\mathcal{L} \in \{0, \ldots, L-1\}$. In this simulation, L is set to 4.

We set the weights $(\lambda_1, \lambda_2) = (0.43, 0.16)$ for both the P-DQN with the cluster size constraint and the P-DQN without the cluster size constraint, which allows us to observe the effect of the cluster size constraint in C_4. Figure 3 depicts the average normalized reward versus steps over 500 realizations during the training phase. It shows the convergence of the user association and power allocation algorithm using the proposed P-DQN. Note that, while the P-DQN approach without cluster size constraint provides higher reward as can be seen from Figure 3, it cannot guarantee all UEs' QoS. Figure 4 shows the effectiveness of the proposed P-DQN algorithm, as it outperforms other approaches in terms of both the reward and the energy efficiency. The numerical values of the reward and energy efficiency obtained in Figure 4 and averaged over all time steps are summarized in Table 5. The improvement in the energy efficiency of the proposed P-DQN on average may reach 77.6% over the traditional DQN and 140.6% over the nearest-distance-based association approaches.

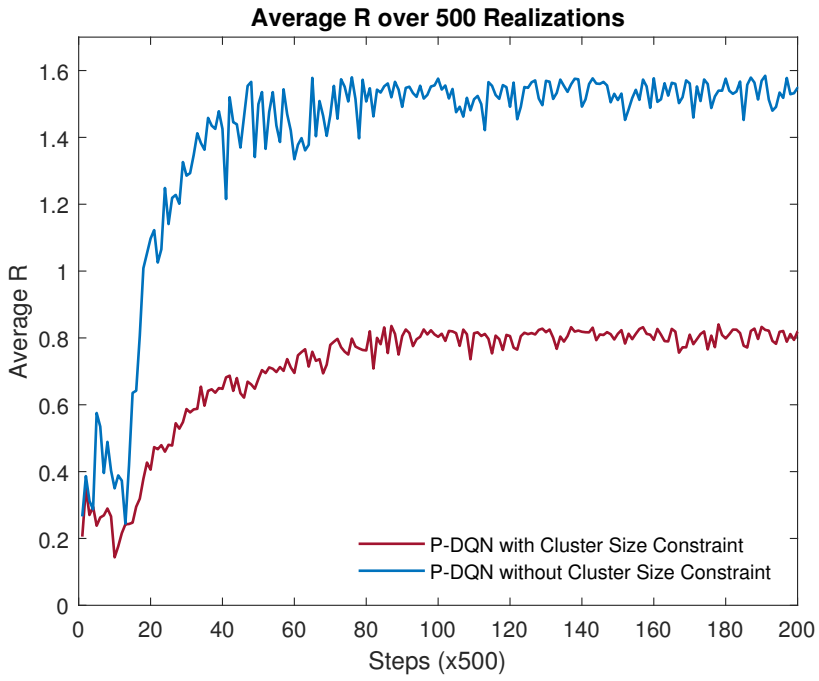
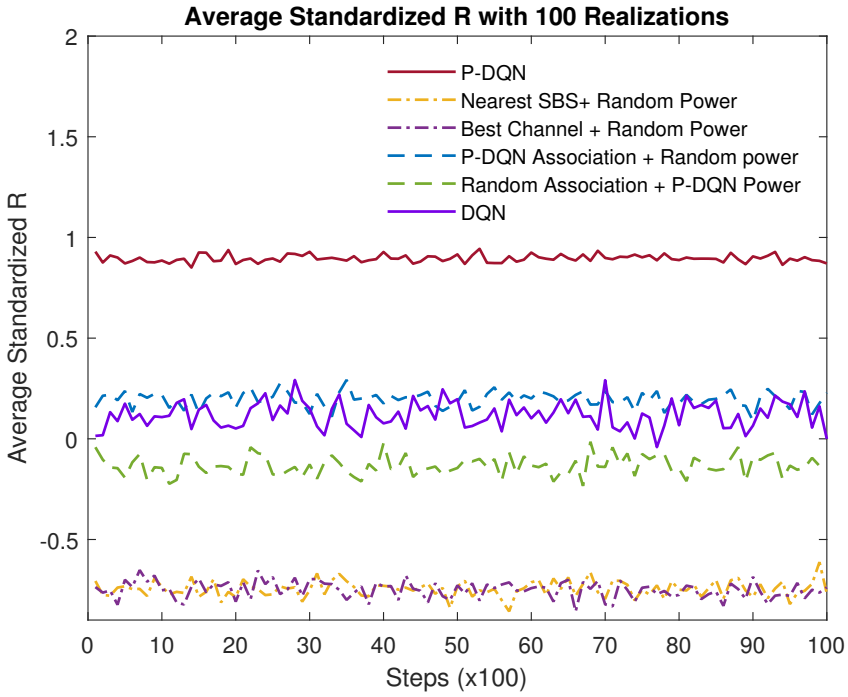
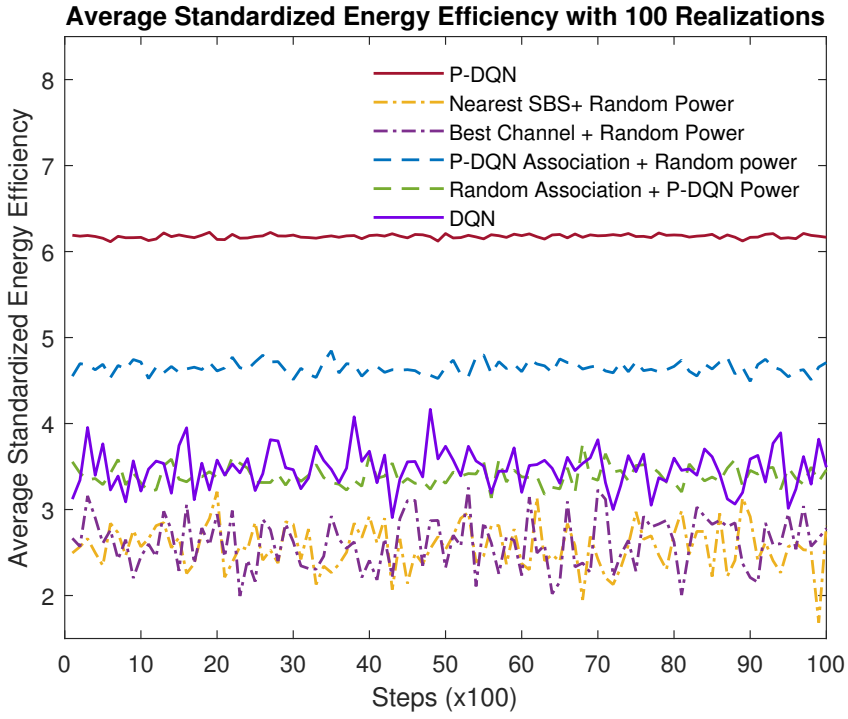


Figure 3. Convergence property of the proposed P-DQN algorithm.

Figure 6c, the P-DQN without the cluster size constraint tends to have all UEs served only by one SBS in the pursuit of small-system power consumption (as large operational power is consumed by any active SBS).



(**a**) Reward.



(**b**) Energy efficiency.

Figure 4. Performance of the proposed reward function. (**a**) Reward. (**b**) Energy efficiency.

Table 5. The average standardized value of the reward and the energy efficiency, obtained from the results in Figure 4 and averaged over all time steps. The method index in the table is: (A) P-DQN with Cluster Size Constraint, (B) Nearest SBS + Random Power, (C) Best Channel + Random Power, (D) P-DQN Association + Random Power, (E) Random Association + P-DQN Power, and (F) DQN with Cluster Size Constraint.

Method Index	Average Reward	Average Energy Efficiency
(A)	0.8957	6.1770
(B)	−0.7453	2.5677
(C)	−0.7499	2.6169
(D)	0.1968	4.6477
(E)	−0.1327	3.3923
(F)	0.1186	3.4773

Table 6 shows that each UE obtains 83.47% of the required QoS, and each SBS obeys the wireless backhaul capacity constraint at each timeslot through the proposed P-DQN with the cluster size constraint. However, the percentage of QoS satisfaction for the methods with random power or random association is about 50%. The percentage of QoS satisfaction of the proposed P-DQN with the cluster size constraint is higher than that of the other approaches, as a penalty term is used in the reward function to realize each UE's QoS as much as possible in the proposed P-DQN. For Random Association + P-DQN Association, the ratio is even smaller than 5%. For DQN with cluster size constraint, the ratio is even smaller than 1%, as the quantized power levels suffer from round-off imperfections, while the continuous action learned by the P-DQN allows for a better QoS with appropriate continuous power allocation.

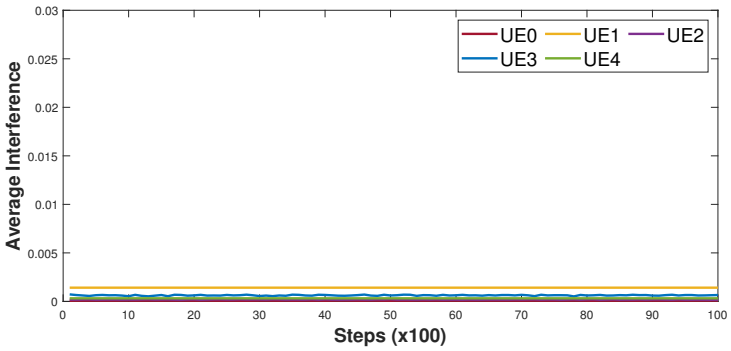
Table 6. The percentage of wireless backhaul link constraint satisfaction and QoS satisfaction. The method index in the table is as follows: (A) P-DQN with Cluster Size Constraint, (B) Nearest SBS + Random Power, (C) Best Channel + Random Power, (D) P-DQN Association + Random Power, (E) Random Association + P-DQN Power, and (F) DQN with Cluster Size Constraint.

Method Index	Percentage of Backhaul Constraint Satisfaction	Percentage of QoS Satisfaction
(A)	100.00%	83.47%
(B)	100.00%	44.91%
(C)	100.00%	44.42%
(D)	100.00%	50.22%
(E)	100.00%	4.87%
(F)	100.00%	0.58%

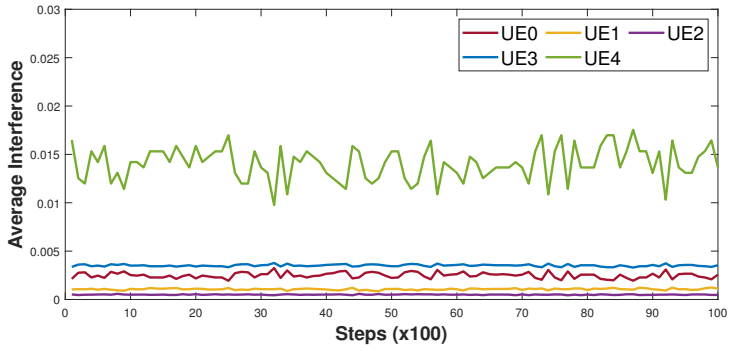
In Figure 5, we analyze the interference power experienced by each UE under various approaches. The results in Figure 5a,c,e demonstrate that the interference each user experiences in the "Nearest SBS + Random Power" approach is higher than that in the "P-DQN with Cluster Size Constraint" approach and "P-DQN Association + Random Power" approach. By comparing the results in Figure 5a,e, it can be seen that random power allocation does not have much impact on each user's received interference level. On the other hand, by comparing the results in Figure 5c,e, we observe a noticeably increased interference level in UE 3 and UE 4, which implies that the interference is largely dominated by the result of user association. The results here indicate that the user association policy learned by the proposed P-DQN with cluster size constraint generally tends to determine the matching between UEs and SBSs such that the inter-cluster interference can be managed to a lower level, as shown in Figure 5a, which results in a higher system throughput.

The increased level of interference in UEs (such as UE 3 and UE 4) observed in Figure 5c can be explained with the aid of Figure 6, which plots the locations of each UE and SBS with their association status being specified by colors. The association results and physical distances between each user and all SBSs in Figure 6 can provide insights into the interference level each UE experiences under different association strategies. For example, the increased interference in UE 4 in Figure 5c can be analyzed by comparing Figure 6b with Figure 6a. More specifically, while UE 4 in Figure 6a associated with SBS 2 under the proposed P-DQN with cluster size constraint is interfered only by inter-cluster signals transmitted from SBS 1, this UE 4 associated with SBS 2 in Figure 6b under the "Nearest SBS" (i.e., nearest distance) association approach can potentially be interfered with by inter-cluster signals from SBS 0 and SBS 1. It can be seen from Figure 6b that SBS 0 could strongly interfere with UE 4, due to the short distance between SBS 0 and UE 4, thus leading to the increased interference level in UE 4 as shown in Figure 6b. Finally, as for the case of P-DQN approach without cluster size constraint, the system suffers from intra-cluster interference, which significantly impacts the interference level in each user as shown in Figure 5b. Since keeping an SBS active demands huge operational power, we can see from Figure 6c that the P-DQN without a cluster size constraint tends to have all UEs served only by one SBS in pursuit of less overall system power consumption.

In contrast with the user association schemes based on the distance or the channel quality between a UE and an SBS, the P-DQN-based user association tends to activate fewer SBSs, which leads to less consumption of overall operational power in the SBSs and results in a higher energy efficiency. More specifically, as illustrated in Figure 6a, where different colors refers to different clusters, we see thatUE 0, UE 2 and UE 3 are associated with SBS 1, and UE 1 and UE 4 are connected to SBS 2, whereas SBS 0 is not active when employing the policy learned by P-DQN with the cluster size constraint.

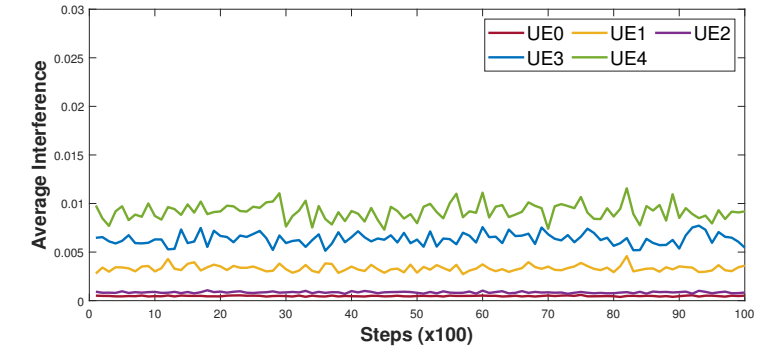


(**a**) P-DQN with Cluster Size Constraint.

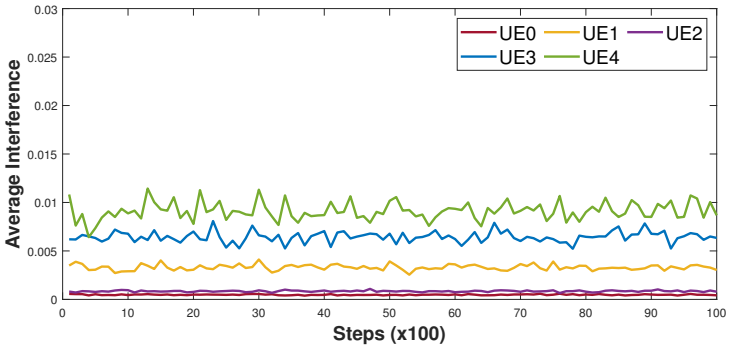


(**b**) P-DQN without Cluster Size Constraint

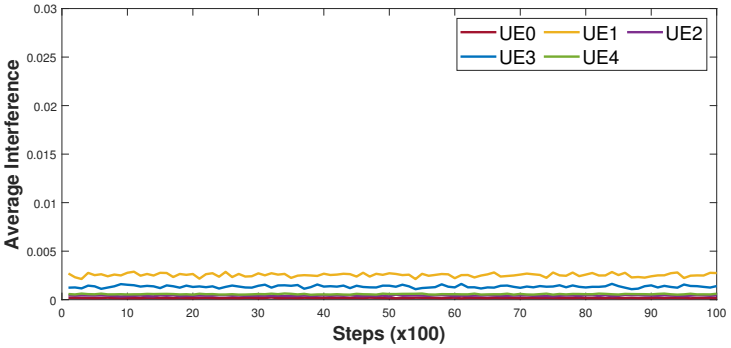
Figure 5. *Cont.*



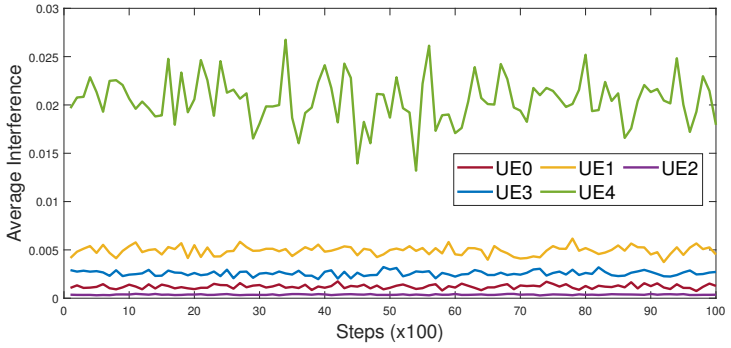
(**c**) Nearest SBS + Random Power.



(**d**) Best Channel +Random Power.

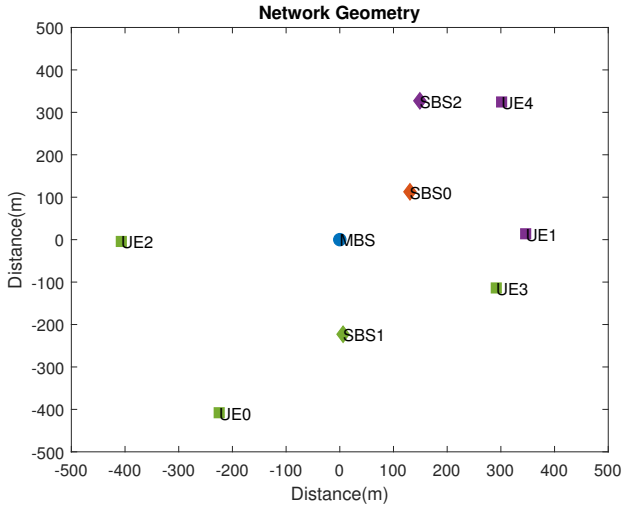


(**e**) P-DQN Association + Random Power.

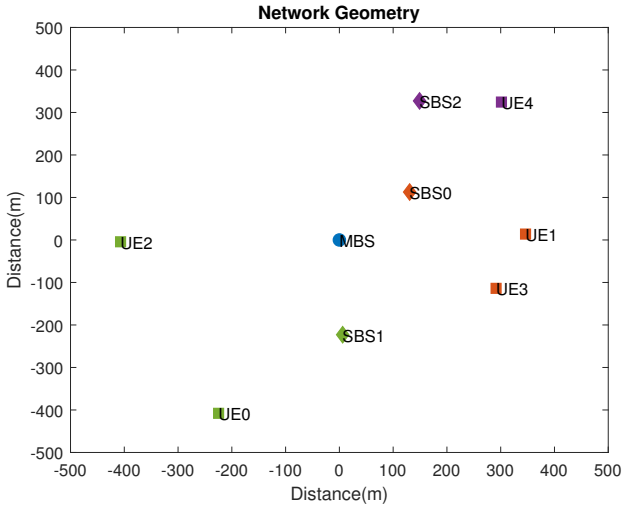


(**f**) Ramdom Association + P-DQN Power.

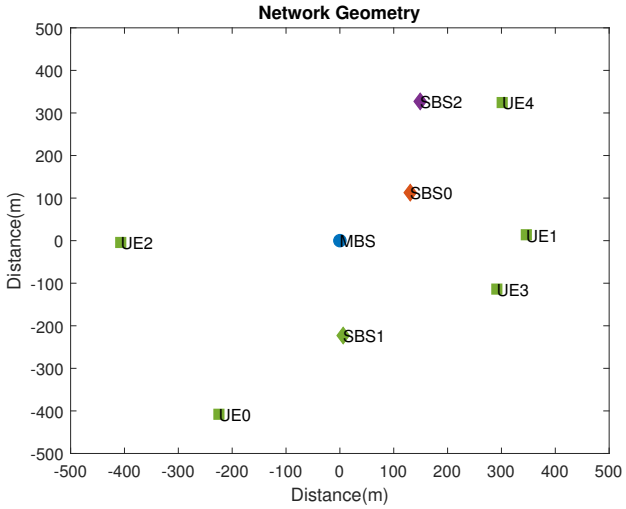
Figure 5. Average interference level of all power allocation and user association approaches. P-DQN association and P-DQN power refer to those of "P-DQN with Cluster Size Constraint".



(**a**) P-DQN with Cluster Size Constraint.



(**b**) The nearest SBS or the best channel policy.



(**c**) P-DQN without Cluster Size Constraint.

Figure 6. The results of user association. (**a**) P-DQN with cluster size constraint. (**b**) The nearest SBS policy. (**c**) P-DQN without cluster size constraint. Each SBS and its associated UEs are shown in the same color.

5. Conclusions

In this paper, we have studied the joint problem of user association and power allocation using P-DQN in the downlink of a two-tier HetNet without knowledge of the environment transition probability. The wireless network has been formulated as a parameterized action Markov decision process with a hybrid (discrete-continuous) action space. The P-DQN has been adopted as a model-free framework to avoid quantization noise resulting from rounding the continuous power space into discrete levels. With the consideration of realistic scenarios, we have designed the reward function as the energy efficiency with QoS constraint per user as well as backhaul capacity constraint. We have introduced a penalty mechanism when the constraints are violated. We have also utilized the cluster size constraint for intra-cluster interference mitigation. In simulations, the proposed P-DQN has been verified to outperform other traditional methods in terms of overall energy efficiency while satisfying QoS requirements and backhaul constraints. The improvement in the energy efficiency of the proposed P-DQN on average may reach 77.6% over the traditional DQN, both with the cluster size constraint. Meanwhile, the proposed P-DQN may still suffer from the curse of dimensionality when dealing with problems with sizable action spaces. It will be worthwhile to investigate advanced DRL techniques (such as the DDPG technique or the multi-agent RL), in future work, capable of handling the problems of joint user association and power allocation, which typically have large action spaces in scenarios of practical interest.

Author Contributions: Conceptualization, F.-T.C.; Investigation, C.-K.H. and F.-T.C.; Methodology, C.-K.H. and F.-T.C.; Resources, F.-T.C.; Writing—original draft, C.-K.H. and K.-L.C.; Software, C.-K.H. and K.-L.C.; Funding acquisition, F.-T.C.; Writing—review and editing, F.-T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology (MOST) in Taiwan under grant number MOST 109-2221-E-009-101.

Institutional Review Board Statement: Not applicable.

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

References

1. Cisco. Cisco visual networking index: Global mobile data trafc forecast update, 2017–2022 white paper. *Update* **2019**.
2. Hu, R.Q.; Qian, Y. An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems. *IEEE Commun. Mag.* **2014**, *52*, 94–101. [CrossRef]
3. Mili, M.R.; Hamdi, K.A.H.; Marvasti, F.; Bennis, M. Joint optimization for optimal power allocation in OFDMA femtocell networks. *IEEE Commun. Lett.* **2016**, *20*, 133–136. [CrossRef]
4. Sambo, Y.A.; Shakir, M.Z.; Qaraqe, K.A.; Serpedin, E.; Imran, M.A. Expanding cellular coverage via cell-edge deployment in heterogeneous networks: Spectral efficiency and backhaul power consumption perspectives. *IEEE Commun. Mag.* **2014**, *52*, 140–149. [CrossRef]
5. Amiri, R.; Almasi, M.A.; Andrews, J.G.; Mehrpouyan, H. Reinforcement learning for self organization and power control of two tier heterogeneous networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 3933–3947. [CrossRef]
6. C., J.; Zhang, H.; Ren, Y.; Han, Z.; Chen, K.C.; Hanzo, L. Machine learning paradigms for next-generation wireless networks. *IEEE Trans. Wirel. Commun.* **2017**, *24*, 98–105. [CrossRef]
7. Zhang, C.; Patras, P.; Haddadi, H. Deep learning in mobile and wireless networking: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2224–2287 [CrossRef]
8. Deng, D.; Li, X.; Zhao, M.; Rabie, K.M.; Rupak, K. Deep learning-based secure MIMO communications with imperfect CSI for heterogeneous networks. *Sensors* **2020**, *20*, 1730. [CrossRef] [PubMed]
9. Liu, S.; He, J.; Wu, J. Dynamic cooperative spectrum sensing based on deep multi-user reinforcement learning. *Appl. Sci.* **2021**, *11*, 1884. [CrossRef]
10. Munaye, Y.Y.; Juang, R.T.; Lin, H.P.; Tarekegn, G.B.; Lin, D.B. Deep Reinforcement Learning Based Resource Management in UAV-Assisted IoT Networks. *Appl. Sci.* **2021**, *11*, 2163. [CrossRef]
11. Meng, F.; Chen, P.; Wu, L.; Cheng, J. Power allocation in multi-user cellular networks: Deep reinforcement learning approaches. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6255–6267. [CrossRef]
12. Tam, H.H.M.; Tuan, H.D.; Ngo, D.T.; Duong, T.Q.; Poor, H.V. Joint load balancing and interference management for small-cell heterogeneous networks with limited backhaul capacity. *IEEE Trans. Wirel. Commun.* **2016**, *16*, 872–884. [CrossRef]

13. Ma, H.; Zhang, H.; Wang, X.; Cheng, J. Backhaul-aware user association and resource allocation for massive mimo-enabled HetNets. *IEEE Commun. Lett.* **2017**, *21*, 2710–2713. [CrossRef]
14. Zhang, H.; Huang, S.; Jiang, C.; Long, K.; Leung, V.C.M.; Poor, H.V. Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations. *IEEE J. Select. Areas Commun.* **2017**, *35*, 1936–1947. [CrossRef]
15. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
16. Nasir, Y.S.; Guo, D. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE J. Select. Areas Commun.* **2019**, *37*, 2239–2250. [CrossRef]
17. He, C.; Hu, Y.; Chen, Y. Joint power allocation and channel assignment for NOMA with deep reinforcement learning. *IEEE J. Select. Areas Commun.* **2019**, *37*, 2200–2210. [CrossRef]
18. Park, H.; Lim, Y. Reinforcement learning for energy optimization with 5G communications in vehicular social networks. *Sensors* **2020**, *20*, 2361. [CrossRef]
19. 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.
20. Ahmed, K.I.; Hossain, E. A deep Q-learning method for downlink power allocation in multi-cell networks. *arXiv* **2019**, arXiv:1904.13032.
21. Xu, Z.; Wang, Y.; Tang, J.; Wang, J.; Gursoy, M.C. A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017, pp. 1–6.
22. Lu, Y.; Lu, H.; Cao, L.; Wu, F.; Zhu, D. Learning deterministic policy with target for power control in wireless networks. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–7.
23. 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]
24. Li, D.; Zhang, H.; Long, K.; Wei, H.; Dong, J.; Nallanathan, A. User association and power allocation based on Q-learning in ultra dense heterogeneous networks. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–7.
25. Liberati, F.; Giuseppi, A.; Pietrabissa, A.; Suraci, V.; Giorgio, A.D.; Trubian, M.; Dietrich, D.; Papadimitriou, P.; Priscoli, F.D. Stochastic and exact methods for service mapping ualized network infrastructures. *Int. J. Netw. Manag.* **2017**, *27*, 872–884. [CrossRef]
26. Gao, J.; Zhong, C.; Chen, X.; Lin, H.; Zhang, Z. Deep Reinforcement Learning for Joint Beamwidth and Power Optimization in mmWave Systems. *IEEE Commun. Lett.* **2020**, *24*, 2201–2205. [CrossRef]
27. Zhang, L.; Tan, J.; Liang, Y.C.; Feng, G.; Niyato, D. Deep Reinforcement Learning-Based Modulation and Coding Scheme Selection in Cognitive Heterogeneous Networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 3281–3294. [CrossRef]
28. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
29. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
30. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic policy gradient algorithms. In Proceedings of the 31st ICML, Beijing, China, 21–26 June 2014; pp. 387–395.
31. Hausknecht, M.; Stone, P. Deep reinforcement learning in parameterized action space. *arXiv* **2015**, arXiv:1511.04143.
32. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
33. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
34. 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.
35. Wang, N.; Hossain, E.; Bhargava, V.K. Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 3251–3268. [CrossRef]
36. 3rd Generation Partnership Project (3GPP). *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*; 3rd Generation Partnership Project (3GPP): Sofia Technology Park, France, 2016.
37. Simulator. Available online: https://github.com/chikaihsieh/Power-Allocation-and-User-Device-Association-with-Deep-Reinforcement-Learning/tree/main (accessed on 27 April 2021).