Learning-Based Joint User-AP Association and Resource Allocation in Ultra Dense Network

Zhipeng Cheng[†], Minghui LiWang[†*], Ning Chen[†], Hongyue Lin[†], Zhibin Gao[†], and Lianfen Huang[†]

[†]Dept. of Information and Communication Engineering, Xiamen University, Xiamen, China
[*]Dept. of Electrical and Computer Engineering, University of Western Ontario, London, Canada
Email:{chengzhipeng@stu.xmu.edu.cn, mliwang@uwo.ca, lfhuang@xmu.edu.cn}

Abstract—With the advantages of Millimeter wave in wireless communication network, the coverage radius and inter-site distance can be further reduced, the ultra dense network (UDN) becomes the mainstream of future networks. The main challenge faced by UDN is the serious inter-site interference, which needs to be carefully addressed by joint user association and resource allocation methods. In this paper, we propose a multi-agent Q-learning based method to jointly optimize the user association and resource allocation in UDN. The deep Q-network is applied to guarantee the convergence of the proposed method. Simulation results reveal the effectiveness of the proposed method and different performances under different simulation parameters are evaluated.

Index Terms—User Association, Resource Allocation, Ultra Dense Network, Multi-agent Q-learning.

I. INTRODUCTION

As one of the key technologies of 5G, Ultra-Dense Network is used to densely deploy a large number of small base stations (SBSs) in hotspots to increase capacity and achieve seamless coverage[1]. However, the dense deployment of SBSs complicates the problems of user association, radio resource allocation, interference control and mobility management [2]. In UDN, SBSs are deployed in overlapping manner, users can choose to be associated with several adjacent SBSs via multi-connectivity solutions [3]. The system performance is greatly influenced by the user association patterns. With the increasing deployment density of SBSs, the network topology becomes very complicated. Moreover, a large number of interference sources with very close signal strength bring huge interference to users. This requires a better resource allocation strategy for interference control.

Recently, several research works have devoted to tackle the user association and resource allocation in UDN [4]-[11]. In [4], a novel modularity-based user-centric (MUC) clustering is proposed for resource allocation in UDN to maximize the sum rate per resource block. The basic idea of MUC clustering is to decompose the UDN into several subnetwork by the group structure of users. A energy-efficient (EE) resource allocation strategy in UDN is presented in [5], the EE optimization problem is decomposed into two sub-optimization problems of sub-channel allocation and power allocation. These two problem are solved by a two-stage Stackelberg game with a uniform pricing scheme. The Millimeter Wave (mmWave)-based UDN is considered in [6] and [7]. In [6], the joint user association and resource allocation problem are modeled as a mixed-integer programming problem, which take multiple factors (e.g., load balance, user quality of service, EE and cross-tier interference) into consideration. In [7], the joint user association and resource allocation problem are considered in mmWave self-backhauling UDN, a coalition game based algorithm is proposed to maximize network sum rate. Similarly, a joint power allocation and user association strategy using non-cooperative game theory is developed in [8]. The joint user association and resource allocation problem are considered in [9], [10] and [11]. In [9], a unified non-orthogonal multiple access (NOMA) in UDN is proposed, which focuses on the user association and resource allocation. Two case studies are presented to demonstrate the effectiveness of the framework. Joint optimization of user association and dynamic time division duplexing (TDD) for UDN are studied in [10]. Authors decompose the problem into separate subproblems that can be solved in a distributed manner and prove convergence to the global optimum. The more similar to our work is [11], they propose a novel method for user association and resource allocation based on coordinated multipoint (CoMP). A cell-filtering and location-load based clustering methods are used to reduce network complexity. Then a competition-based resource allocation scheme is proposed based on the results of clustering.

In the light of previous works, we focus on the joint optimization of user-AP association and resource allocation in UDN rather than solve the joint problem in a separate manner. Due to the complexity of this joint problem, we propose a learning-based joint optimization algorithm. The main contributions of this paper can be summarized as follows:

- We propose a multi-agent Q-learning based solution to solve the joint problem of user-AP association and resource allocation in UDN.
- We apply deep Q-network to avoid the curse of dimensionality and accelerate convergence.
- Demonstrate the effectiveness of the proposed method with simulation results compared with other methods.

The remainder of the paper is organized as follows: In Section II, we present the system model and the joint problem of user-AP association and resource allocation is formulated. In Section III, the multi-agent Q-learning based solution is

Fig. 1. The example layout of UDN.

proposed. Simulation results are shown in Section IV and we conclude the paper in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Network model

N We consider a typical UDN composed of M access points (APs) and N users. All APs are identical in terms of coverage radius, antenna gain, pathloss model and maximum transmission power. The set of APs and users are denoted as $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$, respectively. All APs can operate on L orthogonal subcarriers, the set of subcarriers is denoted as $\mathcal{L} = \{1, 2, \ldots, L\}$. Assume that the user can access at least one AP and a maximum of k APs, each AP can serve up to f users. For the simplicity of discussion, we assume that each user can only choose at most one subcarrier from one AP at any time and can only access to its neighboring APs. Thus, we define the candidate AP set for user i as $S_i = \{j | d_{i,j} \leq r, i \in \mathcal{N}, j \in \mathcal{M}\}$, where $d_{i,j}$ is the distance between user i and AP j, r is the coverage radius of the AP. Similarly, we define the candidate user set for AP j as U_j. The maximum transmission power of the AP is P_{ap}.

Since all APs are closely deployed, the co-subcarrier interference should be considered, the signal to interference plus noise ratio (SINR) at the user i with AP j using subcarrier l at time t is

$$\gamma_{i,j}^l(t) = \frac{P_j G_{i,j}}{\sum_{j' \in M \setminus \{j\}} P_{j'} G_{i,j'} + N_0} \tag{1}$$

where $P_j = P_{ap}/n_j$ is the transmission power of AP j, where n_j is the number of users associated with AP j. $G_{i,j}$ is the channel gain from AP j to user i, which incorporates antenna gain, path loss and shadow fading. We consider a flat fading on all subcarriers and thus the channel gains are same from an AP to a user on all subcarriers. N_0 is the noise power on the subcarrier of bandwidth W.

At any time t, the transmission capacity of user i with AP j on subcarrier l can be denoted as

$$r_{i,j}^l(t) = W \log\left(1 + \gamma_{i,j}^l(t)\right) \tag{2}$$

B. Access point selection model

At any time t, all users make AP selection decisions. Here we take a binary indicator $x_{i,j}$ to denote the user-AP association pattern. Let $x_{i,j} = 1$ if user i is associated with AP j, otherwise $x_{i,j} = 0$. Note that at any time t, we have the following constraints for $x_{i,j}$ as:

$$\sum_{j \in S_i} x_{i,j} \leq k, \forall i \in \mathcal{N} \tag{3}$$

$$\sum_{i \in \mathcal{N}} x_{i,j} \leq f, \forall j \in \mathcal{M} \tag{4}$$

$$x_{i,j} \in \{0, 1\}, \forall i \in \mathcal{N}, \forall j \in S_i \tag{5}$$

where constraint (3) indicates that each user cannot be served more than k APs. Constraint (4) means that one AP can serve simultaneously up to f users.

C. Resource allocation model

As mentioned above, all subcarriers are shared by the APs, thus the co-subcarrier interference needs to be carefully addressed which greatly limits the system capacity of the UDN. How to effectively allocate L subcarriers to M APs will be the major issue. At any time t, users choose to occupy a subcarrier from an AP. Thus, a binary indicator $y_{i,j}^l$ is introduced to indicate the resource allocation strategy. If $y_{i,j}^l = 1$ means the subcarrier l is allocated to user i from the AP j, otherwise $y_{i,j}^l = 0$. The constraints for $y_{i,j}^l$ are as follows:

$$\sum_{l \in \mathcal{L}} y_{i,j}^l \leq 1, \forall i \in \mathcal{N}, \forall j \in S_i \tag{6}$$

$$y_{i,j}^l = 1 - y_{i',j}^l, \forall i, i' \in U_j, \forall j \in \mathcal{M} \tag{7}$$

$$\{y_{i,j}^l = y_{i,j'}^l | x_{i,j} = x_{i,j'}, \forall i \in \mathcal{N}, \forall j, j \in S_i\} \tag{8}$$

$$y_{i,j}^l \in \{0, 1\}, i \in \mathcal{N}, j \in S_i, l \in \mathcal{L} \tag{9}$$

where constraint (6) states that each user can only choose one subcarrier from its associated AP. Constraint (7) ensures that the user is associated with the same AP use orthogonal subcarriers. Constraint (8) indicates that the APs use the same subcarrier to serve the same user.

Thus, we have the total transmission capacity of the user i at time t as:

$$r_i(t) = \sum_{j \in S_i} \sum_{l \in \mathcal{L}} x_{i,j} y_{i,j}^l r_{i,j}^l(t), \forall i \in \mathcal{N} \tag{10}$$

D. Problem formulation

The main goal in this paper is to maximize the aggregate network utility while satisfy user's quality of service (QoS) requirements in the UDN. The joint problem of user-AP association and resource allocation is considered. The utility functions are defined for user i at time t as:

$$U\left(r_i(t)\right) = r_i(t) \tag{11}$$

The utility function is a linear function of user i's transmission capacity. As we aim to maximize the long term network utility, we define the long-term reward of user i R_i as the weighted sum of instantaneous rewards over a finite period T:

$$R_i = \sum_{t=1}^{T} \gamma^t U\left(r_i(t)\right) \tag{12}$$

where $\gamma \in [0,1)$ is the discount factor. Thus, the long-term reward maximization can be formulated as (13).

$$\max_{x_{i,j}, y_{i,j}^l} : \sum_{i=1}^{N} R_i \tag{13}$$
$$s.t. \ (3) - (9)$$

Note that due to the non-convex and combinatorial characteristics of the formulated problem, difficulties exist in obtaining a global optimal strategy of this joint problem. In the following section, the multi-agent Q-learning (MAQL) based solution is proposed.

III. MULTI-AGENT Q-LEARNING BASED SOLUTION

In this section, we first present the basic idea of MAQL method. Then, the deep Q-network is proposed to avoid the curse of dimensionality.

A. Multi-agent Q-learning method

In the joint optimization problem of user-AP association and resource allocation, we can model this problem as a discounted stochastic game. In this game, we assume N users are agents. This N agents stochastic game is defined by the tuple $(S, A_1, \ldots, A_N, r_1, \ldots, r_N, p)$. S is the state space, A_i and r_i are the agent i's action space and reward function. p is the state transition probability. According to the research of stochastic game model, the single-agent Q-learning is extended to a multi-agent scenario, which is called Nash Q-learning. Due to space constraints, refer to [12] for details.

At any time t, the user i will observe the current state of the environment and takes action. When all users have taken actions, they observe the reward and the new state, each user then updates its Q table according to

$$\begin{aligned} Q_i\left(s, a_i\right) = & Q_i\left(s, a_i\right) + \alpha\left[u_i\left(s, a_i, \pi_{-i}\right) + \right. \\ & \left. \gamma \max Q_i\left(s', a_i'\right) - Q_i\left(s, a_i\right)\right] \end{aligned} \tag{14}$$

where α is the learning rate, $u_i\left(s, a_i, \pi_{-i}\right)$ is the agent i's one-period reward in state s adopting the joint strategies.

As the number of APs and users are fixed in the UDN, if each user obtains the information about reward function and

state transition, the Nash equilibrium (NE) can be found to maximize the network utility through message passing within the finite time period T [13]. In the following, we define the agents, states, actions and reward function of our MAQL algorithm.

- **Agents:** All N users.
- **States:** At time t for user i, the state is defined as $s_{i,t} \in \{0,1\}$, indicates whether the user meets its QoS requirement:

$$s_{i,t} = \begin{cases} 1, & r_i(t) \geq r_{QoS} \\ 0, & r_i(t) < r_{QoS} \end{cases} \tag{15}$$

  where r_{QoS} is the minimum data rate to satisfy the QoS requirement of the user. The number of possible states is 2^N and this will be very large with large N. The state vector can be denoted as $S_t = \{s_{1,t}, s_{2,t}, \ldots, s_{N,t}\}$.
- **Actions:** At time t, each user can select up to k APs from the candidate AP set first, then occupy at most one subcarrier from every selected AP. Therefore, the number of possible actions for each user is $M * L$ with one-hot coding. However, the actual action space for AP selection depends on the candidate AP set S_i. Then, we define the action for user i as

$$a_{i,t} = \{m_{i,t}, l_{i,t}\} \tag{16}$$

  The action vector of N users can be denoted as $A_t = \{a_{1,t}, a_{2,t}, \ldots, a_{N,t}\}$.
- **Reward Function:** As we want to maximize the aggregate network utility, then the reward function can be define as $\Psi_t = \sum_{i=1}^{N} U\left(r_i(t)\right)$.

B. Multi-agent deep Q-network for the joint problem

As can be seen from above, the number of states and actions of the MAQL for the joint problem can be very large for a large N, M and L. Thus, it is no longer feasible to store the state-action pairs in a Q-table and deep Q-network (DQN) is a better method. The basic idea of DQN is to use the deep neural network (DNN) to represent action and state spaces. DQN takes the advantages of neural network to approximate the action-value function, and uses memory replay to improve the learning performance. Two different neural networks with the same structure, called target-network and evaluated network, are used in DQN. The parameters of the two networks are alternately updated every several steps to improve the learning stability. In each episode, the evaluated Q-network is trained to adapt its parameters to decrease the loss function as follows:

$$L_i(\theta) = E\left[\left(y_i - Q_i\left(s, a_i; \theta\right)\right)^2\right] \tag{17}$$

$$y_i = u_i\left(s, a_i\right) + \gamma \max_{a_i' \in \mathcal{A}_i} Q_i\left(s', a_i'; \theta^-\right) \tag{18}$$

where $L_i(\theta)$ is the loss function, y_i is the estimated Q-value of target network. θ, θ^- are the weights of evaluated network and target network, respectively.

The multi-agent DQN (MADQN) algorithm for the joint problem is summarized in Algorithm 1.

Algorithm 1 MADQN for joint User-AP Association and Resource Allocation

---

1: Initialize learning rate α, discount factor γ, exploration rate ϵ, maximum learning episode EP, maximum training steps T per episode.
2: Initialize the replay memory D, evaluated network $Q(s, a; \theta)$ parameters with random weight θ.
3: Initialize the target network $Q_i(s', a_i'; \theta^-)$ with weights $\theta^- = \theta$.
4: **for** episode=1 : EP **do**
5:     Initialize the network state s.
6:     **for** each step=1 : T **do**
7:         Each user takes action a_i using the ϵ-greedy policy from $Q_i(s, a_i; \theta)$.
8:         Each user obtains the immediate reward u_i and new state s', and let $s = s'$.
9:         Each user stores the transition (s, a_i, u_i, s') in D.
10:         Each user samples random minibatch of transitions (s, a_i, u_i, s') from D.
11:         Each user set y_i according to (18).
12:         Each user performs a gradient descent step on $(y_i - Q_i(s, a_i; \theta))^2$ with respect to the network parameters θ.
13:         Every C steps update $\theta^- = \theta$
14:     **end for**
15: **end for**

IV. PERFORMANCE EVALUATION

A. Simulation setup

In this Section, we conduct the simulation to evaluate the performance of our proposed scheme. We consider a simulation area with a length and width of 50 meters. The APs and users are uniformly distributed within the simulation area. The pathloss model for all APs is $PL = \alpha + 10\beta \log_{10}(d) + \xi$[dB] $\xi \sim \mathcal{N}(0, \sigma^2)$, d in meters. $\alpha = 72.0, \beta = 2.92, \sigma = 8.7$dB for NLOS and $\alpha = 61.4, \beta = 2, \sigma = 5.8$dB for LOS [14]. Other important simulation parameters are listed in table I.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Carrier frequency	28 GHz
Number of subcarriers L	4
Subcarrier bandwidth	180 KHz
Number of APs	10
Number of users	2:2:10
AP radius	15 m
AP transmission power/Gain	23 dBm/ 5 dBi
Maximum APs for one user k	4
Maximum users for one AP f	4
Noise power density	-174 dBm/Hz
minimum Qos data rate(r_{QoS})	2 Mbps

Fig. 2. Normalized reward for each episode. The AP number is 5 and the user number is 10.

The DQN for each learning agent consists of 3 fully connected hidden layers, containing 100, 200, and 50 neurons. The ReLU activation function and RMSProp optimizer are used [15]. We train the learning agent for a total of 400 episodes and 500 steps per episode with a learning rate 0.0001 and a exploration rate from 0.99 to 0.0001. We set the discount factor γ to be 0.9.

Due to space constraints, instead of using the state of the art as a comparison, a Max_RSRP based method is simulated as the baseline to compare the performance with our proposed MADQN method. Users choose to be associated with the k APs with the largest reference signal receiving power (RSRP) and randomly choose a subcarrier in the Max_RSRP based method. All simulation results are the average result of 50 different user and AP distributions.

B. Simulation results

Fig. 2 demonstrates the convergence behavior of our proposed MADQN method. As can be seen from the figure, the horizontal axis is the number of training episodes and we take the normalized reward for each episode as the vertical axis. The reward increases with the number of training episodes in the first 150 episodes. when the training episode approximately reaches 200 episodes, the performance gradually converges despite some fluctuations.

(a) Total throughput (b) Average user throughput

Fig. 3. The throughput versus the user number of two different methods.

The performances of the two different methods are drawn in Fig. 3. The total throughput and average throughput per user versus the user number of the two methods are compared. As can be seen from Fig. 3, our proposed MADQN

(a) Total throughput   (b) Average user throughput

Fig. 4. The throughput versus the user number under different number of APS.

(a) Total throughput   (b) Average user throughput

Fig. 5. The throughput versus the user number under different values of k and f.

method outperforms the Max_RSRP based method both in total throughput and average user throughput. In particular, the difference between the total throughput is more obvious as the number of users increases. The rationale behind this is that the interference of the Max_RSRP based method increases rapidly with the number of users.

We further analysis the performance of our proposed MADQN method under different simulation parameters. In Fig. 4, we evaluate the throughput versus the user number for different number of APs. It can be seen that both total throughput and average user throughput increase with the number of APs. The performance of different k and f are evaluated in Fig. 5. The larger k and f, the larger the throughput, but this growth is limited as the interference management becomes more complex.

V. CONCLUSION

In this paper, we propose a multi-agent Q-learning (MAQL) based method for the joint problem of user-AP association and resource allocation in ultra dense network. Furthermore, we use the deep Q-network to accelerate the convergence of MAQL. The simulation results demonstrate the feasibility and effectiveness of the proposed method. Moreover, we analysis the performance of our proposed method under different key simulation parameters. This enlightens us to further analyze the relationship between different parameters.

VI. ACKNOWLEDGMENT

REFERENCES

[1] W. Yu, H. Xu, H. Zhang, D. Griffith and N. Golmie, "Ultra-Dense Networks: Survey of State of the Art and Future Directions," *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, Waikoloa, HI, USA, Aug. 2016, pp. 1-10.

[2] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song and Y. Zhang, "Resource Allocation for Ultra-Dense Networks: A Survey, Some Research Issues and Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134-2168, thirdquarter 2019.

[3] F. B. Tesema, A. Awada, I. Viering, M. Simsek and G. P. Fettweis, "Evaluation of Adaptive Active Set Management for Multi-connectivity in Intra-frequency 5G Networks," *2016 IEEE Wireless Communications and Networking Conference*, Doha, Qatar, Apr. 2016, pp. 1-6.

[4] Y. Lin, R. Zhang, L. Yang and L. Hanzo, "Modularity-Based User-Centric Clustering and Resource Allocation for Ultra Dense Networks," in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12457-12461, Dec. 2018.

[5] L. Xu, Y. Mao, S. Leng, G. Qiao and Q. Zhao, "Energy-efficient Resource Allocation Strategy in Ultra Dense Small-cell Networks: A Stackelberg Game Approach," *2017 IEEE International Conference on Communications (ICC)*, Paris, France, May. 2017, pp. 1-6.

[6] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung and H. V. Poor, "Energy Efficient User Association and Power Allocation in Millimeter-Wave-Based Ultra Dense Networks With Energy Harvesting Base Stations," in *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1936-1947, Sept. 2017.

[7] Y. Liu, X. Fang, P. Zhou and K. Cheng, "Coalition game for user association and bandwidth allocation in ultra-dense mmWave networks," *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, Qingdao, China, Oct. 2017, pp. 1-5.

[8] A. Khodmi, S. B. Rejeb, N. Agoulmine and Z. Choukair, "A Joint Power Allocation and User Association Based on Non-Cooperative Game Theory in an Heterogeneous Ultra-Dense Network," in *IEEE Access*, vol. 7, pp. 111790-111800, Aug. 2019.

[9] Z. Qin, X. Yue, Y. Liu, Z. Ding and A. Nallanathan, "User Association and Resource Allocation in Unified NOMA Enabled Heterogeneous Ultra Dense Networks," in *IEEE Communications Magazine*, vol. 56, no. 6, pp. 86-92, Jun. 2018.

[10] N. Sapountzis, T. Spyropoulos, N. Nikaein and U. Salim, "Joint Optimization of User Association and Dynamic TDD for Ultra-Dense Networks," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, USA, Apr. 2018, pp. 2681-2689.

[11] W. K. Lai, C. Hsu and Y. Kuo, "QoS-guaranteed User Association and Resource Allocation with CoMP JT in Ultra-Dense Networks," *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, Matsue, Japan, Sept. 2019, pp. 1-6.

[12] J. Hu, and M. P. Wellman,"Nash Q-Learning for Genera-Sum Stochastic Games," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 1039-1069, 15. Aug. 2004.

[13] 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," in *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141-5152, Nov. 2019.

[14] M. R. Akdeniz et al., "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164-1179, Jun. 2014.

[15] L. Liang, H. Ye and G. Y. Li, "Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning," in *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282-2292, Oct. 2019.