Physical Communication 61 (2023) 102206

Contents lists available at ScienceDirect

Physical Communication

journal homepage: www.elsevier.com/locate/phycom

Full length article

Energy-efficient joint resource allocation in 5G HetNet using Multi-Agent Parameterized Deep Reinforcement learning☆

Amna Mughees [a], Mohammad Tahir [b,*], Muhammad Aman Sheikh [c], Angela Amphawan [a], Yap Kian Meng [a], Abdul Ahad [d,e], Kazem Chamran [f]

[a] Department of Computing and Information Systems, Sunway University, Selangor 47500, Malaysia
[b] Department of Computing, University of Turku, Turku 20014, Finland
[c] School of Technologies, Cardiff Metropolitan University, Cardiff CF5 2YB, UK
[d] Department of Electronics & Communication Engineering, Istanbul Technical University, Istanbul 34469, Turkey
[e] School of Software, Northwestern Polytechnical University, Xian, Shaanxi 710072, P.R. China
[f] Research Management Centre, City University, Petaling Jaya 46100, Malaysia

ARTICLE INFO

Keywords:
5G
Energy efficiency
Ultra-dense network
DRL
Resource allocation machine learning
HetNet
User association
Power allocation
Reinforcement learning

ABSTRACT

Small cells are a promising technique to improve the capacity and throughput of future wireless networks. However, user association and power allocation in heterogeneous networks is complicated by the dense deployment of small cells, resulting in non-convex and combinatorial problems. Conventionally, machine learning techniques are applied to the joint optimization problem, which has different action spaces. Gauging the continuous spaces to discrete spaces results in the loss of granularity due to discretization (e.g. potential power values in power allocation). Due to its hybrid action space, it is sub-optimal to solve joint user association (discrete spaces) and power allocation (continuous spaces) problems by applying traditional machine learning approaches. This work proposes a Multi-Agent Parameterized Deep Reinforcement Learning (MA-PDRL) approach to address the joint user association and power allocation problem efficiently. According to simulation results, the proposed multi-agent PDRL performs better in energy efficiency and QoS satisfaction than WMMSE, game theory, Q-learning, and DRL techniques.

1. Introduction

The growth in wireless communication due to pervasive access to digital services and bandwidth-intensive applications results in massive data traffic and capacity demands. In the pursuit of delivering enhanced data rates, minimal latency, and accommodating a multitude of connected devices, 5G network aims toward diverse use cases which includes Enhance Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URRLC) and Massive Machine Type Communications (mMTC). 5G leverage dynamic spectral resource utilization augmented through advanced architectural design approach [1]. In order to achieve extended coverage and facilitate network offloading, network capacity expansion emerged as a pivotal approach.

Heterogeneous networks (HetNets) have been considered an optimal approach to increase the spectral efficiency and network capacity expansion, as it shifts the load of macro base stations (MBSs) to small cell base stations (SBSs) which cover a smaller geographical location [2]. The reduction of cell radius to accommodate more cells within a given area having diverse transmission power yields benefits such as enhanced coverage, improved indoor connectivity, load distribution, and deployment flexibility. However, the coexistence of MBSs and SBSs in HetNets can lead to resource management and allocation, interference, security, quality of service (QoS), scalability, and energy efficiency issues.

Energy efficiency is crucial for future green communication as adding more BSs may aggravate a proportional increase in power consumption and also result in insufficient resource allocation contributing to CAPEX/OPEX [3]. Approximately, ~30% of the infrastructure energy consumption can be addressed using various approaches such as intelligent management of network resources, energy harvesting, hardware solutions, and unconventional network architecture. Hardware upgrades and network architecture changes can be costly, whereas efficient resource allocation and management solutions tend to be less expensive. Therefore, addressing energy consumption issues by deploying efficient resource allocation strategies is not only cost-efficient but also flexible and adaptable to changing network conditions.

☆ The research is supported by the Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme (FRGS/1/2021/ICT02/SYUC/02/1).
* Corresponding author.
E-mail address: tahir.mohammad@utu.fi (M. Tahir).

https://doi.org/10.1016/j.phycom.2023.102206
Received 27 July 2023; Received in revised form 17 September 2023; Accepted 4 October 2023
Available online 12 October 2023

Resource allocation strategies ensure the required QoS by efficiently assigning the resources among users/network nodes. Efficient resource allocation, such as power allocation with user association, can significantly reduce interference and energy-related issues. User association is important before data transmission, regardless of the adopted user association technique. Efficient distribution of users among different BSs assists in optimizing network resources, delivering high-quality user experience, load balancing, and improving energy efficiency. Also, allocating power optimally enhances network performance, mitigates interference, and enhances energy efficiency. The resource allocation schemes are critically associated with QoS and energy efficiency, where optimization of one parameter can result in the degradation of other parameters. Therefore, joint optimization of user association and power allocation is important to enhance the energy-efficiency.

Over the decades, the power allocation problem has been evaluated on model-based optimization techniques, making it easy to manage and evaluate the set of data because of its mathematical modeling; however, the computational complexity is relatively high [4]. Unlike model-based power allocation schemes, data-driven based power allocation schemes contribute to improved performance with efficient computational complexity. Keeping in view the dense unknown environment of wireless networks, reinforcement learning (RL) is considered an optimal solution due to its ability to learn and the use of reward function [5]. However, it is challenging to apply in the case of a dense wireless network with diverse users and BSs distribution with dynamic transmission power. Such large-scale networks result in infinite states and action spaces, making RL suffer with the large state space problem [6]. The power control problem in the dense network requires continuous allocation, which makes Q-learning infeasible as it suffers from slow convergence issues.

Apart from independent allocation problems, joint resource optimization problems such as user association, power allocation, subchannel allocation, and bandwidth allocation have been widely studied. Usually, the resource allocation relates to different action spaces. As the action spaces in joint resource optimization problems are interconnected. The decisions made in one aspect may directly influence decisions in another. Hence, managing hybrid action spaces jointly in their domain is a challenging task. For example, user association and power allocation have discrete and continuous space, respectively and decisions need to be made collectively to ensure optimal performance. Therefore, to address this issue, this paper proposes a joint resource optimization solution (user association + power allocation) by utilizing deep reinforcement learning (DRL) in a hybrid action space to enhance energy efficiency by learning from interactions with the environment.

1.1. Related work

In [7], a non-cooperative game theory problem based on two subgames is proposed to address the joint user association and resource allocation problem. The joint solution is focused on a user-centric approach and considered under QoS constraint and maximized UE rate. Such user-centric approaches are highly sensitive to user association changes, causing difficulties in maintaining a stable and energy-efficient network. This approach also requires access to complete network information, which is challenging to achieve an optimal solution. A joint Poisson point process and Poisson hole process-based optimization algorithms are proposed in [8]. The coverage probability is derived by the mean interference-to-signal ratio, and the repulsion radius technique is used to minimize the interference. However, the Poisson point process considers the nodes independent of each other and repulsion against nodes, making it not suitable for interference management techniques. The relationship between the coverage probability and the average achievable rate is used to derive the EE expression.

Q-learning is another technique to solve the optimization problem in a dense wireless network dependent on reward function. The work in [9,10] proposed distributed resource allocation among femtocells based on cooperative Q-learning. The method lacks the ability to address the continuous action space problem of power allocation. A multi-agent Q-learning approach is proposed in [11] to maximize the energy efficiency in the ultra-dense wireless network. An optimization problem is modeled based on user association and power allocation, where an agent interacts with the environment as a learning process and saves all information in Q-table. However, the Q-table size grows exponentially, creating a challenge to handle dense real-world problems. The energy efficiency trade-off problem is also studied in energy harvesting support systems in [12–14]. In [6], another framework is proposed to improve energy efficiency for Inter-Cell Interference Coordination (ICIC) in wireless networks. Considering the continuous action and state space, Deep Reinforcement Learning with Deterministic Policy and Target (DRL-DPT) is proposed to improve energy efficiency under different ICIC schemes.

To address the increased capacity and energy issues, a deep reinforcement based algorithm is proposed for joint power control and resource allocation in [15] to optimize the EE and the SE. Another work in [16], targets the distributed resource management problem in two-tier HetNet using deep reinforcement learning techniques. The Deep Neural Network (DNN) is used to learn optimal resource allocation policies, and experimental results show the effectiveness of the proposed approach in improving spectrum and power allocation and satisfying QoS. However, the proposed work is focused on locally-observed information, which could potentially lead to decentralized and insufficient solutions in dense HetNets. Another problem of hybrid action space is considered for joint user association and power allocation in [17]. A parameterized deep Q-network (P-DQN) is employed to enhance energy efficiency while considering backhaul capacity constraints and satisfying QoS constraints. Another multi-agent hybrid resource allocation algorithm is proposed in [19], based on a stochastic game where the agents make decisions sequentially. The sequential decisions sometimes may lead to the significant limitation of randomness and uncertainty in the outcome. A RL-based [20] and policy gradient-based actor-critic [18] RL policy is investigated for user association and user scheduling for resource allocation, respectively. In [13,21,22], the energy-efficient resource allocation strategies are further explored, considering the application and connectivity between different entities. The Table 1 presents a comprehensive overview of the related research articles, drawing insightful comparisons with the proposed work.

1.2. Main contributions

Achieving energy efficiency and supporting high data rates in dense scale, versatile, and complex wireless networks is quite challenging [23]. Efficient and advanced resource management and allocation solutions must be deployed to achieve a balance between enhancing network performance and efficiently utilizing network resources. Several optimization techniques are discussed above, but the heterogeneous nature of wireless networks raises several challenges for conventional optimization techniques.

Conventional techniques such as optimization, heuristic, and game theory algorithms require entire or quasi-complete knowledge of wireless networks such as real-time channel state information (CSI) and channel models. Furthermore, conventional techniques are computationally expensive and require considerable overhead processing time. Game theory techniques are more suitable approach for a homogeneous set of players as gathering comprehensive data and unique equilibrium solutions on heterogeneous players can be difficult and may require significant resources. However, HetNets creates an extra burden due to multiple BSs and UEs. Such a huge number of devices creates large overhead that increases delay, memory, and energy consumption of network elements.

Apart from conventional techniques, Deep Reinforcement Learning (DRL) based techniques are used to solve complex network resource

Table 1
Literature comparison.

Reference	Objective	Method	User association	Power allocation	QoS constraint	Interference
[10]	Downlink utility & QoS	Multi-agent DRL	✓	×	✓	×
[6]	Energy Efficiency	DRL-DPT	×	Continuous	✓	×
[7]	Throughput	Game Theory	✓	Non-cooperative	✓	✓
[8]	QoS	Point Hole Process	✓	✓	✓	✓
[9]	Sum capacity	Q-learning	×	Distributed	✓	×
[11]	Energy Efficiency	DQN	✓	Discrete	✓	×
[15]	Energy Efficiency	DRL	✓	Distributed	✓	×
[16]	Data rate	DNN	×	✓	✓	×
[17]	Energy Efficiency	PDQN	✓	✓	✓	×
[18]	Energy Efficiency	Actor-Critic RL	×	Continuous	×	×
Proposed work	Energy Efficiency	Multi-agent PDRL	✓	Continuous	✓	✓

management and allocation problems while utilizing limited information about the network. DRL techniques are model-free and have the ability to learn through interactions with the network without knowing the exact channel models or other network statistics [24]. DRL algorithms can be used as value-based, policy-based, and model-based algorithms in different action spaces. The focus of most of the research work is on the DRL algorithm families, and the importance of action space is neglected.

From an RL point of view, User association and power allocation comprise discrete action space and continuous action space, respectively. Although much work has been done on the joint solution to these optimization problems, the approach of prior discretization is used. In which the continuous action spaces are first quantized to a finite set of discrete values. Quantifying continuous action spaces to discrete action spaces in DRL problems is challenging [25]. This discretization process of action spaces makes the learning process slow and makes it difficult to learn from past experiences [26]. Also, this quantization leads to high computational complexity and loss of information (such as possible power levels in power allocation) due to conversion.

To overcome these challenges, this work proposes a solution to address the practical and important issue of joint user association and power allocation using Multi-Agent Deep Q-Networks (MA-DQN) in a parameterized action space while considering the constraints of the joint action spaces. The proposed energy-efficient framework also considers the QoS constraint that provides better energy efficiency and data rate all over the network. Furthermore, due to the dynamic nature of wireless networks, reaching the maximum reward in the case of single-agent DRL is challenging. Therefore, multi-agent DRL is used as it explicitly prototypes agents and their actions on the network.

This paper makes the following distinct contributions:

- Formulate the joint optimization problem of user association and power allocation that considers the QoS constraint to enhance energy efficiency.
- The formulated problem is explored in terms of multi-agent, where every agent will be evaluated further in terms of their action spaces, rewards, and penalty.
- A Multi-agent Parameterized Deep Reinforcement Learning (MA-PDQN) framework is deployed as a solution to hybrid action space to handle the dense small-cell wireless network heterogeneity and large dimensions that maximize the reward of agents.

1.3. Organization

The rest of the paper is structured as follows: Section 2 defines the system model and problem formulation. In Section 3, the proposed algorithm and its resulting solution are presented. Simulation results are presented in Section 5, and Section 6 concludes the paper.

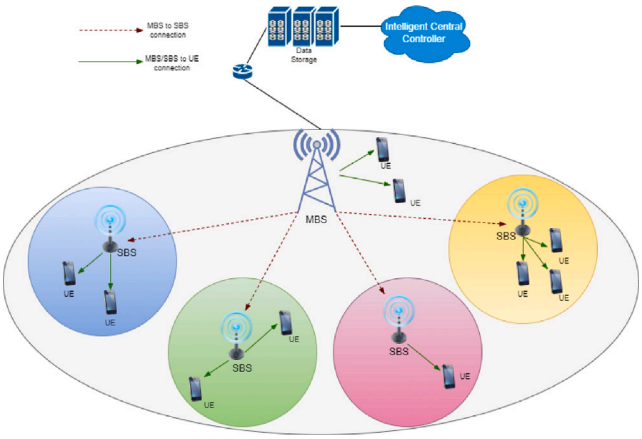

Fig. 1. System model of MBS-SBS architecture based on 5G.

2. System model and problem formulation

The system model for the proposed joint energy-efficient user association and power allocation optimization problem is shown in Fig. 1.

2.1. System model

A small cell network has been considered consisting of one MBS[1] and multiple SBSs as shown in Fig. 1. SBSs are deployed to shift the traffic from MBS as SBSs consume less power, provide small coverage and better QoS. The network is composed of M MBS, F SBS, and U UE with the sets defined below, where B is the set containing all MBS and SBSs.

$$B = \{b_1, b_2, \ldots, b_B\}, \forall M \cup F \in B,$$
$$F = \{f_1, f_2, \ldots, f_F\},$$
$$U = \{u_1, u_2, \ldots, u_U\}.$$

The MBSs are considered to be equipped with an antenna array of size N_T where the total number of sub-channels is denoted by N_{sub}. A fixed bandwidth allocation among each sub-channel for all links is considered. It is assumed that each UE can associate with only one BS at a time, and each UE can also access full channel state information (CSI). The association variable is denoted by $x_{f,u}$. The association strategy between user u and SBS f is expressed as:

$$x_{f,u} = \begin{cases} 1, & \text{if the } u\text{th UE is associated with } f\text{th BS} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

[1] This system model is for single MBS, and can be extended further for multiple MBSs.

Considering different BS serving several different UEs, this binary user association between single BS and multiple UE can be considered as a cluster, where $X_u^f(t) = x_u^1(t), x_u^2(t), \ldots, x_u^{f_F}(t)$ where $f \in F, u \in U$. Thus, the cluster consisting of UE u under SBS f is given by the following equation, where $\mathfrak{R}$ and $\mathfrak{I}$ refer to data rate and sub-channel, respectively.

$$x_{f,u} = 1 \begin{cases} |\mathfrak{R}_u|, \forall u \in U & \text{no. of UE in cluster } l \\ |\mathfrak{I}_f|, \forall f \in F & f \text{ SBS serving } u\text{th UE} \end{cases} \tag{2}$$

The signal-to-interference-ratio (SINR) at MBS UE (MUE) is expressed as:

$$\gamma_{m,u_i,c} = \frac{p_{m_i,u} g_{m,u_i,c}}{\sum_{j \in M, j \neq m} p_{m_j} g_{m,u_i,c} + \sigma^2 + I_{f_i,u_i,c}} \tag{3}$$

Whereas, σ^2, p_m and $g_{m,u,c}$ denotes the additive Gaussian noise, transmitted power by MBS m and channel gain from MBS m to user u with sub-channel c. $I_{f,u,c}$ shows the interference due to neighbor SBSs. p_{f_i} is the transmit power and from $i_t h$ SBS. $g_{f_i,u_i,c}$ is the channel gain between ith SBS and $i-th$ MUE. Eq. (3) represents the generalized form of SINR, due to neighboring cell interference, the SINR γ for (MUE) and (SUE) is calculated as:

$$\gamma_{m,u,c} = \frac{p_{m_i,u} g_{m,u_i,c}}{\sum_{j \in M, j \neq m} p_{m_j} g_{m,u_i,c} + \sum_{j \in F, j \neq 0} p_{f_i} g_{f_i,u_i,c} + \sigma^2} \tag{4}$$

$$\gamma_{f_i,u_i,c} = \frac{p_{f_i} g_{f_i,u_i,c}}{\sum_{j \in M, j \neq m}^{M} p_{m_j} g_{m,u_i,c} + \sum_{j=1, j \neq i, j \in F} p_{f_j} g_{f_j,u_i,c} + \sigma^2} \tag{5}$$

p_f, $g_{f_i,u_i,c}$, p_m, and $g_{m,u_i,c}$ is the transmitted power p and channel gain g from MBS and SBS respectively. It is assumed that all SBS are aware of network channel information.

According to Shannon's capacity formula, considering SBS, UE, and sub-channel, the achievable rate for the considered network can be written as:

$$R_{f,u,c} = \frac{W \log_2 (1 + \gamma_{f,u,c})}{Q_f} \qquad \forall f \in F \tag{6}$$

Where Q_f denotes the maximum number of UE connected to the SBS. The user association between the user and BS is considered a binary variable, where f SBS is associated with u users. Therefore, the transmit power constraint can be written as:

$$\sum_{u \in U} x_{f,u} p_{f,u,c} \leq p_{max} \tag{7}$$

$$\sum_{f \in B} x_{f,u} r_{f,u,c} \geq R_t \tag{8}$$

where p_{max} is the maximum transmit power of cell F and R_t is the minimum transmit data rate. Apart from transmission power, circuits, and related hardware also contribute to total energy consumption. The total power consumption is the sum of average hardware power and overall power consumption for transmission between user u and BS f. The total power consumption of the network is defined as:

$$P^{TOTAL} = P^{FIXED} + |F^{ACTIVE}|.P_{f,u} \tag{9}$$

Where the P^{FIXED} is the operational power that consists of the power required to keep that BS active, $P_{f,u}$ is the data transmission power, and $|F^{ACTIVE}|$ shows the activated or deactivated state of power amplifier efficiency. The power amplifier efficiency depends on the transmitter design [27], defined as $|F^{ACTIVE}| \in [0, 1]$, and is set to $F^{ACTIVE} = 1$ for the sake of simplicity. The goal is to maximize the aggregated data transmission while improving energy efficiency. Therefore, the total power consumption and EE between u user and f BS in the scenario considered can be written as:

$$U_{X,P} = \sum_{f \in F} P_f^{FIXED} + \sum_{f \in F} \sum_{u \in U} \sum_{c \in C_k} x_{f,u} p_{f,u,c} \tag{10}$$

$$EE = x_{f,u} \frac{R_{f,u,c}}{U_{X,P}} \tag{11}$$

2.2. Problem formulation

In order to enhance the energy efficiency of the overall network, a joint user association and power allocation problem is formulated below, considering QoS and transmit power constraints.

$$P1: f(x,p) = \max_{X,P} \sum_{f \in F} \sum_{u \in U} EE(X,P) \tag{12}$$

such that:

$$C1: \quad x_{f,u} \in \{0,1\}, \qquad \forall u \in U, f \in F \tag{13}$$

$$C2: \quad \sum_{f \in F} x_{f,u} = 1, \qquad \forall u \in U \tag{14}$$

$$C3: \quad \sum_{u} x_{f,u} p_{f,u,c} \leq p_{max}, \qquad \forall b \in B \tag{15}$$

$$C4: \quad \sum_{f} x_{f,u} c_{f,u} \geq R_t, \qquad \forall u \in U \tag{16}$$

$$C5: \quad \sum_{u} x_{bu} = Q_b, \qquad 0 \leq Q_b \leq N_u \tag{17}$$

$$C6: \quad \sum_{u} \sum_{k \in F, k \neq f} x_{f,u} p_{f,u,c} g_{f,u,c} \leq I_b, \qquad \forall f \in F \tag{18}$$

The Eq. (13) refers to the user association constraint, the constraint in Eq. (14) indicates the user scheduling constraint that each user can associate with one BS at a time, whereas Eq. (15) defines the total power consumption limitation. The QoS constraint is defined in Eq. (16), and Eq. (17) indicates the number of associated users. Eq. (18) indicates the cross-tier interference. Traditionally, such optimization problems are solved by heuristic search algorithms because of their low complexity nature. However, these algorithms take a considerable amount of time to converge and cannot be implemented in real-time scenarios. The power constraint optimization problem relies on an accurate model due to its model-driven nature. To overcome this inefficiency, parameterized deep Q-learning for the energy-efficiency can be applied due to its adaptive decision-control nature that converges to time and network scalability. The non-linear relationship between power allocation and QoS results in a non-convex problem. Whereas the combination of user association and finding optimal power levels leads to a combinatorial problem due to a large number of possible combinations. Hence, the $P1$ is a non-convex and combinatorial problem. It can be written as:

$$P2: \max_{X,P} \sum_{f \in F} \sum_{u \in U} x_{f,u} \frac{\frac{W \log_2 (1+\gamma_{f,u,c})}{Q_f}}{\sum_{f \in F} P_f^{FIXED} + \sum_{f \in F} \sum_{u \in U} \sum_{c \in C_k} x_{f,u} p_{f,u,c}} \tag{19}$$

such that:

$$C1: \quad x_{f,u} \in \{0,1\}, \qquad \forall u \in U, f \in F \tag{20}$$

$$C2: \quad \sum_{f \in F} x_{f,u} = 1, \qquad \forall u \in U \tag{21}$$

$$C3: \quad \sum_{u} x_{f,u} p_{f,u,c} \leq p_{max}, \qquad \forall b \in B \tag{22}$$

$$C4: \quad \sum_{f} x_{f,u} c_{f,u} \geq R_t, \qquad \forall u \in U \tag{23}$$

$$C5: \quad \sum_{u} x_{bu} = Q_b, \qquad 0 \leq Q_b \leq N_u \tag{24}$$

$$C6: \quad \sum_{u} \sum_{k \in F, k \neq f} x_{f,u} p_{f,u,c} g_{f,u,c} \leq I_b, \qquad \forall f \in F \tag{25}$$

The modified $P2$ Eq. (19) shows the discrete variable $x_{f,u}$ and continuous variable $p_{f,u,c}$, which is further targeted as a decomposition approach. The energy-efficient joint optimization problem of user association and power allocation in their respective action spaces is further elaborated in the subsequent section.

3. Multi-objective optimization problem formulation

The reinforcement learning algorithm enables network entities to learn its best policy from trial-and-error, which allows agents to make optimal decisions [28]. In order to overcome the limitations of RL, DRL has been introduced that takes advantage of Neural Network (NN) to train the learning process [29]. One reason to prefer DRL over traditional optimization methods is that irrespective of dynamic scenarios and network size, DRL leverages the information through past optimizations that help the agent to perceive the information from high dimensional space [29,30].

3.1. Parameterized deep reinforcement learning

Several variations of Q-learning are used for the domain of discrete action spaces, such as Q-Learning [9], DQN [5], double DQN [31]. Also, for continuous action spaces, Actor-Critic RL [18] and DDPG [32] are used. It is possible to discretize the continuous part into a finite number of discrete sets to deal with both discrete actions and continuous parameters domain. This leads to increased processing, resulting in the loss of potential parameters, sub-optimal policy, and requiring too many samples to learn anything functional. Also, this discretization of state–action space in joint resource allocation problems leads to large discretized state–action space due to possible large state–action combinations. In order to take full advantage of continuous parameters, a parameterized action space is considered for the joint user association and power allocation optimization problem. This hybrid action space also assists in dealing with the slow convergence problem and increased power consumption.

To model parameterized DRL, a Markov Game is considered with a parameterized action space consisting of discrete actions and associated continuous parameters. The Eq. (12) is formulated as Markov game, having continuous variable P while X is binary. The Markov game is represented by a tuple consisting of states, actions, transitioned states, and reward. For the assumption purposes the action space $a \in A$ is denoted by $a = (k, x_k)$ such that:

$$A^H = \{(k, x_k) | x_k \in X_k \text{ for all } 1 \leq k \leq K\} \tag{26}$$

where, $[K] = \{1, 2, \ldots, K\}$ is a discrete action and $x_k \in X_k$ is the continuous parameter. As the action space for user association is discrete and power allocation has continuous parameters, the discrete action space is $A^d = \{[x_{f,u}] : x_{f,u} \in \{0, 1\}, \text{where} u \in U, f \in F\}$ and the continuous parameter is the power allocation $p^u(t)$.

3.2. Multi-agent deep reinforcement learning formulation

The network control center will control the user association and assign the transmit power for users in order to improve energy efficiency. The decision depends on the training results that are performed online. Since several BSs and UEs interact actively in the environment, the formulation can be described as a Markov Game and represented as a tuple (S, A, R, S'), where S is defined as the state space, A is action space with possible actions of agents, R is the reward function, and S' is the state transition probability from state s to state s'. In order to solve this optimization problem, we assume each BS is a smart agent equipped with sufficient computational power. The environment comprises both association state and power allocation levels, which are allocated as a result of the agent's action. The state, action, and reward functions are defined as follows:

- **State:** The set of states at the time slot t is defined as $S^T(t) = s_1(t), s_2(t), \ldots, s_i(t)$. The agent interacts with the environment and observes the state while the agent is still unaware of the information on the other agents. The state space represents traffic load and network conditions. According to the constraint (16) and multi-agent environment, a threshold level of R_t is defined. Also, according to the constraint mentioned in Eq. (18), there is no cross-tier interference. However, it is important to consider that each user can only select one sub-channel. The state space from the user association and power allocation perspective while satisfying the QoS can be defined as: $s_i(t) = \{R_{t_i}, X_i, P_i\}$ where $s_i(t) \in S^T(t)$.
- **Action:** The set of actions related to the optimization problem are user association and power allocation, which according to t time slot, is defined as $A^T(t) = \{A^x(t), A^p(t)\}$. Whereas the $A^x(t)$ refers to the action of user association and $A^p(t)$ refers to power allocation. The values of matrix X defines $A^x = [x_{(f,u)}]$, such that $x_{(f,u)} \in \{0, 1\}, \forall f \in F, u \in U$. Similarly the power allocation is defined as $A^p(t) = [a_1^p(n), a_2^p(n), \ldots, a_u^p(n)]$ where, $A^p = [p_{(f,u,c)}], \forall f \in F, u \in U$, and $c \in C$ to map along the state of matrix X. A multi-agent environment is considered, which leads to the power allocation action to distribute transmit power. Hence power function is defined as:

$$p_i(a^p) = p_i^{min} + \frac{a^p}{m_i}(p_i^{max} - P_i^{min}) \tag{27}$$

- **Reward:** The focus is to maximize the EE while satisfying the QoS of UE, thus the reward function $r(s, t)$ is defined as:

$$r(s, a) = \begin{cases} f(x, p), & \text{if } R_f \leq R_f^{max} \forall f \in F \\ 0, & \text{if } R_f > R_f^{max} \forall f \in F \end{cases} \tag{28}$$

where $f(x, p)$ ensures the user association, energy efficiency, and QoS constraint. This reward function maps state–action pair with immediate reward, which is the value obtained from the reward function for a particular state. The aim of the agents is to maximize the cumulative reward, and it is the sum of immediate rewards. The agents achieve cumulative award by selecting actions that lead to long-term consequences in performance.

In order to describe the relationship between the value of the current state and transitioned states, the Bellman equation is used. Bellman equation has the property to describe action value function without anticipating future rewards and is described as follows:

$$\hat{Q}(s, a) = max_\pi \mathbb{E}[r_t + \delta r_{t+1} + \delta^2 r_{t+2} + \cdots [s_t = s, a_t = a, \pi]] \tag{29}$$

where δ is the discount factor and π is the policy.

The $s \in S$ and $a \in A$ are possible sets of states and actions that are the tuple of environment $\{S, A, R, S'\}$ to describe the relation of an agent with the environment. As a hybrid environment state is considered, in order to map the state s and discrete action a to its equivalent parameter, a weight θ is defined such as $x_k(s, \theta)$, which leads the action as a tuple $a = (k, x_k)$ for the policy π. The policy π is defined as $\pi(s, a)$ that depends on the current state s such that $\sum_{a \in A} \pi(s, a) = 1$. Let the agent observe the state s of the environment at a certain time stamp t and take action $a(t) \in A$ according to a policy π. After observing the environment, the agent takes an action a, gets a reward r, and moves from the current state s to s'. This step of environment interaction leads to a transition experience at time t defined as action-value function $E[Q(s, a, x_k(k; \theta); \omega)]$. The goal is to compute an optimal policy π to maximize the reward function without any information on state transition. Hence, the action value-function, according to Bellman Eq. (29) can be written as:

$$Q^\pi(s, a) = R(s, a) + \delta \sum_{s' \in S} P_{ss}^{a\prime} [\sum_{a' \in a} \pi(s', a') Q^\pi(s', a')] \tag{30}$$

where δ is the discount factor, P_{ss}^a is the transition probability from one state to another following an action. The expected reward for the transition from state s to take action a is $R(s, a) = \mathbb{E}[r' | s(t) = s, a(t) = a]$.

A deterministic function is defined with the weights θ to map the discrete state k to continuous parameters x_k.

$$x_k = \mu k(s, \theta) \tag{31}$$

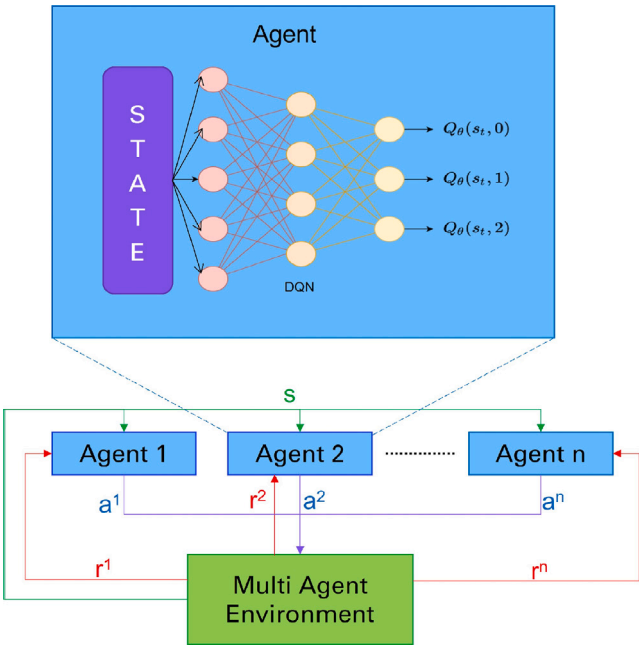

Fig. 2. DQN-based multi-agent system that interacts with the environment to gather data for training and outputs the Q-value of the action.

An action-value Q-function $Q(s,k,x_k,\omega)$ with value network weight ω and the hybrid action (Eq. (31)) is defined to map k states and hybrid actions to $\mathbb{R}$. For larger and hybrid action spaces such as the considered optimization problem, traditional reinforcement learning overlooks the states and encounters storage scarcity issues. DQN is a feasible choice here, as it maintains a lookup table and contains a vector θ for parameters, DQN can be expressed as $Q(s,a,\theta)$. For a balanced multi-agent PDQN, "quasi-static target network" and "experience replay" networks are adopted [33]. Two DQNs are combined as a target network and training network with parameters $Q(s,a,\theta)$ and $Q(s,a,\bar{\theta})$, respectively. A replay buffer D is implemented to train DQN for a random mini-batch. The loss function to train PDQN for actor-parameter and actor can be obtained by:

$$L^x(\theta) = -\frac{1}{N}\sum_{i=1}^{N} Q(s_i,k_i,x_{k_i}(s_i,\theta);\omega) \tag{32}$$

$$L^Q(\omega) = -\frac{1}{N}\sum_{i=1}^{N}(\bar{O} - Q(s_i,k_i,x_{k_i}(s_i,\theta);\omega))^2 \tag{33}$$

The target DQN is used to replace the output of trained DQN, and the target output of PDQN is:

$$\bar{O} = r + \delta max Q(s',k',x'_k(s';\bar{\theta});\omega) \tag{34}$$

N is the size of the mini-batch, θ and ω are the weights for the replay buffer.

4. Multi-agent parameterized deep reinforcement learning solution for user association and power allocation

Multi-agent PDQN is an enhanced version of single-agent PDQN which facilitates multiple agents to learn their optimal policies in parallel while interacting with the environment as shown in Fig. 2. In order to learn and implement the global policy, the agents can be deployed in both cooperative and non-cooperative manner [34]. For MA-PDQN, each agent adopts the same setting as PDQN and learns a policy based on its local observation on the basis of its Q-function.

The training and testing pseudo-code of the proposed MA-PDQN-based algorithm for user association and power allocation to achieve energy-efficient operation is shown in Algorithm 1 and Algorithm 2.

Algorithm 1: Pseudo code: MA-PDQN based testing phase algorithm for user association and power allocation

Initialize: The environment, initial states for each agent, and Q-network weights for each agent on the basis of training networks weights
for *(episode = 1 to number of episodes)* **do**
  Assign Initial state for each agent
  Set loop=0
  while *loop* $\neq 0$ **do**
    for *each agent* **do**
      Get Q-values for all A^x and A^p
      Choose an a using ϵ greedy
      Execute the a: user association and power allocation
      Observe s' and r
      Update s for each agent
    if *any agent reaches a terminal state* **then**
      Set loop=1

Algorithm 2: Pseudo code: MA-PDQN based training phase algorithm for user association and power allocation

Initialize: the DNN (3 layer neural network)
for *each agent* **do**
  Initialize: Learning rate (α_d,α_c), mini-batch size N, replay memory D, exploration parameter ϵ, probability distribution ξ
  Initialize: Random weights θ,ω, enable $\bar{\theta}=\theta$, and number of training episodes
for *episode = 1 to number of episodes* **do**
  Initialize: Environment and initial states for each agent
  Set loop=0
  while *loop =0* **do**
    for *each agent* **do**
      Choose an a using ϵ greedy
      Execute the action (user association and power allocation) in the environment
      Observe the s', r, and equation (23)
      Calculate equation (9) on the basis of action
      if *equation (23) and equation (22) is not satisfied* **then**
        Set a negative reward to penalize the action
      Calculate equation (11)
      Calculate r w.r.t EE
      Store the experience (s, a, r, s') in replay memory D
      if $D \neq full$ **then**
        Sample a batch of experiences from D
        Update weights for parameterized DQN
      Update the s for each agent
    if *any a reaches a terminal state* **then**
      loop = 1
  if *update episode of target Q-network = 0* **then**
    Update the target network with same weights
  Update the trained Q function weights for each agent

5. Performance evaluation

5.1. Simulation setup

The MA-PDQN algorithm is simulated for a two-tier small cell network, with one MBS along 100 antennas, ten SBSs, and fifty UE, as shown in Fig. 3.

The MBS has a coverage area of 500 m, with UEs and SBSs randomly distributed under its coverage area. The area is divided into clusters that are under the coverage of SBS. The channels are considered to be slow Rayleigh fading channels, therefore no change occurs during the

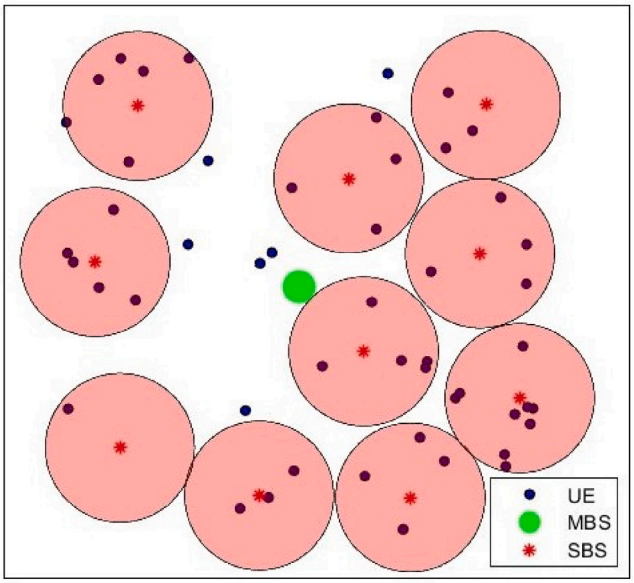

Fig. 3. Network layout with one MBS, ten SBSs and fifty UEs.

Table 2
Simulation parameters.

Parameters	Values
Number of MBS	1
Number of SBS	10
Number of UE	50
Number of channels	10
Transmit power of MBS	46 dBm
Transmit power of SBS	30 dBm
Transmit power of UE	23 dBm
Path loss model for MBS	34 + 46 log(d)
Path loss model for SBS	37 + 30 log(d)
Noise power	−174 dBm/Hz
Radius of MBS	500 m
Radius of FBS	100 m
DQN Parameters	
Episodes	200
Observe steps	200
Mini-batch size	5
Learning rate	0.01
Hidden layers (N_1,N_2, N_3)	64, 64, 32
Discount rate	0.7
Replay memory (D)	500
Maximum ϵ	0.9
Minimum ϵ greedy increment	0.003
Optimizer	RMSProp
Activation function	ReLu

transfer of a block of data. The maximum transmission power of MBS and SBS is considered 46 dBm and 30 dBm, respectively. The path loss factor for MBS and SBS is:

$$MBS = 34 + 46\log(d) \tag{35}$$

$$SBS = 37 + 30\log(d) \tag{36}$$

Where d is the distance between UE and BS in m. In addition to these parameters, other simulation parameters are mentioned in Table 2.

In order to evaluate the proposed algorithm for joint resource allocation for improved energy efficiency, Python and TensorFlow are used

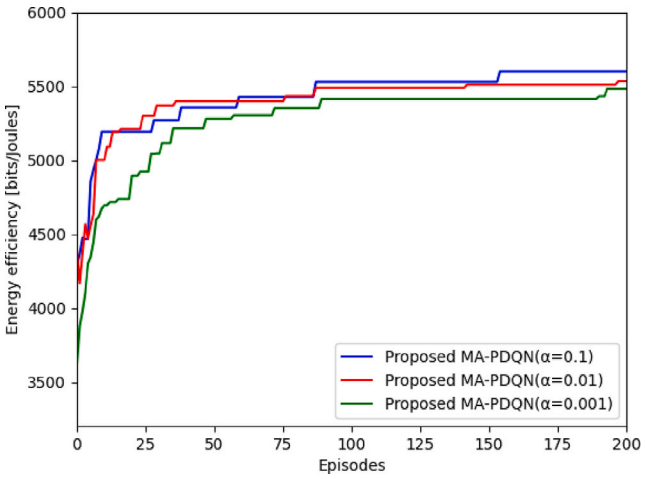

Fig. 4. Energy efficiency v.s different learning rates.

for implementation purposes. The DQN used consists of one input layer, three hidden layers, and one output layer. The hidden layer consists of 64, 32, and 32 neurons, respectively. The sub-channel allocation satisfies the constraint mentioned in Section 2.2. The preference is to use every neuron for learning purposes independent of the size of initial activation values. A truncated distributed function is used instead of a probability distribution function to avoid saturation, as saturation occurs in neural network when activation values are too big or small.

The threshold level is set to 0.1, but the discount factor can be chosen randomly and is considered to be best if between 0 and 1. The discount factor was set to 0.7 because if the discount factor is considerably small the agent will only focus on the immediate reward. The initial weights are set as 0.43 and 0.15. In order to increase the learning rate and speed up the converging speed, the RMSProp optimizer is used with an adaptive learning rate.

To evaluate the performance of the proposed algorithm, we benchmarked the results with four algorithms which are WMMSE, Game theory-based joint user association and power allocation algorithm, Q-learning, and DRL-based power allocation.

5.2. Simulation results

The proposed algorithm is studied with different configurable hyperparameters. The results shown in Fig. 4 show the learning process of DNN on the basis of different learning rates α with an increase in episodes. The training steps are large at the initial stages. However, the convergence speed increases with episodes as the model starts to adapt based on the experience. The learning rate $\alpha = 0.01$ and $\alpha = 0.001$ have faster convergence as compared to $\alpha = 0.1$, which makes 0.001 comparatively less stable. The increased convergence speed is because low α may result in local optimum and may cause the process to get stuck, whereas a high learning rate may result in a quick sub-optimal solution (unstable training process). The energy efficiency is also better for the learning rates 0.1 and 0.01. Thus, for the real-time practical execution of the proposed algorithm, the learning rate is set to 0.1.

The mean square error (MSE) of the network model at different learning rates with 200 epochs are shown in Fig. 5. MSE is given by the loss function equation. The network model is designed with three hidden layers and runs with different learning rates $\alpha \in \{0.1, 0.01, 0.001\}$. The low values of MSE results shown in Fig. 5 suggest that the learning rate $\alpha = 0.1$ should be the optimal choice. High values of loss indicate poor model performance. Further, the different values of discount factors δ are studied to select the value on the basis of its performance.

Fig. 5. The Mean Square Error for different learning rates.

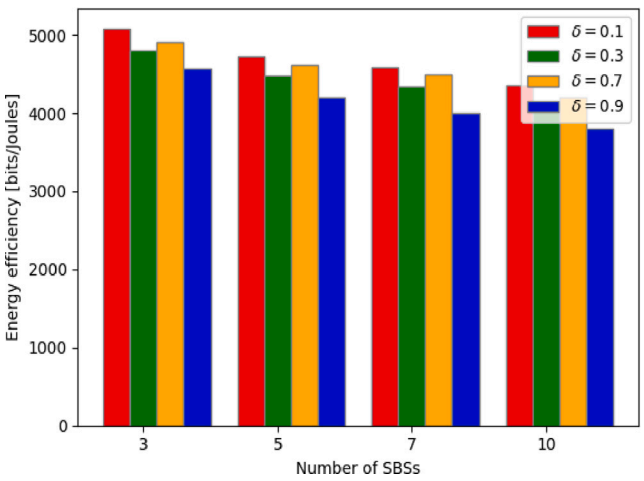

Fig. 7. Energy efficiency v.s network scalability for different discount factor values.

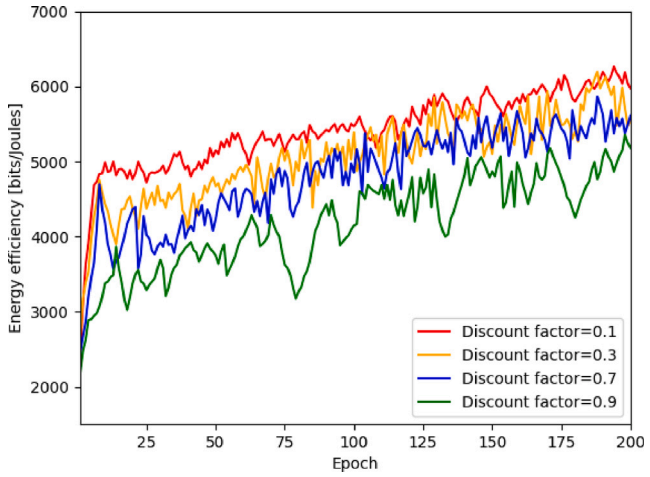

Fig. 6. Effect of discount factors on energy efficiency.

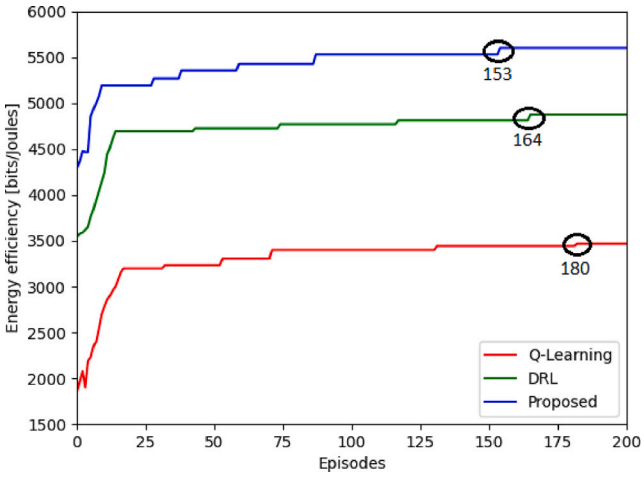

Fig. 8. Proof of convergence.

Fig. 6 shows the performance over the training phase of the proposed algorithm for $\delta \in \{0.1, 0.3, 0.7, 0.9\}$. The $\delta = 0.1$ has the higher value of EE as compared to other values of discount factors, whereas $\delta = 0.9$ has the lowest value. Lower discount factor values tend to focus more on the reward rather than learning. After the training phase, the trained model is further tested on the network with similar discount values.

Fig. 7 shows the importance of the discount factor on the agent's learning. The considered discount factors $\delta \in \{0.1, 0.3, 0.7, 0.9\}$ are studied again for energy efficiency scores for the proposed technique considering the network scalability. The trained DQN is studied for the 4 different SBSs networks. It is visible that the maximum discount factor value $\delta = 0.9$ achieves minimum EE values, and the minimum δ values achieve the highest levels of energy efficiency because of the agent's myopic behavior at low discount factors. The selection of the discount factor is a random process that should be between 0 and 1. However, due to the dependency on the agent's learning, the value should be considered carefully as it improves the learning outcome. In our scenario, the enhancement in energy efficiency is important, but the focus is on the learning of the agent, and for the required correlation between the agent's actions and future reward, the discount factor of $\delta = 0.7$ is considered. It is observed that the lower values of the discount factor are not required as it focuses on the immediate energy efficiency reward rather than learning from the changes. On the contrary, a higher discount factor value decelerates the DQN's response. Hence, a moderate discount factor is desirable that adapts to the network conditions and contributes to the required objectives.

The performance of the proposed MA-PDQN is further studied in terms of energy efficiency and convergence point with increasing episodes. The EE of the proposed algorithm is comparatively much higher than conventional DRL and Q learning algorithms. As shown in Fig. 8, in the start, graphs show the instability of algorithms because of their learning process, as agents are making decisions randomly. MA-PDQN outperforms DRL and Q-learning, as at approximately 153 episodes, the system became more stable. For the case of DRL, the system is more stable as compared to Q-learning, but episode 164 showed change, suggesting that the system has more space for improvement.

In Fig. 9, the energy efficiency is investigated along with four other algorithms i.e., WMMSE, Game theory, DRL, and Q-learning. Initially, the EE is not better with less number of BS, but as the SBS increases, EE starts improving gradually. This is because an increase of SBSs leads to less load on individual BSs. The results are calculated on the basis of $F = 10$. The EE of the machine learning algorithms is much better than the WMMSE and game theory, according to our proposed scenario. DRL-based power allocation scenario allocates power on discrete levels and shows better results as compared to Q-learning. However, when the action space increases with the increase in the number of UEs and SBSs, the performance of DRL starts showing a linear behavior because of a sudden increase of action spaces that is difficult to handle. In the presented scenario, the proposed MA-PDQN shows better results with

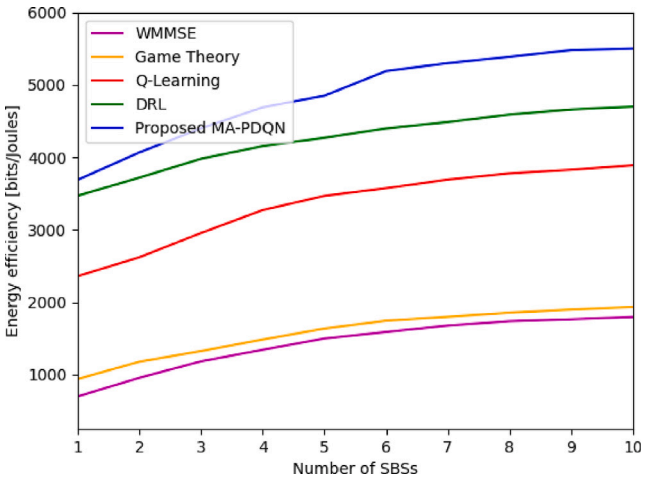

Fig. 9. Energy efficiency with different number of small base stations.

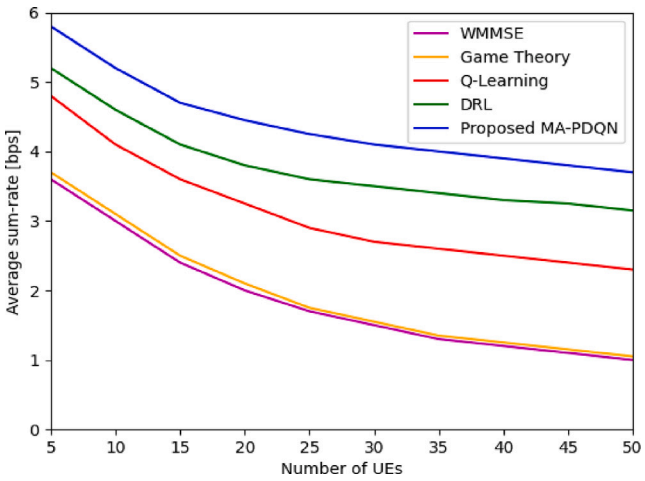

Fig. 11. Average Sum Rate v.s number of user equipment.

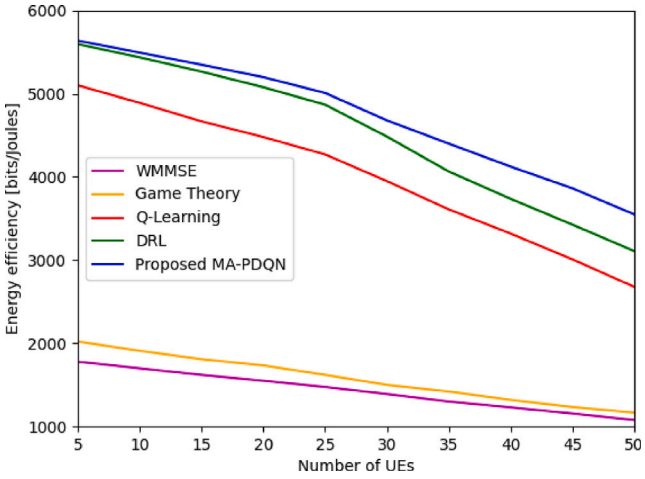

Fig. 10. Energy efficiency with a different number of user equipment.

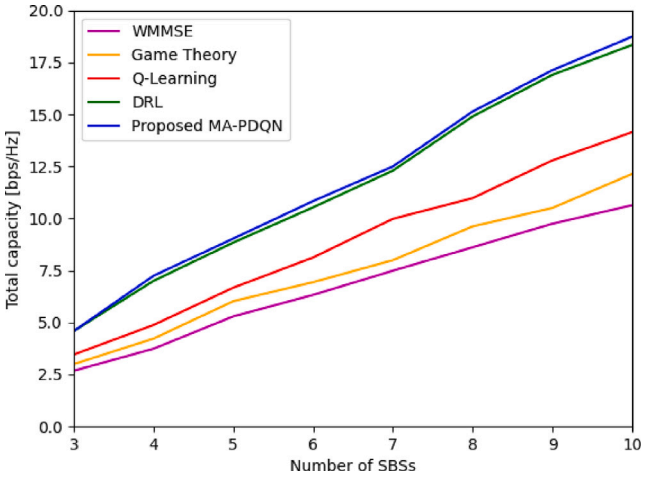

Fig. 12. Sum capacity of network with number of SBSs.

an increase in action and state space, as power allocation is not limited to discrete levels like other DRL algorithms. Also, in the case of Q-learning, the maximum power is allocated for transmission. Using the maximum transmission power for minimum transmission distance results in loss and eventually in degradation of energy efficiency. Hence, while designing a network, the number of SBSs should be considered appropriately.

Fig. 10 shows the results of energy efficiency with an increase in the number of UEs while $F = 10$. Initially, the improvement in EE is much better with less number of UE, but as the UE increases, EE starts decreasing at a gradual speed. This is because the increase of UEs leads to more load on SBSs. As compared to other reference algorithms, our proposed algorithm performs much better in terms of energy efficiency even with the increasing UEs. This is because of joint user association and power allocation strategy, which deals with both allocations on their respective different action spaces. Also, the multi-agent properties of the proposed algorithm optimize the allocation strategy, and with little information transfer, the load balance and interference management among SBSs is achieved. Practically, the users can vary in number according to the scenario, and the dense network can affect the behavior of the algorithm and result in energy and sum rate degradation. The behavior of the average sum rate with UE density is studied in Fig. 11. To show the effect of UE density on the average sum rate, 50 UEs are considered. The plotted result shows the degradation in the average sum rate with the increase in UE. However, the proposed algorithm has observed linear behavior in sum rate drop as compared to other benchmarked techniques. A slight increase in capacity is also observed with an increase in the number of SBSs, compared to other techniques, as shown in Fig. 12. Initially, the sum rate capacity was less with the three deployed BSs. However with the increase in number of SBSs, the sum capacity increases as a result of properly distributing traffic more evenly.

5.3. Complexity

The MA-PDQN employs a neural network to train the agents in a model-free framework, whereas the time complexity of the DNN can be represented in Floating-point Operations (FLOPs) [35]. According to [35] for each layer of two-sided matching, the FLOPs are represented as:

$$FLOPs = 2l_i A_i = O(BS!2A/B)$$

where, l is the input dimensions for the ith layer and O is the output layer for the ith layer. In our case, the neural network is employed for

training the RL. Hence, the training complexity is given by [35]:

$$\Omega = O(e \times t(S \times n \times l_i \times A))$$

where e is the number of episodes of training, t is the iterations, S is the input layer size, A is the output layer size, and l is the hidden layer, which is three in our scenario. Hence time complexity for the single-agent RL that employs the NN architecture with three hidden layers will be:

$$\Omega = O(e \times t(S \times n \times l_1 + n \times l_2 + n \times l_3 \times A))$$

Hence, in our proposed scenario, time complexity with multi-agent and three hidden layers DQN is represented as:

$$\Omega = O(e \times t((N \times (S + A) \times n \times l_1 + n \times l_2 + n \times l_3)))$$

Hence, the training complexity will be $O(N\Omega)$ for multi-agent PDQN.

6. Conclusion

This paper studied a joint user association and power allocation framework in a small-cell-based wireless network using MA-PDQN. The formulated wireless network improves energy efficiency while considering the maximum transit power, association, and QoS constraint. To jointly optimize the user association and power allocation, a parameterized action space has been formulated to update the system for optimal hybrid actions. The proposed MA-PDQN is implemented in a centralized manner for the coordination among agents achieved through information exchange among agents to implement a global strategy. The MA-PDQN has been implemented as a model-free framework, such that the framework does not have any dependency on the number of UE and BSs and does not require re-training for the new model. The simulation results show improved results of our proposed MA-PDQN is 33% as compared to DRL. The convergence speed shows that the proposed algorithm has better energy efficiency. In addition, the proposed method resulted in better performance on energy efficiency in terms of the number of users, SBSs, and different learning rates. This research can help practitioners improve resource utilization and address the critical concern regarding energy-efficient practical wireless network deployments. The proposed solution possesses the potential for future expansion across multiple user associations, uplink considerations, seamless integration of renewable energy sources, and innovative energy harvesting techniques.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Amna Mughees reports financial support and article publishing charges were provided by Ministry of Education Malaysia.

Data availability

Data will be made available on request.

Acknowledgment

All authors approved the final version of the manuscript.

Appendix. Symbol table

(See Table A.3)

Table A.3
Symbols definitions.

Symbol	Definition
B	Set of all base stations
F	Set of all SBSs
U	Set of all UEs
N_T	Antenna array size
N_{sub}	Total number of sub-channels
$x_{f,u}$	User association
γ	SINR
p	Transmission power
g	Channel gain
R	Sum rate
Q_f	Maximum number of associated users
EE	Energy efficiency
δ	Discount factor
π	Policy
$p^u(t)$	Power allocation
A^H	Hybrid action space
s	State
a	Action
r	Reward
s'	State transition
θ	weight in actor-parameter
ω	Weight in actor-network
D	Replay buffer
α	Learning rate
N	mini-batch size
ϵ	Exploration parameter
ξ	Probability distribution

References

[1] A. Ahad, M. Tahir, M.A.S. Sheikh, N. Hassan, K.I. Ahmed, A. Mughees, A game theory based clustering scheme (GCS) for 5G-based smart healthcare, in: 2020 IEEE 5th International Symposium on Telecommunication Technologies, ISTT, IEEE, 2020, pp. 157–161.

[2] A. Mughees, M. Tahir, M.A. Sheikh, A. Ahad, Energy-efficient load-aware user association in ultra-dense wireless network, in: 2021 26th IEEE Asia-Pacific Conference on Communications, APCC, 2021, pp. 254–259, http://dx.doi.org/10.1109/APCC49754.2021.9609810.

[3] F. Marzouk, J.P. Barraca, A. Radwan, On energy efficient resource allocation in shared RANs: survey and qualitative Analysis, IEEE Commun. Surv. Tutor. 22 (3) (2020) 1515–1538.

[4] Y.S. Nasir, D. Guo, Deep actor-critic learning for distributed power control in wireless mobile networks, in: 2020 54th Asilomar Conference on Signals, Systems, and Computers, IEEE, 2020, pp. 398–402.

[5] F. Meng, P. Chen, L. Wu, Power allocation in multi-user cellular networks with deep Q learning approach, in: ICC 2019-2019 IEEE International Conference on Communications, ICC, IEEE, 2019, pp. 1–6.

[6] Y. Lu, H. Lu, L. Cao, F. Wu, D. Zhu, Learning deterministic policy with target for power control in wireless networks, in: 2018 IEEE Global Communications Conference, GLOBECOM, IEEE, 2018, pp. 1–7.

[7] A. Khodmi, S.B. Rejeb, N. Agoulmine, Z. Choukair, A joint power allocation and user association based on non-cooperative game theory in an heterogeneous ultra-dense network, IEEE Access 7 (2019) 111790–111800.

[8] Y. Chen, L. Xun, S. Zhang, The energy efficiency of heterogeneous cellular networks based on the Poisson hole process, Future Internet 15 (2) (2023) 56.

[9] R. Amiri, H. Mehrpouyan, L. Fridman, R.K. Mallik, A. Nallanathan, D. Matolak, A machine learning approach for power allocation in HetNets considering QoS, in: 2018 IEEE International Conference on Communications, ICC, IEEE, 2018, pp. 1–7.

[10] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, Y. Jiang, Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks, IEEE Trans. Wireless Commun. 18 (11) (2019) 5141–5152.

[11] D. Li, H. Zhang, K. Long, W. Huangfu, J. Dong, A. Nallanathan, User association and power allocation based on Q-learning in ultra dense heterogeneous networks, in: 2019 IEEE Global Communications Conference, GLOBECOM, IEEE, 2019, pp. 1–5.

[12] Y.-H. Xu, Q.-M. Sun, X.-R. Xu, W. Zhou, G. Yu, Energy efficiency and delay determinacy tradeoff in energy harvesting-powered zero-touch deterministic industrial M2M communications, Eng. Appl. Artif. Intell. 121 (2023) 105997.

[13] Y.-H. Xu, Y.-B. Tian, P.K. Searyoh, G. Yu, Y.-T. Yong, Deep reinforcement learning-based resource allocation strategy for energy harvesting-powered cognitive machine-to-machine networks, Comput. Commun. 160 (2020) 706–717.
[14] X.-R. Xu, Y.-H. Xu, W. Zhou, A. Nallanathan, Energy efficient resource allocation for UAV-served energy harvesting-supported cognitive industrial M2M networks, IEEE Wirel. Commun. Lett. (2023).
[15] Z. Rezaei, B.S. Ghahfarokhi, Energy and spectrum efficient cell switch-off with channel and power allocation in ultra-dense networks: A deep reinforcement learning approach, Comput. Netw. (2023) 109912.
[16] H. Yang, J. Zhao, K.-Y. Lam, Z. Xiong, Q. Wu, L. Xiao, Distributed deep reinforcement learning-based spectrum and power allocation for heterogeneous networks, IEEE Trans. Wireless Commun. 21 (9) (2022) 6935–6948.
[17] C.-K. Hsieh, K.-L. Chan, F.-T. Chien, Energy-efficient power allocation and user association in heterogeneous networks with deep reinforcement learning, Appl. Sci. 11 (9) (2021) 4135.
[18] Y. Wei, F.R. Yu, M. Song, Z. Han, User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach, IEEE Trans. Wireless Commun. 17 (1) (2017) 680–692.
[19] Z. Cheng, M. Liwang, N. Chen, L. Huang, N. Guizani, X. Du, Learning-based user association and dynamic resource allocation in multi-connectivity enabled unmanned aerial vehicle networks, Digit. Commun. Netw. (2022).
[20] Z. Cheng, N. Chen, B. Liu, Z. Gao, L. Huang, X. Du, M. Guizani, Joint user association and resource allocation in HetNets based on user mobility prediction, Comput. Netw. 177 (2020) 107312.
[21] Y.-H. Xu, W. Zhou, Y.-G. Zhang, G. Yu, Stochastic game for resource management in cellular zero-touch deterministic industrial M2M networks, IEEE Wirel. Commun. Lett. 11 (12) (2022) 2635–2639.
[22] Y.-H. Xu, J.-H. Li, W. Zhou, C. Chen, Learning-empowered resource allocation for air slicing in UAV-assisted cellular V2X communications, IEEE Syst. J. 17 (1) (2022) 1008–1011.
[23] A. Alwarafy, M. Abdallah, B.S. Çiftler, A. Al-Fuqaha, M. Hamdi, The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions, IEEE Open J. Commun. Soc. 3 (2022) 322–365.
[24] A. Mughees, M. Tahir, M.A. Sheikh, A. Ahad, Energy-efficient ultra-dense 5G networks: recent advances, taxonomy and future research directions, IEEE Access 9 (2021) 147692–147716.
[25] J. Zhu, F. Wu, J. Zhao, An overview of the action space for deep reinforcement learning, in: 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence, 2021, pp. 1–10.
[26] A. Lazaric, M. Restelli, A. Bonarini, Reinforcement learning in continuous action spaces through sequential Monte Carlo methods, in: Advances in Neural Information Processing Systems. Vol. 20, 2007.
[27] S. Cui, A.J. Goldsmith, A. Bahai, Energy-constrained modulation optimization, IEEE Trans. Wirel. Commun. 4 (5) (2005) 2349–2360.
[28] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: A survey, J. Artif. Intell. Res. 4 (1996) 237–285.
[29] N.C. Luong, D.T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, D.I. Kim, Applications of deep reinforcement learning in communications and networking: A survey, IEEE Commun. Surv. Tutor. 21 (4) (2019) 3133–3174.
[30] A. Mughees, M. Tahir, M.A. Sheikh, A. Ahad, Towards energy efficient 5G networks using machine learning: Taxonomy, research challenges, and future research directions, IEEE Access 8 (2020) 187498–187522.
[31] Y. Xiao, J. Wu, J. Liu, Power allocation for device-to-multi-device enabled HetNets: A deep reinforcement learning approach, in: 2021 IEEE Global Communications Conference, GLOBECOM, IEEE, 2021, pp. 01–06.
[32] C. Wang, D. Deng, L. Xu, W. Wang, F. Gao, Joint interference alignment and power control for dense networks via deep reinforcement learning, IEEE Wirel. Commun. Lett. 10 (5) (2021) 966–970.
[33] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.
[34] A. Alwarafy, M. Abdallah, B.S. Ciftler, A. Al-Fuqaha, M. Hamdi, Deep reinforcement learning for radio resource allocation and management in next generation heterogeneous wireless networks: A survey, 2021, arXiv preprint arXiv:2106.00574.
[35] H. Zhang, H. Zhang, K. Long, G.K. Karagiannidis, Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation, IEEE Trans. Netw. Sci. Eng. 7 (4) (2020) 2406–2415.

Amna Mughees is working as a Graduate Research Assistant with Sunway University, Malaysia. She received the M.S. degree in Electrical Engineering from the Department of Electrical and Computer Engineering, Comsats University Pakistan. She is pursuing a Ph.D. with the School of Engineering and Technology, Sunway University, Malaysia. She is the author of many peer-reviewed research articles that have been published at various top conferences and journals. Her research interests include 5G, wireless networks, HetNets, energy efficiency, radio resource management, cellular handover, the Internet of Things, smart cities, and artificial intelligence.

Mohammad Tahir received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Department of Electrical and Computer Engineering, International Islamic University Malaysia, in 2011 and 2016, respectively. Before joining academics, he worked with the Research and Development Division of the industry for seven years on several projects related to the Internet of Things and cognitive radio. His research interests include 5G, the Internet of Things, game theory for wireless networks, wireless security, blockchain, and autonomic computing.

Muhammad Aman Sheikh is a lecturer in Electronics and Computer Systems Engineering at Cardiff School of Technologies at Cardiff Metropolitan University, United Kingdom. Prior to joining Cardiff Metropolitan University UK, he was Program Leader and a Senior Lecturer at the School of Engineering and Technology, Sunway University. Previously he was a Senior Scientist in the R&D sector and was involved in enormous projects for PETRONAS. He is a Fellow of the Institute of Electrical and Electronic Engineers, Malaysia (MIEEE) and, since 2010, a Registered Engineer with the Board of Engineers. An active researcher, he has published extensively and established global collaborations by working with international collaborators and securing international grants. He was awarded the Graduate Assistant Scholarship during his M.Sc. and Ph.D. in Electrical and Electronics Engineering in 2013 and 2018, respectively. He also received prestigious awards during his Ph.D. candidature, including Best Graduate Assistant Award and High Achiever Award.

Angela Amphawan received the Ph.D. degree in optical engineering from the University of Oxford, U.K. She is currently leading the Photonics Research Laboratory at the School of Engineering and Technology, Sunway University. Prior to this, she was the Deputy Vice Chancellor of the University Malaysia of Computer Science and Engineering. Her research projects have been funded by the U.S. Department of States, German Government, Malaysian Ministry of Education, and Telekom Malaysia. Her research interests include optical communications and sensing, covering optical fibers, free-space optics, radio-over free space optics, digital imaging, and the Internet of Things. She has won several Best Paper Awards and exhibition medals. She was a recipient of the Fulbright Award from the Research Laboratory of Electronics, Massachusetts Institute of Technology. She serves on the Editorial Board of the APL Photonics journal with the American Institute of Physics. She is on the National 5G Task Force. She was previously on the Editorial Board of Transactions on Emerging Telecommunications Technologies (Wiley) and the Publicity Co-Chair of the IEEE Wireless Communications and Networking Conference. She has given keynote addresses at several Fulbright and IEEE events.

Yap Kian Meng is currently the Founder of the House of Multimodal Evolution (H.O.M.E.) Laboratory, which is to provide novel and innovative multimodal applications and communications toward the technology world. He is also the Founder of the Human–Machine Collaboration Research Centre (HUMAC) which aims to be the nation's main technology hub and to demonstrate its commitment to sustainable development. He has spent 20 years working in data networks (IT), telecommunications, computer networking, haptics, manufacturing industries, and electronic control system in machinery. His research project on DHVE is meant to have multi-sensory feedback data, such as voice, audio, and force, over fixed and wireless networks. He is a principal investigator for a few projects that are supported by the Ministry of Higher Education (MOHE), ERGS, MCMC, Lancaster University, U.K., PPRN, industrial partners, and Sunway University Internal Grant, respectively. His current research interests include haptic over the distributed virtual environment (DVE) in HCI environment, haptics and odour

sensing, drones, odour sensing/tracking, tele-haptics, tele-robotic, materials sensing, assistive technology for limited vision and hearing impaired, and AR&VR.

Abdul Ahad is working as an Assistant Professor at the University of Management and Technology, Sialkot Campus, Pakistan and Visiting Researcher at Istanbul Technical University of Turkey. He received M.Sc. degree in computer science from Swat University, Pakistan, in 2014, M.S. degree in computer science from Virtual University, Pakistan, in 2017, Ph.D. degree in computer science from Sunway University, Malaysia in 2022. His research interests include 5G, 6G, Internet of Things, wireless networks, Wireless Body Area Networks (WBAN), Smart Cities, Intelligent Transportation, Intelligent Healthcare and artificial intelligence. He is the author of many peer-reviewed research articles that have been published at various top conferences and journals. He is a regular reviewer of many conferences and journals, including IEEE Access, Computer Networks, International Journal of Distributed Sensor Networks, Sensors and others.

Kazem Chamran received a B.Eng. degree (honor) in Electrical and Electronic Engineering from Sheffield Hallam University, in 2010, a M.Sc. degree in Network and Communication Engineering from University Putra Malaysia (UPM), in 2015, and Ph.D. in Computing from Sunway University Malaysia in 2022. Currently, he is head of research at City University Malaysia and senior lecturer at the faculty of IT. He is also a member of several national and international committees and scientific societies like IEEE and Malaysian research committees. His research interests are in the areas of Artificial Intelligence, 5G-6G, Energy harvesting, Machine Learning, and IoT.