IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 October 2022, accepted 5 November 2022, date of publication 10 November 2022, date of current version 16 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221455

RESEARCH ARTICLE

Multi-Agent Reinforcement Learning With Measured Difference Reward for Multi-Association in Ultra-Dense mmWave Network

XUEBIN LI[1], TERRY N. GUO[2], (Senior Member, IEEE), AND ALLEN B. MACKENZIE[1], (Senior Member, IEEE)

[1]Department of Electronics and Communication Engineering, Tennessee Tech University, Cookeville, TN 38505, USA
[2]Center for Manufacturing Research, Tennessee Technological University, Cookeville, TN 38501, USA

Corresponding author: Terry N. Guo (NGuo@tntech.edu)

This work was supported in part by the National Science Foundation under Grant 2135275.

ABSTRACT Millimeter Wave (mmWave) communication technology is anticipated to play a vital role in meeting the growing demand for the scarce bandwidth in wireless communications. However, mmWave networks are highly susceptible to blockage. Thus, some mitigation techniques, such as multi-connectivity, need to be considered. Densely deploying mmWave base stations (mBSs) to form an ultra-dense network (UDN) also helps. With a mix of different technologies, optimally allocating resources becomes challenging. In this paper, we study mmWave user multi-association in a two-tier heterogeneous ultra-dense network (HetUDN) with a relatively large number of user equipments (UEs). We propose a framework of multi-agent reinforcement learning (MARL) to tackle the complicated optimization problem, leveraging its adaptivity to the communication environment. The proposed scheme considers mmWave beam-division based multi-connectivity and takes advantage of a macro base station (MBS) for indirect cooperation among agents (UEs). In particular, we borrow a credit-assignment technique called difference reward (DR) to deal with a relatively large MARL system with a large action space, which, to the best of our knowledge, is the rst time to apply MARL with DR in user association. Furthermore, the proposed schemes are scalable mainly due to xed observation dimensions and individual actions taken by UEs independently, ensuring that the operation is independent of the numbers of mBSs and UEs. Numerical results suggest that the two MARL schemes with measured DR could achieve a good balance between energy efciency and QoS outage, and the one using extended DR (EDR) offers additional performance improvement.

INDEX TERMS Ultra-dense network (UDN), millimeter wave (mmWave), heterogeneous network, multi-association, multi-agent reinforcement learning (MARL).

I. INTRODUCTION

As communication systems continue to evolve to meet the growing demand for high bandwidth and low latency, we have seen technology trends and enablers for upcoming Beyond 5G (B5G) and 6G communication systems, such as the use of higher frequency bands, multi-connectivity, heterogeneity and network densication, etc. Synergistically utilizing different technologies is vital but challenging. Densely deploying mmWave base stations (mBSs) is quite reasonable since this favors mmWave communication that has a short-range line-of-sight (LOS) propagation. In an ultra-dense network (UDN) [1] with mBSs, a user equipment (UE) capable of multi-connectivity can be associated with multiple mBSscalled multi-association [2]. Multi-association is considered a good strategy to increase channel capacity and reduce link blockage [3]. An example of a promising solution is a two-tier heterogeneous network (HetNet) with densely

The associate editor coordinating the review of this manuscript and approving it for publication was Xujie Li.

deployed low-power mBS within the coverage of an existing cellular network, where UEs can connect to the macro base station (MBS) and multiple mBS simultaneously [4].

In order to leverage the full potential of UDNs, the network architecture is transformed from traditional cell-centric to user-centric [5], and the UE can be associated with multiple mBSs in a multi-cell setup [2]. The user association problem is NP-hard in general, thus it is difcult to achieve strictly optimal association given some constraints. Recently, researchers have proposed some solutions to multi-association, such as user-centric clustering [6], [7], multi-label classication methods [8], nonlinear programming method [9], and heuristic algorithm [10], [11]. Chen, et al. investigate optimal multi-connectivity and downlink power allocation using a few mathematical techniques to deal with the non-convex Mixed-Integer Nonlinear Programming (MINL) with summation of fractions [9]. Crowd intelligence has been applied to tackle multi-connectivity issues [12], [13]. In [12], a non-dominated sorting genetic algorithm is proposed to obtain the near-optimal solution of multiple associations; the algorithm provides maximum energy efciency while balancing user rate and base station load under QoS constraints. In [13], the authors introduce a Multi-Objective Harris Hawk Optimization algorithm designed to achieve near-optimal performance. Note that most of the above works rely on accurate models of the UDN environments. Environment online modeling requires intensive involvement of UEs and information exchanges, which tends to be impractical for normal mmWave UDNs.

Mainly because of its support for autonomous behavior, reinforcement learning (RL) has been successfully applied to many different elds [14]. In particular, multi-agent Reinforcement Learning (MARL) is gaining increasing attention in different elds, such as communications and network-ing [15], [16], caching and computation ofoading [17], [18], [19], [20], [21], resource management and allocation [22], [23], [24], etc. Recently, a number of works have applied MARL to solve the user association problem, and many of them consider mmWave UDN [25], [26], [27], [28], [29], [30].In [25], the authors design a historical information-based MARL method to solve the vehicle association problem to achieve load balancing in vehicular networks, considering the periodicity of urban trafc and the constraints of vehicle travel paths. In [28], the authors study the symbiotic relationship between cellular and IoT networks and infer real-time channel information by using historical channel information. Sana et al. [29] designed an adaptive user association algorithm to maximize the sum-rate. Dinh et al. [30] investigated user-to-multiple sub-6GHz/mmWave access points (APs) association and solved the problem in a distributed fashion with the goal of maximizing the long-term throughput of the whole system. MARL has shown effectiveness in addressing decision making under uncertainty [31]. Therefore, it is promising to apply MARL to mmWave UDN to achieve optimal multi-association.

MARL can have different forms, such as distributed learning and centralized learning, and each type has its advantages and disadvantages. In particular, MARL algorithms with centralized training decentralized execution (CTDE) [32] have gained attention in recent years. In CTDE, global information is required for centralized training but not for executing tasks, and all agents make individual decisions based on mutual understanding gained during previous training. Among many CDTE algorithms, VDN [33], QMIX [34], COMA [32], etc., have shown impressive performances. However, the cellular network is an open network architecture that requires scalability for BSs and UE, and centralized training methods often cannot meet the very essential scalability requirements since CTDE requires the number of agents is xed.

One important issue investigated in this paper is related to reward denition. Reward functions employed in some work become ineffective as the number of UEs increases [29], [30]. Indeed, in a relatively large system with a large action space, the use of a global reward as an agent's individual reward often leads to the ''lazy agent'' problem[1] [33], [35]. On the other hand, when using a selsh local reward without cooperation among agents, it is not guaranteed that pursuing maximum individual return will lead to a maximum global return. Regarding fair assignment of credit to an agent in proportion to its contribution, existing work includes local reward design [22], [25], [28]. The concept of difference reward (DR) is introduced by Wolpert et al. [36] as a solution to the MARL credit assignment. DR is a decoupled individual reward reecting an individual contribution of the current agent's action to the global reward [32], [37]. In this paper, we investigate a user multi-association problem to maximize the energy efciency of mmWave UND while minimizing QoS outage in a relatively large MARL system with a large action space. The joint optimization problem is t into a MARL framework, and MBS is used to accommodate information exchange and indirect cooperation among agents (UEs). In particular, we consider DR to assign credit to an agent as a response to its action.

To the best of our knowledge, it is the rst time that MARL with DR is considered in user association. Different from model-based algorithms such as crowd intelligence [12], [13], nonlinear programming method [9] and heuristic algorithm [10], [11]. Our proposed schemes belong to the model-free online learning category which does not require infeasible online modeling based on real world systems. On the other hand, compared to some centralized-training approaches, such as VDN [33], QMIX [34], and COMA [32], etc., our schemes are more scalable for cellular systems with a varying number of UEs.

The main contributions of this paper are summarized as followsV

[1]A lazy agent is one who gets spurious rewards because of other agents' efforts in MARL, i.e., it takes advantage of the successful actions of other agents. This phenomenon is due to partial observability.

MBS-aided MARL Framework for User Multi-association in Heterogeneous mmWave UDN With Relatively Large Number of UEs: The framework is proposed to address a few pressing demands and challenges, including multi-connectivity for mitigating mmWave blockage, scalability and effectiveness of MARL as the number of UEs increases. The proposed framework is systematically studied and validated.

Effective MARL Leveraged by DR: As the number of UEs (agents) increases, many existing MARL schemes become ineffective, especially if multi-connectivity is considered. We borrow the concept of DR for proportionally assigning a credit to an action, and use it as an agent's individual reward. A DR measurement technique is proposed as an attempt to apply DR to real-world systems.

Extended DR (EDR): The traditional DR is extended for further performance improvement. A concept of compound actions, along with a conict resolution mechanism, is introduced. EDR is a combination of DR and compound-action-based DR (CDR). The conict resolution mechanism is to deal with beam contention in the concurrent follow-up actions in the second part of the compound action.

Table 1 summarizes the symbols used in this paper. The rest of this paper is organized as follows. The system is described and modeled in the next section. In section III, the multi-association problem is formulated as an optimization problem and then ts into the MBS-aided MARL Framework. The application of DR, including traditional DR and its extended version, is presented in Section IV. Section V reports simulation results and provides some remarks, followed by conclusions in Section V.

TABLE 1. Major symbols used in this paper.

Symbol	Remarks
I	Number of UEs
J	Number of mBSs
$\mathcal{I}=\{1,2,...,I\}$	The set of UEs
$\mathcal{J}=\{1,2,...,J\}$	The set of mBSs
$\mathcal{L}_i$	UE i's neighbor mBS set
$RSSI_{i,j}$	The RSSI between UE i and mBS j
$RSSI_{limit}$	The threshold of the maximum transmission range
λ	Wavelength
d_0	Close-in reference distance
$d_{i,j}$	Distance between UE i and mBS j
$G^r_{i,j}$	Receive antenna gain
$G^t_{i,j}$	Transmit antenna gain
$SINR_{i,j}$	Signal to Interference and Noise Ratio
$PL(d_{i,j})$	path loss
$X_{i,j}$	Random variable for shadowing
$P^r_{i,j}$	Power received by UE i from mBS j
P_{total}	mBS total transmit power
$c_{i,j}$	Downlink achievable rate of the link between UE i and mBS j
C_i	Downlink achievable rate of UE i
P_j	mBS j transmit power
P_{ceil}	Total transmit power of mBS
$I^{Type1}_{i,j}$	Interference between UEs
$I^{Type2}_{i,j}$	Multi-connection interference on UE i
R	Total network sum-rate
R_j	Sum-rate of mBS j
k	The maximum number of mmWave beams per UE
f	The maximum number of mmWave beams mBS
$x_{i,j}$	Association indicator for UE i and mBS j

II. SYSTEM DESCRIPTION AND MODELING

We consider a two-tier HetNet with mmWave UDN as the second tier illustrated in Fig. 1. In the rst-tier, an MBS is able to communicate with all UEs over two-way channels in the lower frequency band to exchange control information. The second tier is a mmWave UDN with each UE being equipped with a mmWave frontend able to connect to multiple mBSs simultaneously in a beam-division manner [38], [39]. mmWave beamforming and beam alignment are beyond the scope of this paper. We assume that sharp beamforming and even some nulling techniques have been employed to enable mmWave beam-division-based multiple connectivities. We also assume the system is capable of beam alignment and that perfect beam alignment has been achieved.

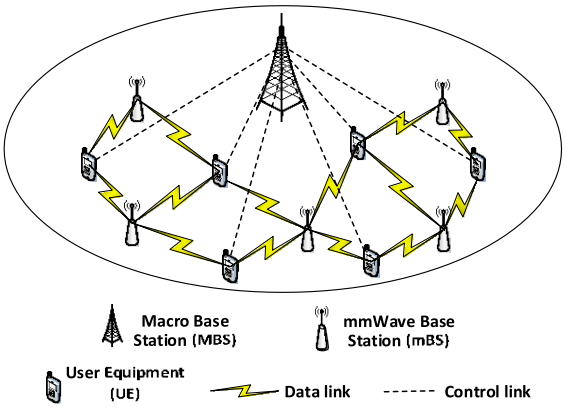


FIGURE 1. Two-tier HetNet with mmWave UDN as the second tier.

Consider a system with I UEs and J mBSs; let $\mathcal{I} = \{1,2,\ldots,I\}$ and $\mathcal{J} = \{1,2,\ldots,J\}$ be the index sets of UEs and mBSs, respectively. In practice, each UE can only access its neighboring mBSs. The received signal strength indicator (RSSI) can be used to determine the neighbors (candidate mBSs that can be heard by a UE). For a given UE i, these candidate mBSs can be represented by a candidate mBS index set dened as

$$\mathcal{L}_i = \{j | RSSI_{i,j} > RSSI_{limit}; j \in \mathcal{J}\}, \quad i \in \mathcal{I} \tag{1}$$

where $RSSI_{i,j}$ is the RSSI for a pair of receiver UE i and transmitter mBS j, and $RSSI_{limit}$ is a threshold corresponding to the minimum required signal strength and maximum transmission range.

The mmWave propagation, antenna radiation pattern, interference, etc., are modeled as follows for performance evaluation. The log-normal path loss model [40] is used to express mmWave propagation

$$PL(d_{i,j}) = 10log_{10}\left(\frac{4d_0}{\lambda}\right)^2 + 10\, n_{i,j} \log_{10}\left(\frac{d_{i,j}}{d_0}\right) + X_{i,j} \tag{2}$$

where d_0 is the close-in-free-space reference distance, $n_{i,j}$ is the path loss exponent, is the wavelength, $X_{i,j}$ is the zero-mean Gaussian random variable expressed in dB with standard deviation , to reect shadowing or blockage effect. Measurement results reported in [41] have suggested that path loss equation (2) can be applied to both LOS and non-line-of-sight (NLOS) conditions with $n_{i,j}$, $X_{i,j}$ being set differently. According to [42], the channel condition can switch randomly between LOS and NLOS following a Bernoulli distribution. In urban areas with regular street layouts, the probability of LOS at a distance d m is given by

$$P_{LOS}(d) = \min(\frac{A}{d}; 1)(1 - e^{-\frac{d}{B}}) + e^{-\frac{d}{B}}$$
$$A = 18 \text{ m}; B = 63 \text{ m}: \quad (3)$$

In suburban areas [42] suggests

$$P_{LOS}(d) = e^{-d/C}; \quad C = 200 \text{ m} \quad (4)$$

In our case, the two probabilities in Bernoulli distribution are denoted by $P_{LOS}(d_{i,j}; n_{i,j}; X_{i,j})$ and $P_{NLOS}(d_{i,j}; n_{i,j}; X_{i,j}) = 1 - P_{LOS}(d_{i,j}; n_{i,j}; X_{i,j})$.

To model mmWave antennas, we adopt the simplied and commonly used sectored antenna model [43], [44] with antenna pattern dened as

$$G(;^t) = \begin{cases} \frac{2\ (2\ )}{}; & \text{if } j^t j < \frac{}{2} \\ ; & \text{otherwise} \end{cases} \quad (5)$$

where , $0 < 1$, is the gain of side-lobe, ^{t} is the beam offset angle with respect to the broadside in radian, and is the main-lobe beamwidth in radian.

As mentioned earlier, we assume each UE is able to connect multiple mBSs simultaneously, which implies that the mmWave frontend has multiple channels that transmit or receive signals in different directions. For any downlink pair of UE and mBS, there are four types of possible beam combinations and the joint transmit-receive antenna gain in a given direction is expressed in (6), as shown at the bottom of the page, and superscripts t and r have been used to represent ''transmitter'' at a mBS and ''receiver'' at a UE, respectively. In (6), ^{t} and ^{r} are the main-lobe beamwidths, ^{t} and ^{r} are the beam offset angles, and $G^t(^t)G^r(^r)$ is the joint downlink transmit-receive antenna gain, corresponding to a pair of UE i and mBS j. If the beams between UE i and mBS j are aligned (corresponding to the rst line on the right-hand side of equation (6)), then, according to (5), $G^t_{i,j}(^t)$, and $G^r_{i,j}(^r)$ are given by

$$G^t_{i,j}(^t) = \frac{2\ (2\ ^t)}{^t_j} \quad (7)$$

$$G^r_{i,j}(^r) = \frac{2\ (2\ _i)}{^r_i} \quad (8)$$

To enable a multi-association process, downlink testing signals are generated by mBSs and measured by UEs. Normally each mBS beam transmits a constant-power P_{beam}. Denoted by $P^t_{i,j}$ the transmit power of mBS j toward UE i, and considering path loss dened in (2), the received power corresponding to the two aligned beams is given by

$$P^r_{i,j} = P^t_{i,j} G^t_{i,j}(0) G^r_{i,j}(0) PL(d_{i,j}) \quad (9)$$

where $^t = 0$ and $^r = 0$ have been used.

In this paper, we consider mmWave beam pattern model with both mainlobe and sidelobe, and it has been considered by many researchers, because this model is more general and accurate, and can be benecial for future work in this line. Indeed, some existing work suggests that when the UE is associated with multiple mmWave base stations simultaneously, the interference cannot be ignored [45]. Note that there can possibly be three types of downlink interference, corresponding to the second, third, and fourth lines in the right-hand side of (6): Type 1mainlobe to sidelobe; Type 2sidelobe to the mainlobe, and Type 3sidelobe to sidelobe. Obviously, Type-3 interference is negligible and will not be considered in our performance evaluation. Type 1 and Type 2 interference powers at UE i can be expressed as

$$I^{Type1}_{i,j} = \sum_{j' \in J, j' \neq j} x_{i,j'} P^t_{j'} G^t_{i,j'}(^t) G_{i,j'}(^r) PL(d_{i,j'})$$
$$j^t j < \frac{}{2}; \ j^t_j j > \frac{}{2} \quad (10)$$

$$I^{Type2}_{i,j} = \sum_{i' \in I, i' \neq i} x_{i',j} P^t_{i',j} G^t_{i',j}(^t) G^r_{j}(^r) PL(d_{i,j})$$
$$j^t j > \frac{}{2}; \ j^t_j j < \frac{}{2} \quad (11)$$

where $G^t_{i,j}(^t) G_{i,j}(^r)$ is the joint transmit-receive antenna gain for Type 1 downlink interference, $G_{i',j}(^t) G_{i,j}(^r)$ is

$$G^t_{i,j}(^t)G^r(^r_{i,j}) = \begin{cases} \frac{2\ (2\ _j)}{2\ (2\ ^t)} \frac{2\ (2\ _i)}{^t_j}; & \text{if } ^t \ 2 \ \frac{^t_j}{} \text{ and } ^r \ 2 \ \frac{^r_i}{} \\ \frac{2\ (2\ ^t)}{^t_j}; & \text{if } ^t \ 2 \ \frac{^t_j}{} \text{ and } ^r > 2 \ \frac{^r_i}{} \\ \frac{2\ (2\ _i)}{^r_i}; & \text{if } ^t > \frac{^t_j}{2} \text{ and } ^r \ \frac{^r_i}{2} \\ ; & \text{if } ^t > 2 \ \frac{^t_j}{} \text{ and } ^r > 2 \ \frac{^r_i}{} \end{cases} \quad i \in I; j \in J \quad (6)$$

the joint transmit-receive antenna gain for Type 2 downlink interference, $x_{i;j} \in \{0;1\}$ indicates the UE-mBS association, i.e., $x_{i;j} = 1$ if UE i is associated with mBS j, otherwise $x_{i;j} = 0$.

The Signal-to-Interference-and-Noise-Ratio (SINR) at UE i for the link (mBS j, UE i) is given by

$$SINR_{i;j} = \frac{x_{i;j}P^{r}_{i;j}}{I^{Type1}_{i;j} + I^{Type2}_{i;j} + WN_0} \tag{12}$$

where W denotes the channel bandwidth, and N_0 denotes the noise power spectral density. The downlink achievable rate of each associated link is dened as

$$c_{i;j} = W\log_2(1 + SINR_{i;j}) \tag{13}$$

The downlink achievable rates of UE i can be expressed as

$$R_i = \sum_{j \in J} x_{i;j}c_{i;j} \tag{14}$$

The overall energy efciency (EE) of the system can be formulated as

$$EE = \frac{\sum_{j \in J}\sum_{i \in I} x_{i;j}c_{i;j}}{\sum_{j \in J}\sum_{i \in I} P_{i;j}} \tag{15}$$

In addition, we introduce QoS coverage probability as a metric of network service performance

$$C_p = Pr[R_i \geq R_{QoS}] = \sum_{i} \frac{\mathbf{1}(R_i \geq R_{QoS})}{N_{UE}} \tag{16}$$

where N_{UE} is the number of UEs in the system, $\mathbf{1}()$ is an indicator function, and R_{QoS} is the minimum data rate requirement for each user.

III. PROPOSED MARL FRAMEWORK

In this section, we formulate the optimization problem rst and then introduce the MARL model as well as the architecture of the proposed scheme.

A. PROBLEM FORMULATION

In general, user association is an optimization problem that may have multiple objectives (such as sum-rate, QoS, load balancing, etc.). In this paper, our goal is to achieve high EE without sacricing QoS. Different from the traditional associative optimization objective, QoS is not regarded as a constraint, but as an optimization objective. Communication systems usually have some limitations due to hardware limitations. Denote by k the maximum number of mBSs a UE can link to, and f the maximum number of UEs an mBS can serve. The following constraints are applied to user association

$$\sum_{j \in L_i} x_{i;j} \leq k; \quad i \in I \tag{17}$$

$$\sum_{i \in I} x_{i;j} \leq f; \quad j \in L_i \tag{18}$$

$$x_{i;j} \in \{0;1\}; \quad \forall i \in I;\ j \in L_i \tag{19}$$

where constraint (17) indicates that each UE cannot be connected to more than k mBSs, constraint (18) means that one BS can simultaneously serve up to f UEs. The binary indicator $x_{i;j} \in \{0;1\}$ reects UE actions. Then, the optimization problem can be formulated as

$$P1 \max_{\{x_{i;j}\}} \{EE; C_p\}$$
$$s.t.\ (17) \quad (19) \tag{20}$$

The optimization problem dened above is a multi-objective optimization problem that is difcult to solve directly. As a matter of fact, this problem can be converted to single objective optimization. By dening a utility function

$$0 = EE \cdot e^{-C_p} \tag{21}$$

with being a tuning constant for balancing between EE and QoS, Finally, the problem can be rewritten as

$$P2 \max_{\{x_{i;j}\}} 0$$
$$s.t.\ (17) \quad (19) \tag{22}$$

In the following, we will need to convert the optimization P2 to a MARL problem.

B. MARL MODEL

(1) **Agent**: Each agent is played by a UE.

(2) **Observation**: It is the union of two subsets $\boldsymbol{o}^{0}_{i}$ and $\boldsymbol{o}^{00}_{i}$, where the former is a set of selected RSSI values observed by UE i, and the latter is a composition of global network parameters observed by the MBS. Specically, we have

$$\begin{aligned} \boldsymbol{o}_i &= \{o^{0}_{i}; o^{00}\} \\ \boldsymbol{o}^{0}_{i} &= \{RSSI_{i;j}; j \in L_i\}; \quad i \in I \\ \boldsymbol{o}^{00} &= \{SR; N_{UE}; N_{QoS}; R_j; N_j; N_{QoS;j}; j \in L_i\}; \quad i \in I \end{aligned} \tag{23}$$

where SR is the whole network's sum-rate, N_{UE} is the number of UEs in the area covered by the MBS, N_{QoS} is the number of UEs that do not satisfy QoS requirement, R_j is the sum-rate of mBS j, N_j is the number of UEs served by mBS j, and $N_{QoS;j}$ is the number of UEs in the area of mBS j that do not meet QoS requirement. For the convenience of description later, the state is dened as follows

$$\boldsymbol{s} = [o_1; o_2; \cdots; o_I] \tag{24}$$

(3) **Action**: It is the action taken by a UE (agent) to associate with mBSs. An action $\boldsymbol{a}_i$ expressed in (25) is equivalent to an association option and can be represented by a vector consisting of association indicator $x_{i;j} \in \{0;1\}$.

$$\boldsymbol{a}_i = [x_{i;1}; x_{i;2}; \cdots; x_{i;J}]^{T} \tag{25}$$

When multiple vectors are combined into a matrix form, a joint action $\boldsymbol{u}$ is formed as

$$\boldsymbol{u} = [a_1; a_2; \cdots; a_J] \tag{26}$$

(4) **Reward**: Dening an effective reward function for MARL is essential and tricky. Typically the reward function

r D $R(s_t; a_t; s_{tC1})$ is determined by the current state s_t, the actions a_t taken at state s_t, and the next state s_{tC1} after taking action a_t.The goal of multi-association in this paper is to maximize the utility function 0 dened in (21). In other words, 0 can simply serve as a global reward, i.e., r_G D 0. However, achieving this global goal in a multi-agent system requires all agents to act smartly and harmoniously, and each agent needs a ''local'' reward to guide its action toward the global goal. It is desired to align individual goals with the global goal. Achieving this alignment is very challenging and related to the credit assignment [46]. Credit assignment in MARL is to determine the individual contribution of the current agent's action to the global reward. Two credit assignment methods are described in the next section.

C. OVERALL ARCHITECTURE OF PROPOSED SCHEMES WITH DR

The optimization problem in (21) and (22) is very difcult to solve. In this paper, we try to solve the problem using MARL which is essentially a ''static game of incomplete information'' according to [47]. Multiple agents (played by UEs) participating in the game try to improve their rewards via iterative exploration and exploitation. Specically for our application, the reward is formulated based on the utility function dened in (21), and the action taken by an agent is to get associated with some mBSs. Depicted in Fig. 2 is a conceptual illustration of the MBS-aided MARL framework applied to a two-tier HetNet, where a common deep neural network[2] at MSB is shared by all UEs. It is worth noting that although the network under our consideration is a single-MBS system, it can be extended to a multi-MBS system with some modication.

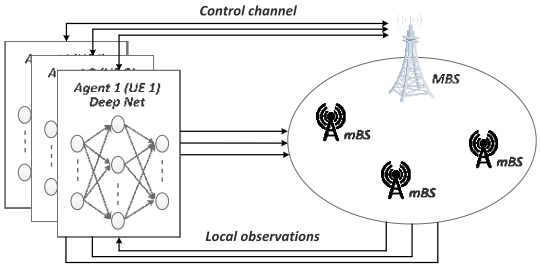


FIGURE 2. Conceptual illustration of MBS-aided MARL framework.

MBS has four main tasks: 1) scheduling all UEs to ensure only one UE is in the exploration state in a time step; 2) broadcasting network parameters obtained from mBS, assist UE in observing the network environment; 3) conducting collision avoidance according to some predened conict resolution rule, and 4) Train and update the neural network weights. Two MARL-based schemes incorporating DR concept are briey described in the following:

DR-MARL: a UE performs exploration by taking an action in the learning period, which is simply to connect to or disconnect from one or more BSs; we name such an action as **simple action**.

CDR-MARL: similar to DR-MARL, UEs perform exploration by taking simple actions in the learning period; in addition, followed by each simple action of a UE, all other UEs simultaneously adjust their connections accordingly; we name such a two-step action as **compound action** (to be discussed further in the next section).

[2]This is because all UEs are homogeneous and use a common utility function.

IV. APPLICATION OF DIFFERENCE REWARD

With the preparation of some basics given in the previous section, we are able to discuss how to apply and implement the proposed ideas in detail.

A. MEASURED DR

The standard DR for agent i is dened as followsV

$$DR_i(a_i|s^t; u_{\ i}) \text{ D } r_G(s^t; (u_{\ i}; a_i)) \quad r_G(s^t; (u_{\ i}; c_i)) \quad (27)$$

where a_i is an action made by agent i, c_i is a baseline action (an action as a reference) of agent i, $u_{\ i}$ is the joint action of all agents except agent i, and $r_G(s^t; (u_{\ i}; a_i))$ and $r^t_G(s^t; (u_{\ i}; c_i))$ are the global rewards for joint actions $(u_{\ i}; a_i)$ and $(u_{\ i}; c_i)$, respectively. Then, the optimization problem dened in (22) can be interpreted in terms of rewards

asV

$$r_i \text{ D } \begin{cases} DR_i; & \text{if constraints (17)-(18) are met} \\ 0; & \text{if constraints (17)-(18) are not met} \end{cases} \quad i \text{ 2 I} \quad (28)$$

where r_i is the individual reward of agent i.

Unfortunately, this DR denition is not perfect for all practical applications since in some scenarios the reference total reward $r^t_G(s^t; (u_{\ i}; c_i))$ cannot be accurately obtained without change of state (to be discussed later). In previous studies [48], [49], the estimation method is often used to calculate r_G. In our work, standard DR is applied to the rst scheme (DR-MARL): $r^t_G(s^t; (u_{\ i}; c_i))$ and $r^t_G(s^t; (u_{\ i}; a_i))$ are measured in two consecutive steps, respectively, and then DR $D^t_i(a_i|s^t; u_{\ i})$ is calculated based on (27) and used as reward r_i. Here the baseline action c_i refers to ''no connection to any mBS.'' Rigorously speaking, the reference total reward $r^t_G(s^t; (u_{\ i}; c_i))$ can only be obtained approximately since the state cannot be kept the same as s^t.The pseudo code for two-step DR measurement is shown in Algorithm 1. The agents execute association actions sequentially, and the global reward can be calculated by MBS in online via communicating with UEs for each action. For each agent in a Round-robin cycle, rstly, the reference global reward is measured with the given agent performing the default action c_i; then, the regular global reward is measured with the given agent performing the action a_i according to the -greedy rule, and the DR estimate is obtained by subtracting the former measurement from the latter measurement. Although the regular DR estimation takes two steps, for a new UE to join the system, only the second step is required.

Algorithm 1 Two-Step Measurement of DR

for t=1:T **do**
  for i 2 I **do**
    The ith agent take action c^i.
    MBS accumulative $r^t_G(s^t; (u\ \ ^i; c^i))$.
    With probability select a random action a_i otherwise select a_i D $maxQ(o_i; a$ V$w)$.
    MBS accumulative $r^t_G(s^t; (u\ \ ^i; a^i))$.
    MBS calculate DR_i based on (27).
  end for
end for

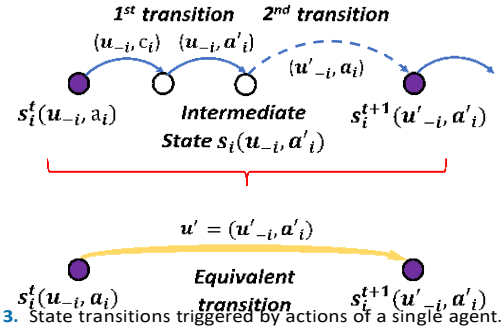


FIGURE 3. State transitions triggered by actions of a single agent.

The approximated and estimated (standard) DR has been proved effective. As an attempt to improve performance further, we extend the DR-based reward by incorporating concepts of compound action and compound-action-based DR.

1) COMPOUND ACTION

It is described in section IV-A and further explained in the following. As illustrated in Fig. 3, at current state $s^t_i(u\ \ _i; a_i)$, agent i takes an action a^c_i according to the -greedy rule, and then the system reaches an intermediate state $s_i(u\ \ _i; a^0_i)$. This process is called the 1^{st} transition. Immediately after this transition, all the other agents take a joint action $u^0\ \ _i$ which is a composite of individual actions executed concurrently based on the deterministic policy (w). This follow-up process is called the 2^{nd} transition. Because of the follow-up joint action, the outcome of a compound action provides additional information which benets the learning process. Two notes to the follow-up individual actions are: 1) in practice, likely only a few affected UEs need to take actions (change connections); 2) these follow-up actions need to be approved by an arbitrator at MBS, and when contention happens, i.e., multiple UEs content for a beam, a conict resolution mechanism (to be explained in subsection B) will be performed.

2) COMPOUND-ACTION-BASED DR (CDR)

The fundamental DR concept is to decouple the dependency of UEs' actions, though in practice perfect decoupling is hard to achieve. CDR is an extension of traditional DR and is dened as a compound action, anticipating some improvement over DR. Furthermore, we lose the requirement of common default action (c_i in the standard DR denition) by allowing to use any executed action as a reference, which simplies the CDR calculation. CDR can be expressed asV

$$CDR_i(_w(s^0)) \text{ D } r^{t\text{C}1}_G(s^{t\text{C}1}; (_w(s^0)\ \ _i; a^0))\ \ _i\ r^t\ (s^t_{\dot{G}}(u\ \ _i; a_i)) \quad (29)$$

Similar to DR measurement, CDR can be measured at MBS via communicating with all UEs.

3) EXTENDED DR (EDR)

It is expected that CDR can be more effective than the traditional DR. However, CDR is a function of $_w(s_0)\ \ _i$ which uctuates signicantly at the beginning of the learning phase. Based on the discussion above, we propose an EDR that combines DR and CDRV

$$EDR_i(; _w(s^0)) \text{ D } DR_i \text{ C } \quad CDR_i(_w(s^0)) \quad (30)$$

where (>0) is a variable weight that needs to increase gradually in the neural network training phase, and $_w(s^0)$ is a compound action based on a deterministic decision $_w(s)$. EDR is applied to EDR-MARL scheme. It used as the individual reward: r_i D $EDR_i(; _w(s^0))$ if constraints (17)-(18) are met; r_i D 0 if constraints (17)-(18) are not met. The acquisition of EDR is described in pseudo code in Algorithm 2.

Algorithm 2 EDR Acquisition Process

for i 2 I **do**
  Take the ith agent as the learner and the rest of agents as the actor.
  for tD1:T **do**
    MBS accumulative $r^t_G(s^t; u)$.
    With probability select a random action a_i otherwise select a_i D $maxQ(o_i; a$ V$w)$.
    Calculate the DR for the action a_i base on Algorithm 1.
    Actors take actions according to deterministic policy (w).
    MBS accumulative $r^{t\text{C}1}_G(s^{t\text{C}1}; u^0)$.
    MBS calculate the EDR_i, based on (30), variable weight D 1 .
  end for
end for

B. EDR-MARL SCHEME

EDR-MARL is relatively more complicated than DR-MARL and described in detail in this subsection. The roles and relationships in the proposed EDR-based scheme are shown in Fig. 4, where learner and actor are used for better explanation [50], r_G is the total reward mentioned earlier, and s is the joint state dened in (24). In the learning phase of this scheme, in each epoch,[3] all agents are divided into two parts, i.e., the learner played by one of the agents and actors

[3] An epoch is a section of learning period in which a learner along with all actors take many cycles of actions until not much improvement can be achieved.

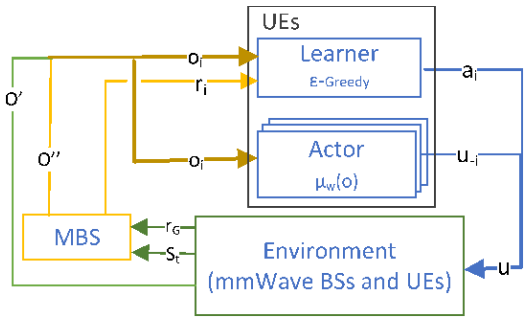


FIGURE 4. Different roles and relationships among them in the proposed CRD-MARL scheme (training phase).

played by the rest of agentsall actors are treated as a whole. Recalling Fig. 3, in the 1^{st} transition, the learner performs exploration, interacts with actors to generate trajectories and uploads the exploration experience to MBS. In the 2^{nd} transition, the actors execute (follow-up) actions reexively based on their policy learned so far. Obviously, an actor needs to retrieve the latest policy parameters from MBS before each follow-up action. The agents take turns playing the learner role from one epoch to another epoch. In other words, this MARL scheme is performed sequentially. MBS conducts a number of functions, including providing each agent with network information and informing the learner of the reward for each joint action, all via broadcasting.

Fig. 5 shows the proposed EDR-MARL scheme with detailed interaction ow. In the initial stage of the system, all UEs download the learned neural network parameters (w) from the MBS, and MBS selects a UE as the learner. The learner observes the environment and performs the exploration action a_i; with the execution of a_i, the environment changes, and then the actors (the rest of agents) observe this change independently; each corresponding follow-up action plan and its Q-value are uploaded to MBS, and then these individual actions of all agents form a joint action u. As mentioned earlier, an arbitration process is needed to ensure collision-free follow-up actions. When a conict is detected by MBS, the involved agent with a lower Q-value

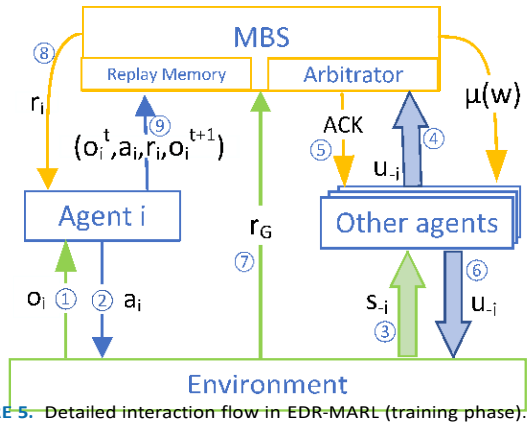


FIGURE 5. Detailed interaction flow in EDR-MARL (training phase).

TABLE 2. Simulation parameters.

Parameter	Value
Carrier frequency	28 GHz
bandwidth W	1 GHz
Per-UE power ceiling P_{ceil}	23 dBm
Noise figure+implementation losses	10 dB
Max number of associated mBSs per UE	3
Number of mBSs	15
Number of beams per BS	22
Area	100 m × 100 m
Maximum number of neighboring mBSs to consider	5
Close-in reference distance (d_0)	1 m
mBS mainlobe gain	24 dBi
mBS sidelobe gain	-16 dBi
UE mainlobe gain	9 dBi
UE sidelobe gain	-6 dBi
UE thermal noise density N_0	-174 dBm/Hz
Type of area	urban (use of (4))
Path loss exponent $n_{i,j}$	2.17 (LOS) and 2.51 (NLOS) [41]
Standard deviation of shadowing parameter $\Lambda_{i,j}$ (dB)	3.94 (LOS) and 6.95 (NLOS) [41]
Tuning constant κ	10
Learning rate α	0.01
Discounting factor γ	0.9
Batch size	200
Activation function	ReLU
Number of hidden layers	6
ϵ	$1 \rightarrow 0.1$
Target-DQN updating frequency T_u	50
Size of replay memory $\mathcal{D}$	1000

receives a rejection message and is required to resubmit a new action plan. If not a single collision is detected, the actors are allowed to take the approved follow-up actions. MBS obtains the global reward 0, calculates the learner's EDR using (30), and informs the learner of it. Then, the learner uploads the action trace $(s; a; r; s^0)$ to MBS, and MBS updates the neural network parameters with this data and sends the new parameters to all UEs.

The detailed process of EDR-MARL can be seen in the pseudo code of EDR-MARL shown in Algorithm 3 on the next page, where double deep Q-learning (DDQN) [51] is employed. DDQN is a widely-used type of RL and the goal of RL is to nd the optimal policy (s) for each state s to maximize the expected return. The update rule of DDQL algorithm is as followsV

$$Q(\boldsymbol{o}_i^t; \boldsymbol{a}_i^t) \mathrel{D} (1 \quad) Q(\boldsymbol{o}_i^t; \boldsymbol{a}_i^t) \mathrel{C} \boldsymbol{r}_i(\boldsymbol{o}^t; \boldsymbol{a}^t) \mathrel{C} \max_{\boldsymbol{a}_i^{\zeta}} Q^{0i}(\boldsymbol{o}_i^{t\mathrm{C}1}; \boldsymbol{a}^0)_i \qquad (31)$$

where is the learning rate, is the discount factor. The training process is to adjust the neural network to minimize the loss function $L()$,

$$L_i() \mathrel{D} \mathsf{E}[(\boldsymbol{r}_i(t) \mathrel{C} \max Q(\boldsymbol{\delta}^{t\mathrm{C}1}; \boldsymbol{a}^{t\mathrm{C}1}{}_i | \quad) Q(\boldsymbol{o}_i^t; \boldsymbol{a}_t |))^2] \qquad (32)$$

where is the weight of the target network, is the weight of the behavior network, and Q is the target Q-function.

V. EXPERIMENTAL VALIDATION

This section demonstrates the effectiveness of the proposed EDR-MARL scheme via evaluating the performance with

Algorithm 3 EDR-MARL for User Multi-Association

Initialize the mmWave UDN environment
Initialize repay memory D to capacity N.
Randomly initialise the Q-network $Q(o; a|w)$
Randomly initialise the target Q-network $\hat{Q}(o; a|w^0)$
while Not convergence **do**
  Download Q-network weight for each UE from MBS.
  for i 2 I **do**
    Take the ith agent as the learner and the rest of agents as the actor.
    for tD1:T **do**
      Agent i observe $\boldsymbol{o}_i(t)$ from the environment and MBS
      With probability take a random action a_i otherwise take a_i D $argmax Q_k(o_i; a$ V$w)$.
      while Not conict **do**
        for k 2 I,k D i **do**
          Agent k select action a_k D $argmax Q(o_k; a$ V $w)$.
          Upload action a_k and $max(Q_k)$ to MBS.
          if ACKD0 **then** {/*Conict*/}
            Remove $max(Q_k)$ from Q_k.
            Upload action a_k D $argmax Q_k^c(o_k; a$ V$w)$ and $max(Q_k^0)$ to MBS.
          end if
        end for
      end while
      Take compound action u_i.
      Calculate the EDR based on (30).
      Store tuple $(\boldsymbol{o}_i(t); \boldsymbol{a}_i(t); \boldsymbol{r}_i(t); \boldsymbol{o}_i(t$ C 1)) in D
      Sample random batch from D.
      Perform a gradient descent step on (31) with respect to the network parameters w.
      Minimize the loss (32) using stochastic gradient.
      The agent updates the target Q-network weights w^0 once per T_u steps with w^0 D w.
      Upload Q-network weight w to MBS.
    end for
  end for
  Update new location for each UE
end while

numerical results. We consider a simplied heterogeneous mmWave UDN with one MBS, 15 mBSs, and 40-180 UEs randomly placed in a 100 100 m^2 area. All simulation parameters are listed in Table 2. For UE deployment, two different distributions are considered. In the rst one, UEs are uniformly scattered over the entire area whereas, in the second one, UEs is distributed by following a 2-dimensional symmetric Gaussian distribution with a random center and standard deviation around 47 m. In the following, all simulation results are generated by averaging over at least 50 independent realizations.

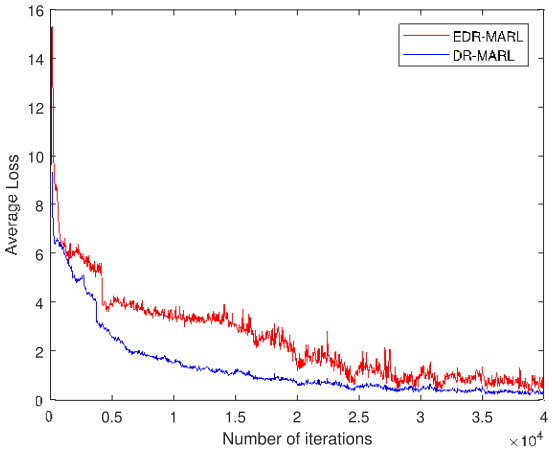


FIGURE 6. Convergence of DR-MARL and EDR-MARL.

In addition to the proposed two MARL schemes, DR-MARL and EDR-MARL, the following three schemes (user association strategies) are considered for comparison purposesV

Max-SNR single connectivity: Each UE connects to a single mBS that leads to the strongest SNR;
Max-SNR double connectivity: Each UE connects to two mBSs that correspond to the rst two strongest SNR;
Max-SNR triple connectivity: Each UE connects to three mBSs that correspond to the rst three strongest SNR;

For the three schemes using xed numbers of connections, when a UE cannot be provided the required number of beams (due to poor link quality or no available beams), the system will simply stop accepting the UE and accordingly its QoS requirement is not satised.

Firstly, Fig. 6 shows the convergence behaviors of the proposed DR-MARL and EDR-MARL algorithms at typical learning rates 0.01. We observe that both learning curves converge, but the curve of DR-MARL is relatively smoother, while the other one has spikes and converges a little slowly. Recalling (29) and (30), the less smoothness is mainly due to that EDR-MARL involves policy (w) which is more random in the early learning phase. Note that although EDR-MARL behaves less smoothly, it leads to better outcomes thanks to the compound actions.

Fig. 7 shows the QoS dissatisfaction rate (dened as OoS outage probability) for QoS_{req} D 1.2 and 1.8 Gbit/s, and different numbers of UEs. Among the three schemes with xed-number connectivity, the single-link one is a safer option, but, because of mmWave blockage, it cannot meet QoS well even when there are a very small number of UEs; while both the double-link and triple-link schemes demonstrate dramatically increased QoS dissatisfaction as the numbers of UEs increase and excess certain values, because equally assigning beams to all UEs will quickly use up the limited number of available beams. In contrast, the proposed DR and EDR methods perform unanimously well for either

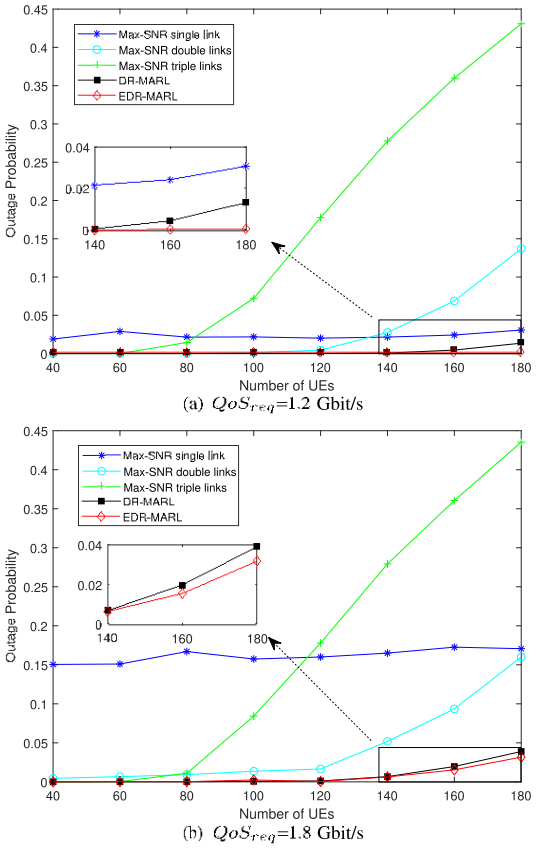


FIGURE 7. QoS outage probability versus number of UEs (uniform distribution).

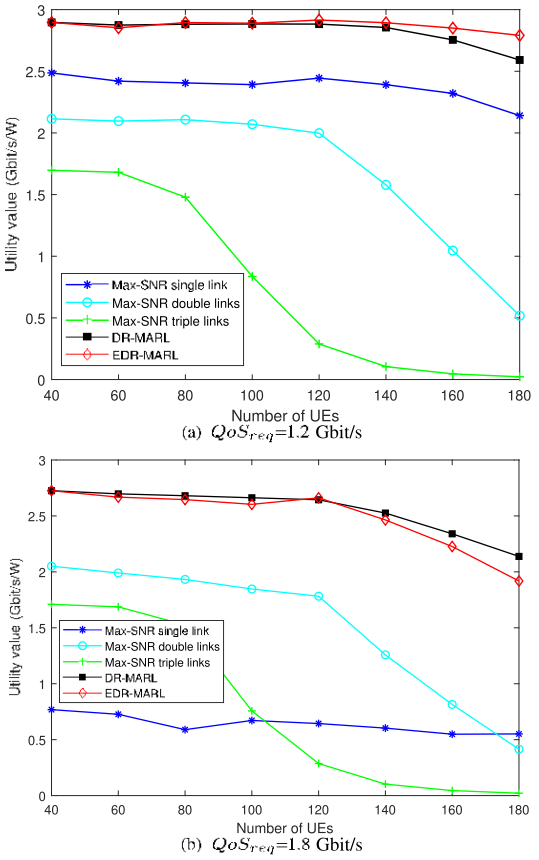


FIGURE 8. Comparison of utility values under different numbers of UEs (uniform distribution).

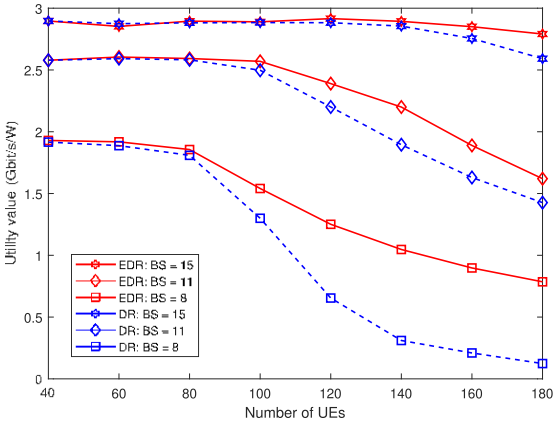


FIGURE 9. Impact of mBS density and UE density (QoS_{req} 1.2 Gbit/s, uniform UE distribution).

QoS_{req} D 1.2 and 1.8 Gbit/s over a wide range UE density, implying their excellent adaptability to the environment. Furthermore, the EDR outperforms the DR, especially when the number of UEs reaches the high end of the considered range: the QoS outage ratio of the EDR to the DR is only 23% when there are 180 UEs in the considered example.

The superiority and robustness of proposed schemes can be observed as well in Fig. 8 that compares the utility values of various algorithms, where the utility function considers both energy efciency and QoS outage. Again, the results suggest that a connection strategy with a xed number of connections is unsuitable in the mmWave UDN environment. The energy efciency of a single-connection scheme is usually the highest, but it suffers higher blockage; in consequence, its utility is low though it is not quite sensitive to the change of UE density. Both the double-link and triple-link schemes are sensitive to the change of UE density and drop quickly as the number of UEs increases, but the latter exhibits worse utility mainly due to higher unavailability of mmWave beams.

In the case of a small number of UEs, the performances of DR and EDR algorithms are close, indicating that their adaptive RL decision methods can achieve the same level of exible association when resources are sufcient. As the number of UEs is further increased, CDR becomes more effective, which might be due to that the compound action provides more information, of course, at the cost of slightly increased complexity.

The number of UEs and the number of BSs jointly affect the performance, which can be seen in Fig. 9. As expected, the EDR outperforms the DR, and adding more BSs is benecial, but increasing the number of UEs degrades the performance. Interestingly, with fewer BSs, the EDR becomes signicantly better than the DR as the number of UEs increases (see the utility value gas for 8 BSs and 160 UEs).

Finally, Fig. 10 shows the behaviors of the proposed schemes in the two UE distributions, i.e., uniform and nonuniform. Interestingly, for either DR or EDR schemes, there is a performance switch phenomenon: the performance

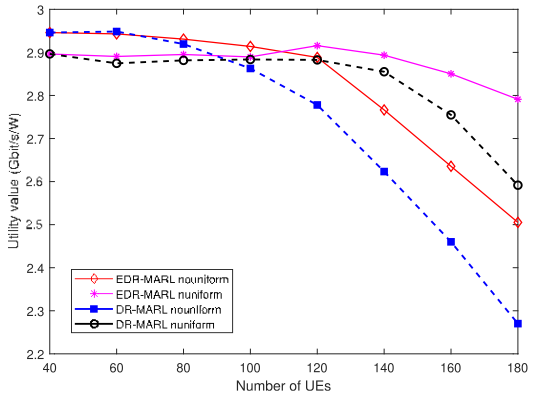


FIGURE 10. Performance comparison of uniform and nonuniform distributions of UEs (QoS_{req}1.2 Gbit/s).

of the nonuniform case is better than that of the uniform case at a small number of UEs (low UE density), but as the number of UEs is over a turning point (about 92 for the DR and 110 for the EDR), the performance of the nonuniform case becomes worse than the other one. This phenomenon can be explained as follows. With nonuniform UE distribution, there is a better chance that many UEs can connect to BSs at short distances if there are sufcient resources (mmWave beams); as the UE density increases, UE concentration in an area causes more resource contention which has a negative impact on the performance.

VI. CONCLUSION

We have systematically studied two mmWave multi-association schemes, DR-MARL and EDR-MARL. An important and essential feature, scalability, is enabled mainly by two mechanisms: a) the observation vector is designed such that the dimension is xed, and b) individual actions are taken by UEs independently. To effectively deal with a relatively large MARL system with a large action space, the traditional DR technique is implemented via a proposed measurement procedure. With the introduced compound action and compound-action-based DR, an extended version of DR is proposed to improve performance further. Simulation results have validated the superiority of the two proposed schemes over other schemes and demonstrate their robustness over a large range of UE density.

Our future work in this line of research includes: 1) Assessment of overall complexity, including computational complexity, communication overhead and run time in the training phase. 2) The use of a digital twin in the training phase to accelerate the process, if the environment can be modeled. 3) The use of transfer learning is to accelerate the learning process and mitigate the impact of UE mobility.

REFERENCES

[1] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 25222545, 4th Quart., 2016.

[2] M. I. Kamel, W. Hamouda, and A. M. Youssef, "Multiple association in ultra-dense networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 16.

[3] M. Saimler and S. Coleri, "Multi-connectivity based uplink/downlink decoupled energy efcient user association in 5G heterogenous CRAN," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 858862, Apr. 2020.

[4] G. Yu, R. Liu, Q. Chen, and Z. Tang, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 7985, Jun. 2018.

[5] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 7885, Apr. 2016.

[6] Y. Lin, R. Zhang, C. Li, L. Yang, and L. Hanzo, "Graph-based joint user-centric overlapped clustering and resource allocation in ultradense networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 44404453, May 2018.

[7] Z. Wu, Z. Fei, Z. Han, and L.-C. Wang, "Distributed user-centric clustering and base station mode choose in ultra dense networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 17.

[8] R. Liu, G. Yu, and G. Y. Li, "User association for ultra-dense mmWave networks with multi-connectivity: A multi-label classication approach," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 15791582, Dec. 2019.

[9] A. Chen, S. Li, K. Jin, and Z. Tang, "Energy-efcient multi-connectivity enabled user association and downlink power allocation in mmWave networks," in *Proc. Wireless Telecommun. Symp. (WTS)*, Apr. 2022, pp. 16.

[10] G. Simsek, H. Alemdar, and E. Onur, "Multi-connectivity enabled user association," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 16.

[11] C. Chaieb, Z. Mlika, F. Abdelke, and W. Ajib, "On the optimization of user association and resource allocation in HetNets with mm-wave base stations," *IEEE Syst. J.*, vol. 14, no. 3, pp. 39573967, Sep. 2020.

[12] X. Cai, A. Chen, L. Chen, and Z. Tang, "Joint optimal multi-connectivity enabled user association and power allocation in mmWave networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 16.

[13] K. Jin, X. Cai, J. Du, H. Park, and Z. Tang, "Toward energy efcient and balanced user associations and power allocations in multi-connectivity enabled mmWave networks," *IEEE Trans. Green Commun. Netw.*, early access, May 3, 2022, doi: 10.1109/TGCN.2022.3172355.

[14] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 2638, Nov. 2017.

[15] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 31333174, 4th Quart., 2019.

[16] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 22242287, 3rd Quart., 2019.

[17] J. J. Hernandez-Carlon, J. Perez-Romero, O. Sallent, I. Vila, and F. Casadevall, "Deep learning-based multi-connectivity optimization in cellular networks," in *Proc. IEEE 95th Veh. Technol. Conf., (VTC-Spring)*, Jun. 2022, pp. 15.

[18] X. Liu, J. Yu, Z. Feng, and Y. Gao, "Multi-agent reinforcement learning for resource allocation in IoT networks with edge computing," *China Commun.*, vol. 17, no. 9, pp. 220236, Sep. 2020.

[19] A. Sacco, F. Esposito, G. Marchetto, and P. Montuschi, "Sustainable task ofoading in UAV networks via multi-agent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 50035015, May 2021.

[20] X. Huang, S. Leng, S. Maharjan, and Y. Zhang, "Multi-agent deep reinforcement learning for computation ofoading and interference coordination in small cell networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 92829293, Sep. 2021.

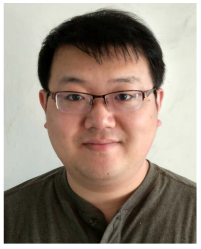
[21] H. Lu, C. Gu, F. Luo, W. Ding, S. Zheng, and Y. Shen, "Optimization of task ofoading strategy for mobile edge computing based on multi-agent deep reinforcement learning," *IEEE Access*, vol. 8, pp. 202573202584, 2020.

[22] J. Moon, H. Ju, S. Kim, and B. Shim, "Energy-efcient mmWave UDN using distributed multi-agent deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 16.

[23] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 35073523, Jun. 2021.

[24] N. Yang, H. Zhang, and R. Berry, "Partially observable multi-agent deep reinforcement learning for cognitive resource management," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2020, pp. 16.

[25] Z. Li, C. Wang, and C.-J. Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 22172228, Aug. 2017.
[26] G. Kwon and H. Park, "Joint user association and beamforming design for millimeter wave UDN with wireless backhaul," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 26532668, Dec. 2019.
[27] 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," *IEEE Access*, vol. 7, pp. 111790111800, 2019.
[28] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 45354548, Jul. 2020.
[29] M. Sana, A. D. Domenico, W. Yu, Y. Lostanlen, and E. C. Strinati, "Multi-agent reinforcement learning for adaptive user association in dynamic mmWave networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 65206534, Oct. 2020.
[30] T. H. L. Dinh, M. Kaneko, K. Wakao, K. Kawamura, T. Moriyama, H. Abeysekera, and Y. Takatori, "Deep reinforcement learning-based user association in sub 6GHz/mmWave integrated networks," in *Proc. IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2021, pp. 17.
[31] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.
[32] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 19.
[33] P. Sunehag, G. Lever, A. Gruslys, W. Marian Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning," 2017, *arXiv:1706.05296*.
[34] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 42954304.
[35] S. Vanneste, A. Vanneste, S. Bosmans, S. Mercelis, and P. Hellinckx, "Learning to communicate with multi-agent reinforcement learning using value-decomposition networks," in *Proc. Int. Conf. P2P, Parallel, Grid, Cloud Internet Comput.* Cham, Switzerland: Springer, 2019, pp. 736745.
[36] D. H. Wolpert and K. Tumer, "Optimal payoff functions for members of collectives," in *Modeling Complexity in Economic and Social Systems*. Singapore: World Scientic, 2002, pp. 355369.
[37] J. Castellini, S. Devlin, F. A. Oliehoek, and R. Savani, "Difference rewards policy gradients," 2020, *arXiv:2012.11258*.
[38] C. Sun, X. Q. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam division multiple access transmission for massive MIMO communications," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 21702184, Jun. 2015.
[39] H. S. Vu, K. Truong, and M. T. Le, "Beam division multiple access for millimeter wave massive MIMO: Hybrid zero-forcing beamforming with user selection," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 12, no. 1, p. 445, Feb. 2022.
[40] J. Ko, Y.-J. Cho, S. Hur, T. Kim, J. Park, A. F. Molisch, K. Haneda, M. Peter, D.-J. Park, and D.-H. Cho, "Millimeter-wave channel measurements and analysis for statistical spatial channel model in in-building and urban environments at 28 GHz," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 58535868, Sep. 2017.
[41] K. Zhang, R. Zhang, J. Wu, Y. Jiang, and X. Tang, "Measurement and modeling of path loss and channel capacity analysis for 5G UMa scenario," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 15.
[42] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, Jr., "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403430, Jan. 2016.
[43] M. Dianati, X. Shen, and S. Naik, "A new fairness index for radio resource allocation in wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, vol. 2, May 2005, pp. 712717.
[44] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, "Deep learning-based beam management and interference coordination in dense mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592603, Jan. 2019.
[45] M. Kamel, W. Hamouda, and A. Youssef, "Performance analysis of multiple association in ultra-dense networks," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 38183831, Sep. 2017.
[46] R. S. Sutton, "Temporal credit assignment in reinforcement learning," Ph.D. dissertation, Dept. Comput. Inf. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, 1984.
[47] B. Slantchev, "Game theory: Static and dynamic games of incomplete information," Dept. Political Sci., Univ. San Diego, San Diego, CA, USA, White Paper, 2008.
[48] S. Proper and K. Tumer, "Modeling difference rewards for multiagent learning," in *Proc. AAMAS*, 2012, pp. 13971398.
[49] M. K. Colby, W. J. Curran, and K. Tumer, "Approximating difference evaluations with local information," in *Proc. AAMAS*, 2015, pp. 16591660.
[50] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. AAAI/IAAI*, nos. 746752, 1998, p. 2, 1998.
[51] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 17.



XUEBIN LI received the M.S. degree from the Changchun University of Science and Technology, Changchun, China, in 2010. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Tennessee Technological University, USA. His research interests include deep reinforcement learning, resource allocation for wireless communication systems, and combinatorial optimization.

TERRY N. GUO (Senior Member, IEEE) received the M.S. degree in telecommunications engineering from the Beijing University of Posts and Telecommunications, Beijing, in 1990, and the Ph.D. degree in communications and electronic systems from the University of Electronic Science and Technology of China, Chengdu, in 1997. From January 1997 to December 1999, he was a Postdoctoral Research Fellow at the University of California at San Diego, San Diego. He worked for a few technology companies in NS, USA, in the early 2000s. Since 2004, he has been with Tennessee Technological University, where he is currently a Research Professor. He has authored or coauthored more than 90 peer-reviewed articles and is a well-established researcher in wireless communications and sensing. His recent research interests include integrated sensing and communications, the IoT security, resource allocation, and digital twin for wireless networks.

ALLEN B. MACKENZIE (Senior Member, IEEE) joined Tennessee Tech as the Chair, in August 2019, and a Professor with the Department of Electrical and Computer Engineering. Prior to joining Tennessee Tech, he was a Professor with the Department of Electrical and Computer Engineering, Virginia Tech, where he was a Faculty Member, from 2003 to 2019, where he was the Associate Director of Wireless @ Virginia Tech. From 2012 to 2013, he was an E. T. S. Walton Visiting Professor at the Trinity College Dublin. He is the author of more than 90 refereed conference and journal articles and a coauthor of the book *Game Theory for Wireless Engineers*. His research interests include wireless communications systems and networks. His current research interests include integration of millimeter wave technology into networks, cognitive radio and cognitive network architectures, and the analysis of wireless systems and networks using game theory and stochastic optimization. His past and current research sponsors include the National Science Foundation, Science Foundation Ireland, the Defense Advanced Research Projects Agency, and the National Institute of Justice.

He is a member of the ASEE and ACM. He was previously on the Editorial Board of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE TRANSACTIONS ON MOBILE COMPUTING. He was a member of the U.S. Department of Commerce's Spectrum Management Advisory Committee (CSMAC), from 2016 to 2018.