Deep Reinforcement Learning Based Resource Allocation for Heterogeneous Networks

Helin Yang[1], Jun Zhao[1,2], Kwok-Yan Lam[1,2], Sahil Garg[3], Qingqing Wu[4], and Zehui Xiong[5]
[1]Strategic Centre for Research in Privacy-Preserving Technologies and Systems, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Electrical Engineering Department, École de technologie supérieure, Université du Québec, Montréal, Canada
[4]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China
[5]Pillar of Information Systems Technology and Design Singapore University of Technology and Design, Singapore
Email: {hyang013, junzhao, kwokyan.lam, zxiong002}@ntu.edu.sg, sahil.garg@ieee.org, qingqingwu@um.edu.mo

***Abstract*—This paper investigates the problem of distributed resource management (i.e., joint device association, spectrum allocation, and power allocation) in two-tier heterogeneous networks without any central controller. Considering the fact that the network is highly complex with large state and action spaces, a multi-agent dueling deep-Q network-based algorithm combined with distributed coordinated learning is proposed to effectively learn the optimized intelligent resource management policy, where the algorithm adopts dueling deep network to learn the action-value distribution by estimating both the state-value and action advantage functions. Under the distributed coordinated learning manner and dueling architecture, the learning algorithm can rapidly converge to the optimized policy. Simulation results demonstrate that the proposed distributed coordinated learning algorithm outperforms other existing learning algorithms in terms of learning efficiency, network data rate, and QoS satisfaction probability.**

***Index Terms*—Heterogeneous wireless networks, distributed resource management, dueling deep reinforcement learning.**

I. Introduction

DENSE deployment of small base stations (BSs) in multi-tier heterogeneous networks has been considered as one of important effective solutions to meet the ever-increasing wireless communication demands in fifth-generation (5G) and beyond networks [1]-[2]. The heterogeneous networks can deploy a number of small-coverage micro or pico cells within a macrocell for enhancing spectrum utilization efficiency and network coverage [1], [2]. Specifically, micro BSs and pico BSs reuse and share the same spectrum with macro BSs to improve the spectrum efficiency. Thus, heterogeneous networks not only enhance the network capacity, but also satisfy the growing communication demands of users in the future wireless networks [1]-[3].

Although the wireless network performance can be improved, the deployment of heterogeneous networks also brings new challenges, such as mobile device association, inter-cell interference (ICI) management, spectrum allocation, and power allocation [2]. In this context, network resource optimization becomes a critical issue, and there is need for advanced solutions to overcome it. Two joint device association and resource allocation algorithms were proposed to maximize the network capacity [4], [5], where the mixed-integer nonlinear optimization problem was addressed by decoupling it into two sub-problems. Considering that mobile devices have different quality-of-service (QoS) requirements, i.e., delay constraint and minimum rate requirement, QoS-aware device association and resource management approaches were proposed for heterogeneous wireless networks [6]-[8].

Since the complete information and environment's evolution are generally unknown in mobile environments, model-free reinforcement learning (RL) or deep RL (DRL) [9] has been adopted to address optimization problems in heterogeneous networks. Chen *et al.* [10] designed a centralized Q-learning algorithm for heterogeneous cellular networks, where a central controller gradually improves the global traffic offloading policy for small cells based on collected information. To maximize the overall designed quality of experience (QoE) metric or mean opinion score (MOS) in heterogeneous networks, centralized RL approaches for the optimization of traffic-aware resource allocation strategy were developed [11], [12], however, centralized approaches have the expensive cost of global information sharing and privacy information maybe leaked during this process.

In [13], [14], distributed resource optimization solutions based on multi-agent RL (MARL) were developed to jointly optimize the device association and resource allocation policy in dense heterogeneous networks, through enabling multiple agents to interact with unknown wireless environment. Gu *et al.* [15] and Alnwaimi *et al.* [16] proposed distributed strategies based on RL to enable each BS to operate individually configuration/optimization, which significantly reduce the expensive cost of knowing global information. Furthermore, in [17] and [18], the authors investigated energy-efficient resource management for interference channels in two-layer heterogeneous networks, and distributed RL aided power allocation algorithms were developed to improve the network energy efficiency while guaranteeing devices' QoS requirements. Asheralieva *et al.* [19] proposed a novel Bayesian RL framework to address distributed resource sharing problem in heterogeneous cellular networks.

In this paper, we investigate a joint device association, spectrum allocation, and power allocation optimization problem in two-tier heterogeneous wireless networks. As the network is

978-1-6654-2854-5/21/$31.00 ©2021 IEEE

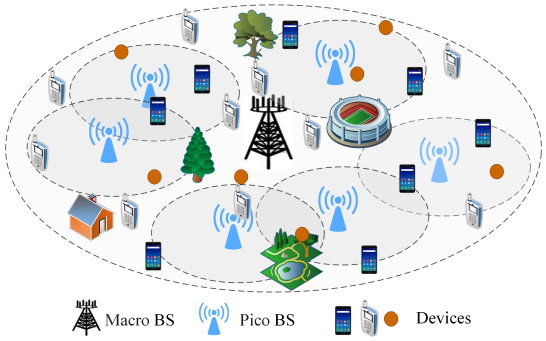

Fig. 1. A two-tier heterogeneous wireless network.

complex and dynamic with large state and action spaces, by using the tools of DRL along with the distributed coordinated learning mechanism, a distributed coordinated multi-agent dueling deep Q-network (DC-MA-DDQN) is proposed to learn the optimized intelligent resource management policy with fast convergence speed. Moreover, we evaluate and compare its superior performance to other existing learning algorithms in terms of network capacity and QoS satisfaction probability.

The rest of this paper is organized as follows. The two-tier heterogeneous network model and problem formulation are provided in Section II. The intelligent resource management based on DC-MA-DDQN algorithm is proposed in Section III. Section IV presents simulation results and analysis. Finally, we conclude this paper in Section V.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, the considered two-tier heterogeneous network is composed of macrocell tier and picocell tier, and each tier is classified based on its coverage range. The macrocell tier is associated with a macro BS (MBS) and it can support a large coverage area as shown in Fig. 1. The picocell tier is associated with multiple pico BSs (PBSs) with smaller coverage areas. As illustrated in Fig. 1, in general, as PBSs are non-uniformly distributed within the macrocell coverage area, a part number of picocells have overlapped area with each other while other picocells may have no overlapped area. In this context, the mobile devices located on the overlapped area suffer ICI from other nearby picocells on the same spectrum.

We consider a downlink transmission scenario with one MBS and C PBSs. There are K mobile devices randomly located in the network. The set of devices and PBSs are respectively denoted by $\mathcal{K} = \{1, 2, ..., K\}$ and $\mathcal{C} = \{1, 2, ..., C\}$. Both the MBS and PBSs provide wireless services for mobile devices by using orthogonal frequency division multiple access (OFDMA) technique, where the bandwidth is equally divided into N subchannels with a set of $\mathcal{N} = \{1, 2, ..., N\}$. Here, the unity frequency reuse (UFR) mechanism is used in two-tier heterogeneous networks to improve the spectrum utilization efficiency, where the communication band is reused across all cells. Here, let $g^{\mathrm{p}}_{c,k}$ denote the channel gain between the c-th PBS and the k-th device, and g^{m}_{k} denote the channel gain from the MBS to the k-th device.

In the picocell tier, if the k-th device is associated with the c-th PBS on the n-th subchannel ($n \in \mathcal{N}$), its received signal-to-interference-plus-noise ratio (SINR) can be expressed as

$$\gamma^{\mathrm{p}}_{c,k,n} = \frac{P^{\mathrm{p}}_{c,n} g^{\mathrm{p}}_{c,k}}{\sum_{c' \in \mathcal{C}} P^{\mathrm{p}}_{c',k,n} g^{\mathrm{p}}_{c',k} + P^{\mathrm{m}}_{n} g^{\mathrm{m}}_{k} + \delta^2_k}, \tag{1}$$

where $P^{\mathrm{p}}_{c,n}$ denotes the transmit power allocated on the n-th subchannel at the c-th PBS, and δ^2_k denotes the background noise power. In (1), if the device is located on the overlapped area, it also suffers the inter-tier interference ($P^{\mathrm{m}}_{n} g^{\mathrm{m}}_{k}$) from the macrocell tier in addition to the received ICI ($\sum_{c' \in \mathcal{C}} P^{\mathrm{p}}_{c',k,n} g^{\mathrm{p}}_{c',k}$) from the adjacent cells in the same tier when it operates in the same band as the macrocell tier.

In addition, if the k-th device is associated with the MBS on subchannel n, its received SINR can be written as

$$\gamma^{\mathrm{m}}_{k,n} = \frac{P^{\mathrm{m}}_{n} g^{\mathrm{m}}_{k}}{\sum_{c \in \mathcal{C}} P^{\mathrm{p}}_{c,k,n} g^{\mathrm{p}}_{c,k} + \delta^2_k}, \tag{2}$$

where P^{m}_{n} denotes the transmit power allocated on the n-th subchannel at the MBS. In (2), for the k-th device, it may suffer the inter-tier interference ($\sum_{c \in \mathcal{C}} P^{\mathrm{p}}_{c,k,n} g^{\mathrm{p}}_{c,k}$) from the picocell tier when it locates in the overlapped area and operates in the same band.

Accordingly, the instantaneous achievable rate in Mbps of the k-th device on the n-th subchannel at the picocell tier or the macrocell tier can be respectively expressed as

$$R^{\mathrm{p}}_{c,k,n} = B_{\mathrm{sub}} \log_2 \left(1 + \gamma^{\mathrm{p}}_{c,k,n}\right), \tag{3}$$

$$R^{\mathrm{m}}_{k,n} = B_{\mathrm{sub}} \log_2 \left(1 + \gamma^{\mathrm{m}}_{k,n}\right), \tag{4}$$

where B_{sub} is the subchannel bandwidth, and $B_{\mathrm{sub}} = B/N$ with B being the available system bandwidth. Here, equal bandwidth is assumed among subcannels for simplicity.

B. Problem Formulation

The achievable rate of device k over its allocated subchannels with its associated BS is expressed as

$$R_k = \sum_{c \in \mathcal{C}} x_k \sum_{n \in \mathcal{N}} \rho_{k,n} R^{\mathrm{p}}_{c,k,n} + (1 - x_k) \sum_{n \in \mathcal{N}} \rho_{k,n} R^{\mathrm{m}}_{k,n}, \tag{5}$$

where x_k is the BS association indicator that it associates with an PBS or an MBS, and it has a binary value of "1" or "0". $\rho_{k,n}$ is also a binary variable, i.e., $\rho_{k,n} \in \{0, 1\}$, $\rho_{k,n} = 1$ indicates that the k-th device is allocated on subchannel n, otherwise, $\rho_{k,n} = 0$.

In our considered two-tier heterogeneous network, mobile devices have QoS requirements. Let $R^{\min}_{k}$ denotes the minimum rate threshold, and the QoS requirement of the k-th device can be given by

$$R_k \geq R^{\min}_{k}, \forall k \in \mathcal{K}. \tag{6}$$

It is inefficient to apply traditional optimization algorithms to search for the optimal policy. Model-free RL is a dynamic programming which can be used to address the problem by

formulating the optimization problem as an MARL [9], where each BS acts as a learning agent.

We adopt MARL to model the optimization problem, which enables each BS to dynamically perform action selection to maximize the long-term benefit. Generally, in MARL, from the view point of the two-tier heterogeneous network, the network acts as an environment which is comprised of multiple picocells and the macrocell. In addition, the MARL can be defined by $\langle \mathcal{J}, \mathcal{S}, \mathcal{A}, \mathcal{P}, r \rangle$ with the help of Markov decision process (MDP) concept, where $\mathcal{J}$ is the set of participating agents, $\mathcal{S}$ denotes a set of network states, $\mathcal{A}$ represents a set of available actions, $\mathcal{P}$ is the state transition probability, and r denotes the reward function of the network. All these mentioned elements are described in detail as follows:

Agent: Multiple agents (i.e., one MBS and C PBSs) participate the learning process in the two-tier heterogeneous network.

State space: We define the state $s_t \in \mathcal{S}$ as one network state and define $s_{k,t}$ as the state of the k-th device. The state $s_{k,t}$ of each device consists of four parts: the previous received SINR values on its allocated subchannels $\gamma_k(t-1) = \{\gamma_{k,n}(t-1)\}_{n\in\mathcal{N}}$, the previous BS selection indicators $x_k = \{x_{k,j}(t-1)\}_{j\in\mathcal{J}}$, the previous subchannel allocation indicators $\rho_k(t-1) = \{\rho_{k,n}(t-1)\}_{n\in\mathcal{N}}$, and the current channel information $g_k(t) = \left\{g^{\mathrm{p}}_{k,c}(t), \mathrm{or}, g^{\mathrm{m}}_{k}(t)\right\}_{c\in\mathcal{C}}$. Hence, the state $s_{k,t}$ of each device and the overall network state s_t at the t-th time slot can be respectively described as

$$\begin{aligned} &s_{k,t} = \{\gamma_k(t-1), x_k(t-1), \rho_k(t-1), g_k(t)\}, \\ &s_t = \left\{\{s_{k,t}\}_{k\in\mathcal{K}}\right\}. \end{aligned} \tag{7}$$

As we consider the MARL framework in the network, each agent j (i.e., BS) only observes its network state from the environment, which is expressed as $s_{j,t} = \{\{s_{k,t}\}_{k\in\mathcal{K}_j}\}$, where $\mathcal{K}_j$ denotes all the associated devices in the j-th agent's cell.

Action space: At the t-th slot, the learning agent j chooses an action $a_{j,t} \in \mathcal{A}_j$ based on its observed state $s_{j,t}$, and it contains the BS association indicator $x_{j,t} = \{x_k(t)\}_{k\in\mathcal{K}_j}$, the subchannel allocation $\rho_{j,t} = \{\rho_{k,n}(t)\}_{n\in\mathcal{N},\, k\in\mathcal{K}_j}$, and the power allocation $P_{j,t} = \left\{P^{\mathrm{p}}_{c,n}(t), \mathrm{or}, P^{\mathrm{m}}_{n}(t)\right\}_{c\in C, n\in\mathcal{N}}$, that is

$$a_{j,t} = \{x_{j,t},\ \rho_{j,t},\ P_{j,t}\}. \tag{8}$$

Transition probability: $\mathcal{P}(s'|s,a)$ shows the state transition probability from the current state s to a next state s' by implementing an action a.

Reward function: In our framework, the reward function includes two parts, namely, the network data rate of all associated devices, and the QoS requirements (i.e., minimum data rate requirements) of these devices. Before building the reward function, let us define p_k^{outage} as the violated QoS constraint, which is mathematically expressed as

$$p_k^{\mathrm{outage}} = \begin{cases} 1, \text{ if } R_k < R_k^{\min}, \\ 0, \text{ otherwise.} \end{cases} \tag{9}$$

In (9), $p_k^{\mathrm{outage}} = 1$ if the QoS constraint of the k-th device is violated; otherwise, $p_k^{\mathrm{outage}} = 0$. According the above analysis, considering the QoS requirements, the QoS-aware reward function of the j-th agent (i.e., BS) with all the associated device set $\mathcal{K}_j$ is defined as

$$r_t = \sum_{k\in\mathcal{K}_j} R_k - \lambda \sum_{k\in\mathcal{K}_j} p_k^{\mathrm{outage}} \tag{10}$$

where λ is a positive weight of the reward function, and it is utilized to balance the revenue ($\sum_{k\in\mathcal{K}_j} R_k$) and the penalty ($\sum_{k\in\mathcal{K}_j} p_k^{\mathrm{outage}}$). In (10), the penalties increase if the QoS requirements of devices are violated frequently.

In order to achieve good performance in the long term, both the immediate reward and the long-term rewards are considered in RL. The goal of each agent is to find an optimal policy π^* (π is a function mapping from sates in to probabilities of choosing each available action in $\mathcal{A}$) to maximize the expected cumulative discounted reward, i.e.,

$$U_t = \sum_{i=0}^{\infty} \xi^i r_{t+i+1}, \tag{11}$$

where $\xi \in (0,1)$ denotes the discount factor.

III. Intelligent Resource Management Based on MARL

In this section, the coordinated multi-agent DDQN framework for device association, spectrum and power allocation in two-tier heterogeneous networks is proposed.

In RL, the expected reward is generally defined as the state-action value function $Q^\pi(s,a)$ and it is given as

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \xi^i r_{t+i+1} | s_t = s, a_t = a\right]. \tag{12}$$

The function $Q^\pi(s,a)$ satisfies the Bellman equation [13]

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[r(s,a) + \xi \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a) \sum_{a'\in\mathcal{A}} \pi(s',a') Q^\pi(s',a')\right]. \tag{13}$$

Q-learning is widely used to learn the optimal policy π^*, in order to achieve the optimal Q-function which is given by

$$Q^*(s,a) = r(s,a) + \xi \sum_{s'\in S} P(s'|s,a) \max_{a'\in\mathcal{A}} Q^*(s,a). \tag{14}$$

Note that Q-learning can effectively search for the optimal policy when the state space and action spaces are small. However, it is practically inapplicable in our presented multi-tier heterogeneous network where the state and action spaces are large, leading to the slow convergence if we employ the Q-learning algorithm. In the sequel, leveraging the DRL, we can exploit a deep Q-network (DQN) $Q(s,a;\boldsymbol{\theta})$ to estimate the Q-function, e.g., $Q(s,a) \approx Q(s,a;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the deep neural network (DNN) parameter. In order to stabilize the

overall learning performance, the DQN updates its parameter $\boldsymbol{\theta}$ by minimizing the following loss function defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(y^{DQN} - Q(s,a;\boldsymbol{\theta})\right)^2\right], \tag{15}$$

where $y^{DQN} = r + \xi \max_{a' \in \mathcal{A}} \hat{Q}(s', a'; \boldsymbol{\theta}^-)$, and $\boldsymbol{\theta}^-$ represents the DNN parameter of the target Q network $\hat{Q}(s,a;\boldsymbol{\theta}^-)$.

In DDQN, instead of directly evaluating the action-value function, the value function and advantage function can be separately estimated by using the two streams of fully connected layers. The value function is used to evaluate the quality of the learning policy with a certain state, while the advantage function captures the relationship between a certain action and other actions, and finally these functions are combined together in the final layer to generate the Q-value [20].

In DDQN, when a policy π is given, the value function of $Q^\pi(s,a)$ of a state-action pair (s,a) and the value function $V^\pi(s)$ of a state s can be respectively expressed as

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}\left\{U_t | s_t = s, a_t = a, \pi\right\}, \\ V^\pi(s) &= \mathbb{E}_{a\sim\pi(s)}\left\{Q^\pi(s,a)\right\}. \end{aligned} \tag{16}$$

According to the value functions of (16), the advantage function of actions is defined as

$$W^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{17}$$

In (17), note that the function $V^\pi(s)$ is used to measure the quality of its performance in a specific state s and the function $Q^\pi(s,a)$ evaluates the Q-value of selecting an action a in the state s. It is worth noting that $\mathbb{E}_{a\sim\pi(s)}\left\{W^\pi(s,a)\right\} = 0$ [20]. Moreover, when a deterministic policy $a^* = \arg\max_{a' \in \mathcal{A}} Q(s,a')$ is given, we have $Q(s,a^*) = V(s)$ and hence $W(s,a^*) = 0$.

In order to estimate the value functions of $V(s)$ and $W(s,a)$, a dueling neural network is generally used to make one stream of connected layers outputs a scalar $V(s;\boldsymbol{\theta},\mu)$ and other streams output an dimensional vector $W(s,a;\boldsymbol{\theta},\alpha)$, where α and μ denote the weights of the fully-connected layers. After combining these streams, the output of the dueling neural network can be expressed as

$$Q(s,a;\boldsymbol{\theta},\alpha,\mu) = V(s;\boldsymbol{\theta},\mu) + W(s,a;\boldsymbol{\theta},\alpha). \tag{18}$$

However, $Q(s,a;\boldsymbol{\theta},\alpha,\mu)$ is only a parameterized approximator of the accurate Q-function. Moreover, given Q-function, we cannot recover $A(s,a)$ and $V(s)$ uniquely, which causes an unidentifiability of (19) and thus the poor performance. In other words, adding a constant to $V(s;\boldsymbol{\theta},\mu)$ and subtract the same constant from $W(s,a;\boldsymbol{\theta},\alpha)$ results in the same Q-value. To address this problem, the dueling neural network can be implemented by the following mapping

$$\begin{aligned} Q(s,a;\boldsymbol{\theta},\alpha,\mu) = V(s;\boldsymbol{\theta},\mu) + (W(s,a;\boldsymbol{\theta},\alpha) - \\ \arg\max_{a' \in \mathcal{A}} W(s,a';\boldsymbol{\theta},\alpha)). \end{aligned} \tag{19}$$

The objective of (19) is to make the advantage function approximator to have zero advantage when selecting actions. Note that given $a^* = \arg\max_{a' \in \mathcal{A}} Q(s,a';\boldsymbol{\theta},\alpha,\mu) = \arg\max_{a' \in \mathcal{A}} W(s,a';\boldsymbol{\theta},\alpha)$, we can achieve $Q(s,a^*;\boldsymbol{\theta},\alpha,\mu) = V(s;\boldsymbol{\theta},\mu)$. Thus, the stream $V(s;\boldsymbol{\theta},\mu)$ estimates the value function of state s, while other streams provide an estimator of the advantage function. In addition, (19) can be transformed into a simple alternative module by replacing the maximum value with an average, i.e.,

$$\begin{aligned} Q(s,a;\boldsymbol{\theta},\alpha,\mu) = V(s;\boldsymbol{\theta},\mu) + (A(s,a;\boldsymbol{\theta},\alpha) - \\ \frac{1}{|\mathcal{A}|}\sum_{a'} A(s,a';\boldsymbol{\theta},\alpha). \end{aligned} \tag{20}$$

According to the above analysis, the proposed distributed coordinated multi-agent DDQN (called DC-MA-DDQN) algorithm for intelligent resource management in the two-tier heterogeneous network is shown in **Algorithm 1**. Note that when a picocell has no overlapped coverage area with any other picocell in the same tier, the action of the agent is selected independently based on its own local observed information without sharing its decision with other picocells. The reason lies in the fact that the decision or action of the independent cell does not affect the reward or strategy of other picocells as they have no overlapped area [22]. Only when the picocell has overlapped coverage area with neighbour cells, its action (i.e., subchannel allocation and power allocation) information for those devices located on overlapped area will be shared with neighbour cells to minimize the overall ICI.

Algorithm 1 DC-MA-DDQN Based Intelligent Resource Management

Input: Two-tier heterogeneous network simulator.
Initialize experience relay memory $\mathcal{D}$ to capacity D, and a mini-batch.
Initialize the primary DDQN Q with parameters α and μ.
Initialize the target DDQN $\hat{Q}$ as a copy of the DDQN with parameters $\alpha^- = \alpha$ and $\mu^- = \mu$.
for each episode = 1, 2, ..., I **do**
  Each agent observe its initial network state s;
  for each time step t = 0, 1, 2, ..., T **do**
    Choose an action a_t at the state s using ϵ-greedy policy:
    $a_t = \arg\max_{a \in \mathcal{A}} Q(s,a;\boldsymbol{\theta},\alpha,\mu)$, with probability $1-\epsilon$;
    $a_t = \text{random}\{a_j\}_{a_j \in \mathcal{A}}$, with probability ϵ;
    Perform the action a_t, and receive a reward r_t and a next state s_{t+1};
    Calculate two streams of DDQN value functions, including $V(s_t;\boldsymbol{\theta},\mu)$ and $A(s_t,a_t;\boldsymbol{\theta},\alpha)$, and combine them into $Q(s_t,a_t;\boldsymbol{\theta},\alpha,\mu)$ by (21);
    Update loss function $\hat{\mathcal{L}}(\boldsymbol{\theta},\alpha,\mu)$ and calculate parameter $\boldsymbol{\theta}$ using gradient descent by (26);
  end for
end for
Each cell or BS share its action with nearby cells or BS.
Return the DDQN learning model.

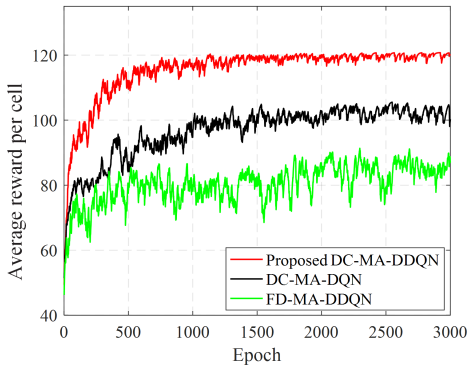

Fig. 2. Algorithm convergence comparisons.

IV. Simulation Results and Discussions

This section evaluates and analyzes the performance of the proposed distributed coordinated learning algorithm for two-tier heterogeneous networks. As shown in Fig. 1, an MBS is located at the center of a macrocell with a covered radius of 250 m, while 6 PBSs are distributed in the coverage area of the macrocell with each picocell being a covered radius of 80 m. Note that as shown in Fig. 1, two groups of picocells have overlapped area so that the picocells in each pair can coordinate with each other to maximize their sum rewards, while one remaining picocell does not have overlapped area with other picocells so that it can perform action selection independently. The path loss in dB between the MBS and the mobile device can be seen in [7]. Tradoff weight in (11) is 10, the minimum data rate requirement is 4.5 Mbps, noise spectral density is -174 dBm/Hz, the number of subchannels per cell or BS is 16, the transmit power of MBS is {25, 30, 35, 40} dBm, and the he transmit power of PBS is $\{20, 25, 30\}$ dBm.

The DDQN model is trained to employ empirically hyper-parameter, where deep neural networks (DNN) trains for 3000 epochs with 64 mini-batches being used in every epoch. The DNN architecture that we adopted has three hidden layers, include 512, 512, and 256 neurons, respectively. We set the learning rate as $\alpha = 0.001$ and the discount factor as $\gamma = 0.98$. We initialize the exploration rate ε as 0.8 at the beginning, and it then gradually anneals from 0.8 to 0.005 over the first 1000 episodes and remains constant afterwards.

We compare the performance of the proposed distributed coordinated MA-DDQN algorithm (denoted by DC-MA-DDQN) with two following algorithms in the two-tier heterogeneous network:

1) The distributed coordinated DQN based intelligent resource management algorithm (denoted by DC-MA-DQN), similar to the work [28].

2) The fully distributed MA-DDQN algorithm based intelligent resource management (denoted by FD-MA-DDQN), similar to the learning idea in [32], where each cell selfishly maximizes its own received reward without considering the generated interference to other cells.

Fig. 2 shows the convergence of the three algorithms in terms of the average reward per cell when the number of devices is $K = 200$. It can be seen that the three algorithms can gradually converge to final levels with different numbers

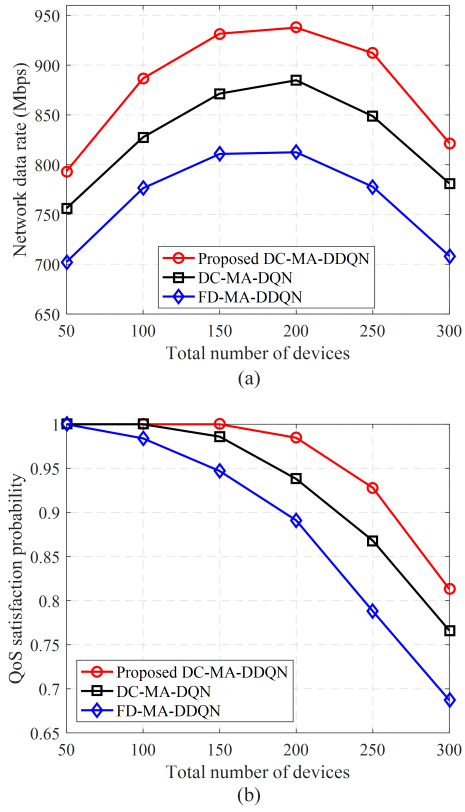

Fig. 3. Performance comparisons versus different total numbers of devices.

of epochs despite some fluctuations due to dynamic environment characteristic and policy exploration. Among the three algorithms, the proposed DC-MA-DDQN algorithm achieves the best convergence rate and the highest average reward. For DC-MA-DDQN and DC-MA-DQN, we can find that the convergence rate of DC-MARL-DQN is approximately similar to that of DC-MA-DDQN, and both of them converge at about 1500 epochs. However, at that time, the average reward calculated by DC-MA-DDQN is better than that of DC-MA-DQN, and it does not reach the final optimized level yet. Thus, when DC-MA-DDQN converges, the optimized reward of the two-tier heterogeneous network is superior to that of DC-MA-DQN. Besides, the FD-MARL-DDQN algorithm achieves the worst reward among the three algorithms, because each cell only aims to selfishly maximize its own received reward without considering the interference generated to other nearby cells.

Fig. 3 plots the network data rate and the QoS satisfaction probability performances of the three algorithms with the various total numbers of devices K. From Fig. 3(a), with the growing number of devices K, we can observe that the network data rate enhances obviously to a peak when K is smaller than a certain value $K \leq 200$, due to the high probability of finding devices having high channel gains to improve the network data rate (this phenomenon also called device diversity). However, further increasing the number of devices after the capacity reaches a maximum value. We observe that the data rate decreases, because the network resource requires to be allocated to the devices with poor

channel quality to meet their QoS requirements which occupy lots of network resource, as well as channel access collisions increase with K, hence reducing the network data rate. In addition, as shown in Fig. 3(b), for all DRL algorithms, the QoS satisfaction probability is favorable when the number of devices is small, but it declines clearly when K is large. The reason lies in that the two-tier heterogeneous network cannot complete all increased services due to the fixed spectrum and power resource.

However, when the number of devices K is large, the performance of all algorithms decreases, but the proposed DC-MA-DDQN algorithm still has the best performance among all algorithms. Moreover, the performance gap between them increases for increasing the number of devices K. Besides, as illustrated from Fig. 3(a), although the DC-MA-DQN algorithm tends to achieve the comparable network data rate performance to the proposed DC-MA-DDQN algorithm with the increased number of devices, DC-MA-DDQN has better performance than that of DC-MA-DQN, especially that the QoS satisfaction probability is significantly improved in dense device region. For example, when $K = 200$, DC-MA-DDQN achieves the network data rate and QoS satisfaction probability improvements of 7.57% and 6.91% compared with DC-MA-DQN, and achieves the improvements of 17.27% and 17.77% compared with FD-MA-DDQN.

V. Conclusions

In this paper, the distributed coordinated multi-agent DRL algorithm has been proposed to achieve the joint device association, spectrum allocation, and power allocation strategy for two-tier heterogeneous wireless networks. As the network is complex and dynamic with huge state and action spaces, by employing the tools of DRL along with the distributed coordinated learning mechanism, a DC-MA-DDQN algorithm has been developed to learn the optimal intelligent resource management policy with fast convergence speed. Simulations have demonstrated the superior performance of the proposed DC-MA-DDQN algorithm compared to other popular DRL algorithms in terms of the improvement of the network capacity and QoS satisfaction probability. We will apply this learning model to address the optimization problem in intelligent reflecting surface 6G networks in the future [22].

VI. Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, the SUTD SRG-ISTD-2021-165, the Macau Science and Technology Development Fund, under Grant 0119/2020/A3 and 0108/2020/A, and the Guangdong NSF under Grant 2021A1515011900. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] H. Zhang, L. Song, Y. Li, and G. Y. Li, "Hypergraph theory: Applications in 5G heterogeneous ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 70-76, Dec. 2017.

[2] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Network*, vol. 34, no. 6, pp. 272-280, November/December 2020.

[3] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[4] A. Khalili, S. Akhlaghi, H. Tabassum, and D. W. K. Ng, "Joint user association and resource allocation in the uplink of heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 804-808, Jun. 2020.

[5] S. Jabeen and P. Ho, "A benchmark for joint channel allocation and user scheduling in flexible heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9233-9244, Sept. 2019.

[6] T. Kim and J. M. Chang, "QoS-aware energy-efficient association and resource scheduling for HetNets," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 650-664, Jan. 2018.

[7] Q. Ye, W. Zhuang, S. Zhang, A. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896-9910, Oct. 2018.

[8] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1809-1822, Mar. 2017.

[9] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 1998.

[10] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 627-640, Apr. 2015.

[11] A. Xiao, X. Huang, S. Wu, H. Chen, and L. Ma, "Traffic-aware rate adaptation for improving time-varying QoE factors in mobile video streaming," Appear in *IEEE Trans. Network Science and Engin.*, 2020. Doi: 10.1109/TNSE.2020.3013533.

[12] F. S. Mohammadi and A. Kwasinski, "QoE-driven integrated heterogeneous traffic resource allocation based on cooperative learning for 5G cognitive radio networks," in *Proc. IEEE 5G World Forum (5GWF)*, Silicon Valley, CA, 2018, pp. 244-249.

[13] D. Li, H. Zhang, K. Long, W. Huangfu, J. Dong, and A. Nallanathan, "User association and power allocation based on Q-learning in ultra dense heterogeneous networks" in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1-5.

[14] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141-5152, Nov. 2019.

[15] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3996-4012, Sept. 2016.

[16] G. Alnwaimi, S. Vahid, and K. Moessner, "Dynamic heterogeneous learning games for opportunistic access in LTE-based macro/femtocell deployments," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2294-2308, Apr. 2015.

[17] J. Wang, C. Jiang, K. Zhang, X. Hou, Y. Ren, and Y. Qian, "Distributed Q-learning aided heterogeneous network association for energy-efficient IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2756-2764, Apr. 2020.

[18] A. H. Arani, A. Mehbodniya, M. J. Omidi, F. Adachi, W. Saad, and I. Güvenç, "Distributed learning for energy-efficient resource management in self-organizing heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9287-9303, Oct. 2017.

[19] A. Asheralieva, "Bayesian reinforcement learning-based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5016-5032, Aug. 2017.

[20] N. Van Huynh, D. Thai Hoang, D. N. Nguyen, and E. Dutkiewicz, "Optimal and fast real-time resource slicing with deep dueling neural networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1455-1470, Jun. 2019.

[21] H. Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977-2990, May 2021.

[22] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375-388, Jan. 2021.