2023 International Conference on Future Communications and Networks (FCN) | 979-8-3503-9603-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/FCN60432.2023.10543489

Deep Reinforcement Learning-Based Access Point Selection for Cell-Free Massive MIMO with Graph Convolutional Networks

Mingjun Du, Xinghua Sun, Wenhai Lin, and Wen Zhan
School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China
E-mail: {dumj3,linwh33}@mail2.sysu.edu.cn; {sunxinghua,zhanw6}@mail.sysu.edu.cn

***Abstract*—In this paper, we investigate the association problem between access points (APs) and users in cell-free massive multiple-input multiple-output (MIMO) systems. To address the limitations of received signal reference power (RSRP) based user-centric approach, we develop a deep reinforcement learning (DRL) approach with graph convolutional networks (GCN) to extract pertinent features encompassing connection status and topological information. The relative positioning of nodes is encoded within an adjacency matrix, while the connection states are captured within a feature matrix. Our empirical results validate the efficacy of the proposed approach, which consistently outperforms both the baseline method without GCN and the conventional user-centric approach.**

***Index Terms*—Deep reinforcement learning, cell-free massive MIMO, AP selection, graph convolutional networks.**

I. Introduction

For beyond fifth-generation (B5G) wireless access technology, cell-free massive multiple-input multiple-output (MIMO) has attracted the attention of researchers due to its high throughput, reliability, and energy efficiency [1]–[3]. In cell-free massive MIMO, all access points (APs) cooperate with each other to serve all users simultaneously by exploiting the same time-frequency resources. These APs are connected to a central processing unit (CPU) through a backhaul network and use conjugate beamforming to transmit data to users.

It is widely recognized that having all APs serve all users can lead to significant backhaul overhead. To mitigate backhaul overhead burden and enhance network practicality, one AP selection approach known as the user-centric approach was proposed in [4], where each AP only serves the users with the strongest channels, i.e., the strongest received signal reference power (RSRP). However, the user-centric approach overlooks the information of network topology, and might perform inadequately in certain topological scenarios. Consider a network topology where numerous users are situated in close proximity to one AP. According to user-centric approach, these user will be connected to the same AP, resulting in insufficient receiving signal strength and severe intra-AP interference. To enhance the user-centric approach in such network topologies, the effective channel gain instead of RSRP is considered during AP selection process in [5] to reduce intra-AP interference. Yet it was considered that each user connects to only one AP, which can be extended to multiple APs.

To leverage the network topology information, machine learning (ML) approaches like graph neural networks (GNN) have been proposed to capture the relative position information between APs and users, which can be effectively transformed into graph structure. Specifically, graph convolutional networks (GCN), as explored in [6], have been employed to extract features from these graph-shaped inputs in various machine learning applications. The study in [7] predicted the potential links between nodes even if the RSRP measurements of few known set of APs are available to the GNN. However, it's worth noting that the performance of supervised learning is significantly influenced by the quality of labels. The labels employed in the aforementioned supervised learning method are still closely tied to suboptimal RSRP, which may pose challenges in achieving optimal network rates.

Reinforcement learning (RL) is a decision-making process that learns which choice provides better benefits based on the experience. In particular, recent studies have proposed deep reinforcement learning (DRL) based methods exploiting GNN to optimize wireless networks. For instance, a centralized channel allocation scheme was introduced in [8] along with DRL-GCN framework in wireless local area networks. Another study [9] adopted a GNN method and policy gradient algorithm for the network node deployment problem. Additionally, in [10], researchers proposed a DRL-based solution that utilizes the underlying graph to learn the weights of GNN for optimal user-cell association in the Open Radio Access Network (O-RAN). Nevertheless, the above methods are rarely applied to cell-free massive MIMO systems. Our previous work [11] investigated clustering and power control problem in cell-free massive MIMO, but did not utilize GNN. In summary, the DRL-GNN based framework represents a promising direction for addressing the AP selection problem due to its effectiveness in function approximation and graph feature extraction.

In this work, we consider the problem of AP selection in cell-free massive MIMO systems. For better fairness, the objective is to maximize average throughput of the lower portion of users rather than focusing solely on the sum rate. We develop a GCN-based deep Q-network (DQN) framework that captures the latent association between APs and users. The key contributions of this paper are summarized as follows:

- We propose a GCN-DQN based approach which can be utilized to problems with graph-shaped states. Only

large-scale fading coefficients are considered in adjacency matrix and connection status is represented in feature matrix.
- Upon comparing proposed GCN-baesd method to simple neural networks based benchmark, we find that it outperforms baselines because of GCN's advantage in graph.

The rest of the paper is organized as follows. Section II describes the system model. The GCN-DQN based method is proposed in Section III. We provide numerical results and discussions in Section IV. Finally, Section V concludes the paper.

Notations: The superscripts * and ^{T} denotes conjugate and transpose. $\mathcal{CN}(0,\sigma^2)$ denotes circularly symmetric complex Gaussian distribution. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation.

II. System Model

We consider a downlink cell-free massive MIMO network with M APs and K users (or user equipments, UEs). Let $\mathcal{M}$ and $\mathcal{K}$ denote the sets of all APs and UEs, respectively. APs and UEs are randomly located and each equipped with a single antenna. All APs connect to a central processing unit via a backhaul network. APs jointly serve UEs using the same time-frequency resource. The channel coefficient between the m-th AP and the k-th UE is modeled as

$$g_{mk} = (d_{mk}/d_0)^{-\alpha} h_{mk}, \tag{1}$$

where d_{mk} is the distance between the m-th AP and the k-th UE, d_0 is the reference distance, α is the path-loss exponent, and $h_{mk} \sim \mathcal{CN}(0,1)$ denotes small-scale fading. We model the relative positional relationship between nodes using a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. The edges $e_{ij} = \{i,j\} \in \mathcal{E}$ of the graph are connected if and only if the Euclidean distance between them is smaller than a threshold, i.e., d_{max}. To represent the graph $\mathcal{G}$, the adjacency matrix is defined as an $(M+K)\times(M+K)$ matrix $\mathbf{A} = (A_{ij})$ as follows:

$$A_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and } \mathrm{e}_{ij} \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Let us define $\mathcal{K}(m)$ as the set of users served by the m-th AP, and $\mathcal{M}(k)$ as the set of APs serving the k-th user. Let the $M \times K$ feature matrix $\mathbf{X} = (X_{mk})$ represent connection status, which is given by

$$X_{mk} = \begin{cases} 1, & \text{If AP } m \text{ serves UE } k, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We have $X_{mk} = 1$ if UE $k \in \mathcal{K}(m)$ for all $m = 1, 2..., M$ or AP $m \in \mathcal{M}(k)$ for all $k = 1, 2..., K$. We consider that all UEs are guaranteed to be served but not all APs serve UEs. Besides, each UE is served by at most N APs, i.e., $|\mathcal{M}(k)|$ does not exceed N. Using conjugate beamforming, the signal transmitted by the m-th AP can be expressed as

$$x_m = \sum_{k\in\mathcal{K}(m)} \sqrt{\eta_{mk}} \widehat{g}^*_{mk} q_k, \tag{4}$$

where q_k is the symbol intended for the k-th user and it satisfies $\mathbb{E}\left\{|q_k|^2\right\} = 1$. For simplicity, we assume perfect channel state information (CSI), i.e., the estimates of channels are true, $\widehat{g}_{mk} = g_{mk}, \forall m, k$. Let P_m denote the transmitted power by the m-th AP and η_{mk} denote the power control coefficients. In this work, we let $\eta_{mk} = \frac{P_m}{\sum_{k'\in\mathcal{K}(m)}|g_{mk'}|^2}$. The corresponding signal received at the k-th UE is given by

$$\begin{aligned} y_k &= \sum_{m=1}^{M} g_{mk} x_m + w_k \\ &= \sum_{m=1}^{M} \sum_{k'\in\mathcal{K}(m)} \sqrt{\eta_{mk'}} g_{mk} g^*_{mk'} q_{k'} + w_k \\ &= \sum_{k'=1}^{K} \sum_{m\in\mathcal{M}(k')} \sqrt{\eta_{mk'}} g_{mk} g^*_{mk'} q_{k'} + w_k \\ &= \underbrace{\sum_{m\in\mathcal{M}(k)} \sqrt{\eta_{mk}} \left|g_{mk}\right|^2 q_k}_{\text{desired signal}} \\ &\quad + \underbrace{\sum_{k'\neq k}^{K} \sum_{m\in\mathcal{M}(k')} \sqrt{\eta_{mk'}} g_{mk} g^*_{mk'} q_{k'}}_{\text{multiuser interference}} + w_k, \end{aligned} \tag{5}$$

where $w_k \sim \mathcal{CN}(0,\sigma_w^2)$ is the additive noise at the k-th UE. The downlink Signal-to-Interference plus Noise-Ratio (SINR) for the k-th UE can be expressed as

$$\gamma_k = \frac{\left|\sum_{m\in\mathcal{M}(k)} \sqrt{\eta_{mk}} \left|g_{mk}\right|^2\right|^2}{\sigma_w^2 + \sum_{k'\neq k}^{K} \left|\sum_{m\in\mathcal{M}(k')} \sqrt{\eta_{mk'}} g_{mk} g^*_{mk'}\right|^2}. \tag{6}$$

The downlink achievable rate for the k-th UE R_k is given by $\log_2(1+\gamma_k)$. In this study, to improve the overall throughput without loss of fairness, we consider the average throughput of the lower n UEs. Let the n-th lowest throughput among K UEs be denoted by ξ_n, then our objection is to maximize $\frac{1}{n}\sum_{l=1}^{n} \xi_l$.

III. Deep Q-Learning Algorithm

We propose a deep Q-learning approach in which a Q-function is learned from the graph and connection status. The action a_t at time step t is to connect a legal AP-UE pair. Illegal actions such as connecting the connected pairs or exceeding the maximum allowable number of connections for any UE will be masked. The state s_t at time step t is composed of the graph adjacency matrix $\mathbf{A}$ and the square matrix $\overline{\mathbf{X}}$, which can be expressed as

$$\overline{\mathbf{X}} = \begin{bmatrix} \mathbf{0}_{K\times K} & \mathbf{X}^T \\ \mathbf{X} & \mathbf{0}_{M\times M} \end{bmatrix}. \tag{7}$$

The initial state can be considered as partially connected networks. The terminal state s_T is achieved when all the UEs reach maximum number of connection. The reward r_t is defined as the average throughput of the lower n UEs after selecting action a_t, i.e., $r_t = \frac{1}{n}\sum_{l=1}^{n} \xi_l^t$.

Following policy π, The action-value function represents the expected accumulative reward of taking action a_t in state s_t:

$$Q^{\pi}\left(s_t, a_t\right)=\mathbb{E}_{\pi}\left[G_t \mid s_t, a_t\right]=\mathbb{E}_{\pi}\left[\sum_{k=0}^{T-t} \gamma^k r_{t+k} \mid s_t, a_t\right], \quad (8)$$

where G_t is the cumulative discounted reward and γ is the discount factor. The optimal policy π^* is commonly learned through the estimation of optimal action-value function, which follows Bellman optimality equation:

$$Q^*\left(s_t, a_t\right)=\mathbb{E}\left[r_t+\gamma \max_{a'} Q^*\left(s_{t+1}, a'\right) \mid s_t, a_t\right]. \quad (9)$$

Optimal action-value function maximizes cumulative reward, then we can obtain an optimal policy through the estimation of the optimal action-value function:

$$Q^*(s, a)=\max_{\pi} Q^{\pi}(s, a), \quad (10)$$

$$\pi^*(a \mid s)=\arg\max_{a} Q^*(s, a). \quad (11)$$

Considering the number of feasible connection states is extremely high, we adopt neural network to approximate $Q^*(s,a)$. DQN exploits the main network with parameter θ and the target network with parameter θ^- to ensure convergence. To break the temporal correlation in the training data, a mini-batch of transitions $\mathcal{B}$ will be randomly sampled from replay buffer $\mathcal{D}$ to update the networks. The loss function and gradient descent are described as follows:

$$L\left(\theta\right)=E_{\mathcal{B}}\left[\left(r+\gamma \max_{a'} Q\left(s', a' \mid \theta^-\right)-Q\left(s, a \mid \theta\right)\right)^2\right], \quad (12)$$

$$\theta \leftarrow \theta-\alpha_\theta \nabla_\theta L(\theta), \quad (13)$$

where α_θ is the learning rate. To balance the exploitation and exploration, we adopt ϵ-greedy policy while training. The detailed learning process is summarized in Algorithm 1.

Fig. 1 indicates the overall structure of the GCN-based network, where "Dense" represents the fully connected layer and "Graph Convolution" represents the graph convolutional layer. By incorporating the GCN layers into the neural network model, we can treat the graph structure of cell-free massive MIMO in a manner analogous to a convolutional neural network (CNN) processing an input image. Traditionally, spectral techniques employ eigen decomposition to produce node representation vectors. However, these methods suffer from computational inefficiency and limited generalizability. GCN effectively addresses these issues through its robust approximation approach, where the hidden representation in the l-th layer is updated as follows:

$$\mathbf{X}^{(l)}=\sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}^{(l-1)} \mathbf{W}^{(l-1)}\right), \quad (14)$$

where $\tilde{\mathbf{A}}=\mathbf{A}+\mathbf{I}$ and $\mathbf{I}$ is the identity matrix. $\tilde{\mathbf{D}}_{ii}=\sum_i \tilde{\mathbf{A}}_{ij}$ is the degree matrix and $\mathbf{W}^{(l)}$ is the trainable weight matrix of l-th layer. $\sigma(\cdot)$ denotes activation function, i.e., $\mathrm{ReLU}(\cdot)=\max(0,\cdot)$ in this work. $\mathbf{X}^{(l)}$ is the output matrix of l-th layer and $\mathbf{X}^{(0)}=\overline{\mathbf{X}}$. This approach normalizes the adjacency matrix and ensures that information about each node is retained during the propagation process across different layers. The hidden dimensions of graph convolution $\mathbf{W}$ is 64 and 32 while the number of neurons in "Dense" is 300.

Algorithm 1 DQN Procedure

Randomly initialize network parameters θ.
Initialize target network parameters $\theta^- = \theta$.
Initialize the replay buffer $\mathcal{D}$ with capacity D.
for episode = 1 to MAX-EPISODE **do**
  Set the initial state s_1.
  for t = 1 to T **do**
    Execute action $a_t \sim \pi(a_t \mid s_t)$.
    $$\pi\left(a_t \mid s_t\right)=\begin{cases}\text{random AP-UE pairing}, & \text{w.p. } \epsilon \\ \arg\max_{a_t} Q\left(s_t, a_t\right), & \text{o.w.}\end{cases}$$
    Obtain the reward r_t and next state s_{t+1}.
    if $|\mathcal{D}| < D$ **then**
      Store experience (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$.
    else
      Replace the old experience by new experience.
    end if
    if $|\mathcal{D}| \geq |\mathcal{B}|$ **then**
      Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$ randomly.
      Calculate target value:
      $$y_i=\begin{cases} r_i, & t=T \\ r_i+\gamma \max_{a'} Q\left(s_{i+1}, a' \mid \theta^-\right), & \text{o.w.}\end{cases}$$
      Calculate the loss function $L(\theta)$:
      $$L(\theta)=\frac{1}{|\mathcal{B}|} \sum_{b_i \in \mathcal{B}}\left(y_i-Q(s_i, a_i|\theta)\right)^2.$$
      Update the weights θ:
      $$\theta \leftarrow \theta+\alpha_\theta \sum_{b_i \in \mathcal{B}}\left(y_i-Q(s_i, a_i|\theta)\right) \nabla_\theta Q(s_i, a_i|\theta).$$
    end if
    $s_t \leftarrow s_{t+1}$.
  end for
  if episode$\%\tau = 0$ **then**
    $$\theta^- \leftarrow \theta.$$
  end if
end for

IV. Simulation Results

We consider a network with 10 APs and 5 UEs randomly distributed in a 1 × 1 km^2 square. To reduce the sparse reward of the GNN-RL inference for better convergence, we divide UEs into two categories. For the first class of UEs, if the difference between the strongest and the second strongest RSRP measurements is more than 3dB, then these UEs are associated with APs with the strongest RSRP at the beginning of each episode. The remaining UEs will be associated to APs during training. We aim at maximizing throughput of the lowest 3 UEs. We test our algorithm in 30 random scenarios

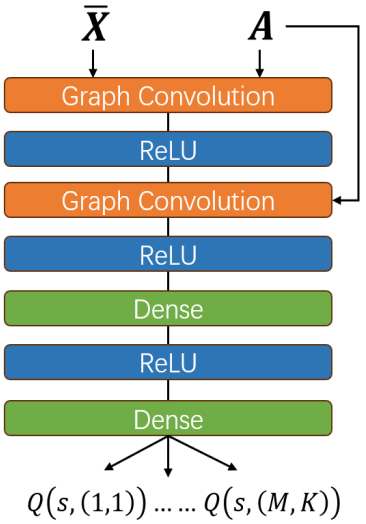


Fig. 1. The architecture of the GCN-based network.

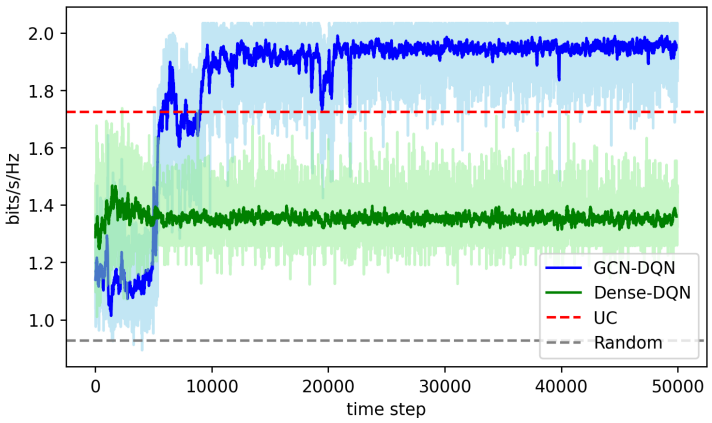


Fig. 2. Training process for GCN-DQN and Dense-DQN when $N = 2$.

with different topology. We train 50000 episodes for each scenario. The essential parameters are listed in Table I.

TABLE I
SYSTEM PARAMETERS

Parameters	Value
Path-loss exponent α	2
distance threshold d_{max}	200m
Transmit power P_m	10 mW
Noise figure	9 dB
Bandwidth	20 MHz
Discount factor γ	0.99
learning rate α_θ	0.1
Update iteration τ	100
Replay buffer capacity $\mathcal{D}$	25000
Batch size $\mathcal{B}$	64
greedy parameter ϵ	0.1

We compare the proposed GCN-DQN method with three other strategies termed as Dense-DQN, UC and Random. In the Dense-DQN strategy, "Graph Convolution" layers are replaced by "Dense" layers while other parameters and algorithm are consistent with GCN-DQN. User-centric approach [4] is adopted in UC, i.e., each UE is served by N APs with the top N strongest RSPR. Random association method with neighboring C APs for each UE is referred as Random. The numerical results reported in this paper are the average of the data obtained during the final ten percent of the simulation time.

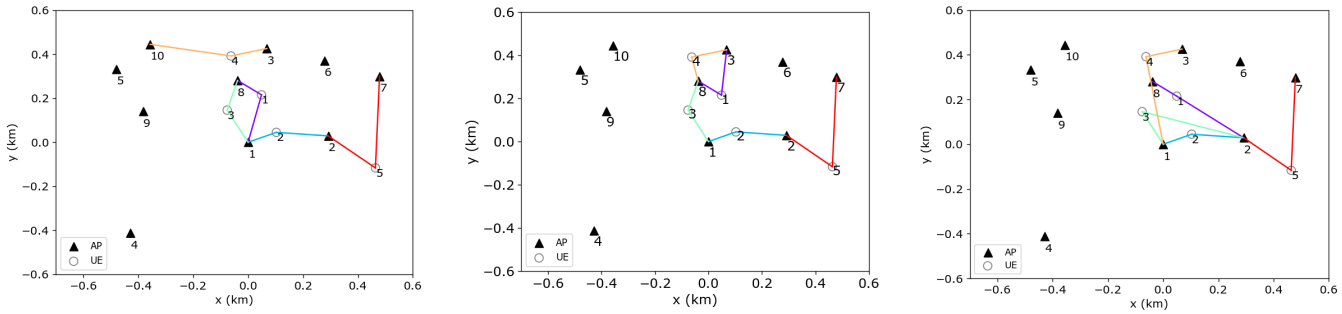
Fig. 2 displays the learning curves when $N = 2$. The dark solid line represents the average calculated per 100 steps. The result shown in the figure is the terminal state of training (all UEs are served by 2 APs). It can be seen that our proposed GCN-DQN method reaches convergence and outperforms other 3 baselines. Fig. 3 shows final connection results provided by GCN-DQN, UC and Dense-DQN in the scenario of Fig. 2. For GCN-DQN, we observe that more APs are exploited than UC, leading to less intra-AP interference and more signal strength for each UE. On contrast, Dense-DQN fails to obtain the optimal solution as UEs seek far from the near in Fig. 3(c).

Fig. 4 shows the cumulative distribution functions (CDFs) of different strategies under different maximum connection numbers. Specifically, "final" denotes that all UEs are fully served by N APs and "best" represents the optimal value achieved in the process (UEs may not reach the maximum number of connection). As indicated in the figure, the average throughput of the lowest 3 UEs are ranked as GCN-DQN (best) $>$ GCN-DQN (final) $>$ UC $\geq$ Dense-DQN (best) $>$ Dense (final) $\gg$ Random. When $N = 1$, the curve for GCN-DQN (best) basically coincides with GCN-DQN (final) as terminal states indicate the best status in most cases. As N increases, the performance gap between "best" and "final" widens while the performance gap between Dense-DQN and UC narrows. Moreover, the performance deteriorates for all methods as N increases. This is because as the loads on each AP grow, intra-AP interference becomes more severe and terminal states may not be the best status.

Fig. 5 denotes the average throughput of the lowest 3 UEs, and we can obtain the same conclusion that as from Fig. 4. GCN-based model exceeds those of the other methods, and learning performance exceeds that of the simple neural network model. Overall, GCN-DQN (best) strategy outperforms UC, Dense-DQN (best) and Random by up to $11.3\%, 20.3\%$ and 210.7% in terms of average throughput of the lowest 3 UEs. These data come to $8.5\%, 23.1\%$ and 202.6% when comparing GCN-DQN (final) with UC, Dense-DQN (final) and Random, respectively.

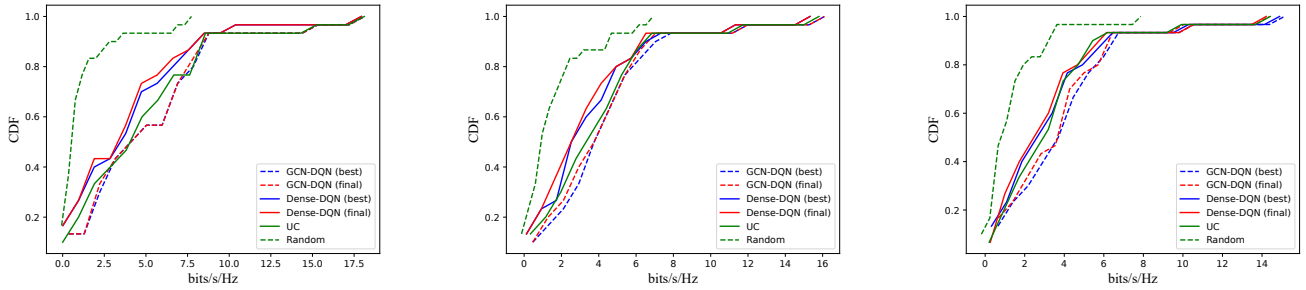
V. CONCLUSION

In this paper, we propose a GCN-DQN approach to solve the AP-UE association problem in cell-free massive MIMO networks. The simulation results demonstrate that our proposed method outperforms fully connected layer based DQN and RSRP based user-centric approach. Looking forward, we



(a) (b) (c)

Fig. 3. A snapshot of a specific scenario when $N = 2$. Hollow circles and solid triangles represent UEs and APs, respectively. Lines with different colors indicate different UEs. (a) GCN-DQN. (b) UC. (c) Dense-DQN.



(a) N=1 (b) N=2 (c) N=3

Fig. 4. CDF for rewards with different N.

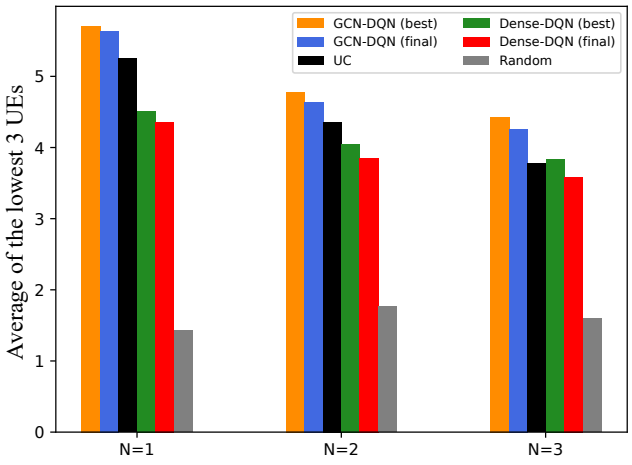


Fig. 5. Average rewards for different schemes with different N.

are interested in load balancing and energy efficiency as our optimization goals.

REFERENCES

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[2] P. Liu, K. Luo, D. Chen, and T. Jiang, "Spectral efficiency analysis of cell-free massive MIMO systems with zero-forcing detector," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 795–807, Feb. 2020.

[3] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[4] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[5] H. T. Dao and S. Kim, "Effective channel gain-based access point selection in cell-free massive MIMO systems," *IEEE Access*, vol. 8, pp. 108 127–108 132, 2020.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[7] V. Ranasinghe, N. Rajatheva, and M. Latva-aho, "Graph neural network based access point selection for cell-free massive MIMO systems," in *Proc. IEEE GLOBECOM*, 2021, pp. 01–06.

[8] K. Nakashima, S. Kamiya, K. Ohtsu, K. Yamamoto, T. Nishio, and M. Morikura, "Deep reinforcement learning-based channel allocation for wireless lans with graph convolutional networks," *IEEE Access*, vol. 8, pp. 31 823–31 834, 2020.

[9] Y. Yang, D. Zou, and X. He, "Graph neural network-based node deployment for throughput enhancement," *IEEE Trans. Neural Netw. and Learn. Syst.*, pp. 1–15, Jun. 2023.

[10] O. Orhan, V. N. Swamy, T. Tetzlaff, M. Nassar, H. Nikopour, and S. Talwar, "Connection management xapp for O-RAN RIC: A graph neural network and reinforcement learning approach," in *Proc. IEEE Int. Conf. Mach. Learn. Appl*, 2021, pp. 936–941.

[11] M. Du, X. Sun, Y. Zhang, J. Wang, and P. Liu, "Joint cooperation clustering and downlink power control for cell-free massive MIMO with deep reinforcement learning," in *Proc. IEEE ICCT*, 2023.