Multi-agent deep reinforcement learning based user association for dense mmWave networks

Mohamed Sana, Antonio de Domenico, Emilio Calvanese Strinati

Multi-Agent Deep Reinforcement Learning based User Association for Dense mmWave Networks

Mohamed Sana[1], Antonio De Domenico[1], Emilio Calvanese Strinati[1]
[1]CEA-Leti Minatec Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 09, France
Email : {mohamed.sana, antonio.de-domenico, emilio.calvanese-strinati}@cea.fr

***Abstract*—Finding the optimal association between users and base stations that maximizes the network sum-rate is a complex task. This problem is combinatorial and non-convex, and is even more challenging in millimeter-wave networks due to beamforming, blockages, and severe path loss. Despite the interest that this problem has gained over the last years, the various solutions proposed so far in the literature still fail at being flexible, computationally effective, and suitable to the dynamic nature of mobile networks. This paper addresses these issues with a novel distributed algorithm based on multi-agent reinforcement learning. More specifically, we model each user as an agent, which, at each time step, maps its observations to an action corresponding to an association request to a base station in its coverage range. Our numerical results show that the proposed solution offers near optimal performance and thanks to its flexibility, provides large sum-rate gain with respect to the state-of-art approaches.**

I. Introduction

The fifth generation (5G) and beyond mobile communication systems promise to meet the increasing demand for ultra-high-speed communications. To achieve this goal, among the different technologies currently under consideration, millimeter-wave (mmWave) communication has attracted a particular attention due to the large and mostly unlicensed spectrum resources available between 24 and 86 GHz. However, transmissions at mmWave frequencies suffer from severe attenuation, blockage, and deafness [1]. Benefiting from the short wavelength characteristic of mmWaves, directional beamforming enables to overcome part of these issues [2]. With the dense deployment of mmWave small cells, though, co-channel interference can be detrimental for the network performance. Optimally associating users in this context can be very challenging. In the literature, efficiently solving the problem of user association in mmWave networks while taking into account inter-cell and intra-cell interference has received little attention. Athanasiou *et al.* have proposed a distributed algorithm for managing the user association [3]. However, they have made the assumption that the interference is negligible, thus achieving sub-optimal performance. Lui *et al.* have designed a decentralized algorithm for beam pair selection between user equipment (UEs) and base stations (BSs), where the sum-rate maximization problem is reformulated as a non-cooperative game with local interactions [4]. This approach achieves good performance in terms of convergence speed and sum-rate; however, each UE has its own utility function whose computation requires information exchange with its neighbouring UEs, thus, leading to significant signaling overhead especially in dense networks.

Recently, the use of machine learning and reinforcement learning has attracted the interest of both academia and industry to handle the network management complexity [5], [6]. Zhou *et al.* proposed a deep neural network (DNN) architecture to optimize the beam management in order to maximize the sum-rate subject to power and beam width constraints [7]. However, this centralized method requires collecting the signal-to-noise ratio (SNR) information of all links. Moreover, the training data is generated using a heuristic algorithm, which reduces the performance of the DNN. Zhao *et al.* designed a distributed algorithm based on multi-agent reinforcement learning (MARL) for user association in heterogeneous network [8]. In contrast to our work, they did not focus on mmWave networks and they considered fully observable environment i.e., agent decisions are based on global state information, which requires message passing among UEs and therefore, induces signaling overhead.

Both distributed and centralized schemes have specific advantages and inconveniences. Centralized schemes allow efficient resource management since information from all nodes is collected and processed globally but may lead to large overhead. In contrast, distributed schemes require less information exchange but may converge to sub-optimal solutions, due to the lack of global information.

Inspired by the aforementioned studies, we propose a distributed deep MARL framework for user association to maximize the network sum-rate. We model each UE as an agent operating in a fully distributed manner. No information is exchanged between UEs, which learn the optimal strategy only from their local state, thus limiting the signaling overhead and the algorithm complexity. In addition, by using reinforcement learning, there is no need of expert database. Moreover, in contrast to some previous works, we consider both intra-cell and inter-cell interference. Our numerical results show that the proposed solution has near optimal performance providing large sum-rate gain with respect to the state-of-art approaches.

The rest of the paper is organized as follows. Section II introduces the system model and formulates the optimization problem to be solved. Section III details the proposed approach while Section IV outlines the simulation results. Section V concludes the paper and gives some perspectives on future work.

II. System Model and Problem Formulation

A. System model

We consider a downlink network where N_s mmWave small cells are deployed in the presence of an overlay macro cell to provide services to K UEs as described in Fig. 1. Let $\mathcal{S} = \{0, 1, 2, ..., N_s\}$ be the set of $N_s + 1$ base stations (BSs) in the network with 0 indexing the macro base station (MBS). Each small cell i covers a set $\mathcal{U}_i$ of UEs. Accordingly, $\mathcal{U} = \cup_{i=1}^{N_s} \mathcal{U}_i = \{1, 2, ..., K\}$ represents the set of all UEs in the macro cell, which in turn uses a sub-6GHz band to provide ubiquitous coverage.

In this architecture with multi-radio access technologies [9], a UE may receive control signals from multiple BSs, i.e., in a multi-connectivity setting [10]. We use $\mathcal{S}_j = \{i, \quad d_{i,j} \leq c_i\}$[1] to denote the set of BSs the UE j could connect to, where c_i is the coverage range of the BS i and $d_{i,j}$ is the distance between BS i and UE j. Let $x_{i,j} \in \{0, 1\}$ be the binary association variable such that $x_{i,j} = 1$ when UE j is served by the BS i and $x_{i,j} = 0$ otherwise. Here we assume that each UE can only received data by one BS at a time; moreover, we consider that each mmWave small cell BS (SBS) cannot serve more than N_i UEs simultaneously, where N_i is the maximum beams available at the SBS i. In our system model, we consider that the SBSs allocate all the available mmWave band to each served UE; in contrast, the MBS equally shares its band across the served UEs. Finally, we consider that the SBSs have already performed beam training and alignment mechanisms in advance and therefore are able to configure the appropriate beam when a data connection is set up.

The downlink signal-to-interference-plus-noise ratio (SINR) between BS i and UE j is evaluated as follows:

$$\mathrm{SINR}_{i,j} = \frac{p_i g^t_{i,j} g^c_{i,j} g^r_{i,j}}{I^{\mathrm{intra}}_{i,j} + I^{\mathrm{inter}}_{i,j} + N_0 B_{i,j}}, \tag{1a}$$

$$I^{\mathrm{intra}}_{i,j} = \sum_{j' \in \mathcal{U}_i / \{j\}} x_{i,j'} p_i g^t_{i,j' \to i,j} g^c_{i,j} g^r_{i,j' \to i,j}, \tag{1b}$$

$$I^{\mathrm{inter}}_{i,j} = \sum_{(i',j') \in \mathcal{V}} x_{i',j'} p_{i'} g^t_{i',j' \to i,j} g^c_{i',j'} g^r_{i',j' \to i,j}, \tag{1c}$$

where p_i is the transmit power of BS i. $I^{\mathrm{inter}}_{i,j}$ and $I^{\mathrm{intra}}_{i,j}$ are the inter-cell and the intra-cell interference on the link (i, j), respectively. $\mathcal{V} = \mathcal{S} \backslash \{i\} \times \mathcal{U} \backslash \{j\}$ is the set of inter-cell interfering links with respect to link (i, j), N_0 is the background noise power spectrum density, and $B_{i,j}$ the bandwidth allocated from the BS i to the UE j. $g^c_{i,j}$ denotes the channel gain between BS i and UE j, which captures the effect of path loss and shadowing as follows:

$$g^c_{i,j}(dB) = -20\log_{10}\left(\frac{4\pi d_0}{\lambda_i}\right) - 10\eta_i \log_{10}\left(\frac{d_{i,j}}{d_0}\right) - X_{i,j}, \tag{2}$$

where d_0 is the close-in-free-space reference distance, η_i the path loss coefficient, λ_i the wavelength, and $X_{i,j}$ the shadowing coefficient. $g^t_{i,j}$ and $g^r_{i,j}$ are the transmitter and receiver antenna directivity gain in the communication link, respectively. We adopt the symbolic notation $g^t_{i',j' \to i,j}$ and $g^r_{i',j' \to i,j}$ to denote the transmitter and receiver directivity gain of the interfering link (i', j') seen from the link (i, j), respectively.

[1] $\mathcal{S}_j$ can also be derived based on links quality, e.g., the received signal strength indicator between UE j and BS i ($\mathrm{RSSI}_{i,j}$) should be greater than a predefined threshold κ_j, i.e., $\mathcal{S}_j = \{i, \quad \mathrm{RSSI}_{i,j} \geq \kappa_j\}$

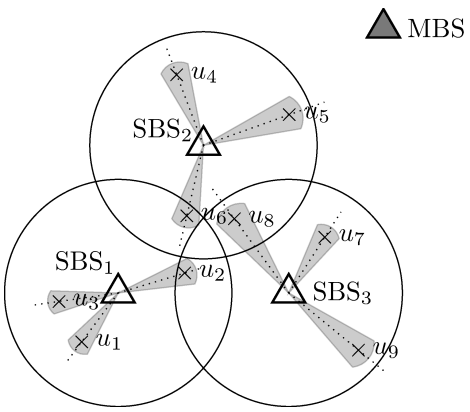

Fig. 1. A downlink network with $N_s = 3$ SBSs, one MBS, and $K = 9$ UEs. $\mathcal{U}_1 = \{1, 2, 3, 6\}$, $\mathcal{U}_2 = \{4, 5, 6, 8\}$, $\mathcal{U}_3 = \{2, 7, 8, 9\}$. As an example, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{1, 3\}$, $\mathcal{S}_8 = \{2, 3\}$.

To evaluate these gains, we adopt the simplified and commonly used sectored antenna model [7], [11] where the beamforming gain is defined as:

$$g(\theta, \alpha) = \begin{cases} g_0 \frac{2\pi - (2\pi - \theta)\xi}{\theta}, & \text{if } |\alpha| \leq \frac{\theta}{2} \\ g_0 \xi, & \text{otherwise} \end{cases}, \tag{3}$$

where ξ $(0 < \xi \ll 1)$ is the side lobe gain, α is the beam offset angle to the main lobe in radian, θ is the beam width in radian, and g_0 is the antenna gain.

B. Problem formulation

According to the system model described in Section II-A, we can formulate the user association problem for sum-rate maximization as follows:

$$\underset{\{x_{i,j}\}}{\text{maximize}} \quad \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{U}} x_{i,j} B_{i,j} \log_2(1 + \mathrm{SINR}_{i,j}), \tag{4a}$$

$$\text{subject to} \quad x_{i,j} \in \{0, 1\}, \tag{4b}$$

$$\sum_{j \in \mathcal{U}_i} x_{i,j} \leq N_i, \ i \in \mathcal{S} \backslash \{0\}, \tag{4c}$$

$$\sum_{i \in \mathcal{S}_j} x_{i,j} = 1, \ j \in \mathcal{U}. \tag{4d}$$

The constraint (4b) ensures that $x_{i,j}$ are binary variables and the constraint (4c) indicates that the maximum number of UEs associated with a given SBS i is limited to N_i. Note that we suppose that the number of UEs that can be simultaneously served by the MBS is larger than K. Finally, constraint (4d) ensures that each UE is served by exactly one BS. It is noteworthy that the SINR of the link (i, j) does not only depend on variable $x_{i,j}$ but also on the association of other users through the interference terms in the denominator. As a result, the objective function (4a) is non-convex, making

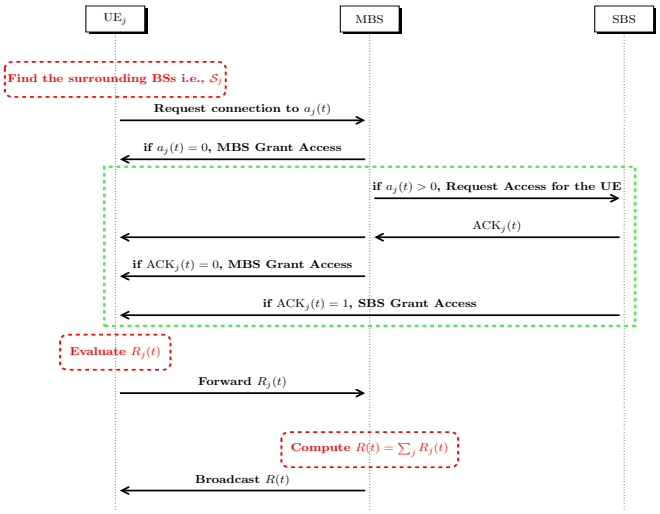

Fig. 2. Message sequence chart of the proposed mechanism.

this problem difficult to solve with the standard optimization techniques.

III. Proposed Solution via Deep multi-agent Reinforcement Learning

A. General framework

We propose a cooperative multi-agent reinforcement learning algorithm for solving the optimization problem (4a)-(4d). Fig. 2 presents the information flow diagram of the proposed algorithm. Each UE j first identifies the set of BSs $\mathcal{S}_j$ it may connect to, which also represents its action space, i.e., the action $a_j(t) \in \mathcal{S}_j$ denotes the index of the BS to which the UE j requests connection at time t. Accordingly, each time step, the UE j takes an action $a_j(t)$ and informs the MBS to which BS it requests connection to. If $a_j(t) = 0$, the MBS grants the connection request and sets up communication. Otherwise, the MBS forwards the connection request to the corresponding SBS. Depending on the overall received requests, the SBS sends an acknowledgement signal ($\text{ACK}_j(t)$) to both the UE and the MBS. If $\text{ACK}_j(t) = 1$, the SBS grants a connection to the UE; otherwise, the MBS establishes the data link with the UE j. Next, each UE j evaluates the perceived data rate, i.e., $R_j(t) = B_{a_j(t),j}\log_2(1+\text{SINR}_{a_j(t),j})$ and forwards this value to the MBS. Then, the MBS computes the network sum-rate $R(t)$ and broadcasts it to the network UEs, which use this information to evaluate the goodness of their action, and to define future actions accordingly.

Following [12], the history $\mathcal{H}_j(t)$ denotes the set of all actions, observations, and measurements collected at UE j up to time t:

$$\mathcal{H}_j(t) = \{a_j(\tau), \text{ACK}_j(\tau), R_j(\tau), R(\tau)\}_{\tau=1}^{t}. \quad (5)$$

Moreover, we define the playing strategy of UE j at time t, $\pi_j(t)$, as a mapping from its history $\mathcal{H}_j(t-1)$ to a probability mass function over its action space $\mathcal{S}_j$. Therefore, each UE takes its actions following its own strategy without being aware of the actions taken by the other UEs.

As formulated, this problem falls in the class of cooperative *independent learners* [13], [14], since no information is exchanged between agents, which operate in a partially observable environment. In fact, at each time, UEs act based only on their local state information including the feedback (ACK, $R(t)$) perceived from the surrounding BSs. They learn to which BS to ask for a connection in order to maximize the overall network throughput. The most challenging issue with this setting is the non-stationarity of the environment, which is due to the simultaneous interactions of all the agents with it. This non-stationarity can lead to *shadowed equilibria* - an agent locally optimal action ends up being a globally sub-optimal action. Omidshafiei *et al.* successfully applied hysteretic Q-learning (first introduced by Matignon *et al.* [14]) to tackle the non-stationarity problem in MARL with partial observability [15]. They use deep recurrent Q-networks (DRQNs), which serves as a basis to our proposed algorithm.

B. Background on Hysteretic Deep Recurrent Q-Network

In the hysteretic deep recurrent Q-network (HDRQN) setting, each UE j acts as an independent learner with its own DRQN $Q_j(o_j(t), h_j(t-1), a_j(t)|\theta_j)$ (see Fig. 3). The input of the DRQN $o_j(t) = \{a_j(t-1), R_j(t-1), R(t-1), \text{ACK}_j(t-1)\}$ is the UE local observation, $h_j(t-1)$ is the recurrent neural network (RNN) hidden state, and θ_j represents the UE local DRQN weights. The use of RNN allows to maintain a local aggregation of previous observed states, i.e., $\mathcal{H}_j(t)$, which improves the average reward perceived when dealing with partial observability.

UEs learn by interacting with the environment. At each time t, from its observation $o_j(t)$, UE j takes action $a_j(t)$ following a policy (e.g., ϵ-greedy), perceives reward $r_j(t)$ and observes the new state $o_j(t+1)$. The resulting experience $e_j(t) = \{o_j(t), a_j(t), r_j(t), o_j(t+1)\}$ is stored into a local memory $\mathcal{M}_j$, which is used to speed up and stabilize the training process [16]. During the learning phase, each UE j concurrently samples a mini batch of experiences $\mathcal{B}_j$ from its local memory $\mathcal{M}_j$ with the help of *concurrent experience replay trajectories* (CERTs) [15] and updates its DRQN weights in order to minimize the hysteretic loss function:

$$L_j(\theta_j) = \mathbb{E}_{e_j^b(t)\sim\mathcal{B}_j}\left[(w_j^b \delta_j^b(t))^2\right]. \quad (6)$$

In (6), b indexes an entry in the mini batch of experiences $\mathcal{B}_j$. $\delta_j^b(t) = y_j^b(t) - Q_j(o_j^b(t), h_j^b(t-1), a_j^b(t)|\theta_j)$ denotes the temporal difference (TD) error between the output of the UE j DRQN and the target value

$$y_j^b(t) = r_j^b(t) + \gamma \max_{a'} Q_j(o_j^b(t+1), h_j^b(t), a'|\hat{\theta}_j), \quad (7)$$

where the parameter γ is referred as the discounting factor and $\hat{\theta}_j$ are the weights of the target DRQN. In the hysteretic Q-learning, the network weights are updated via a gradient descent algorithm with two distinct learning rates α and β

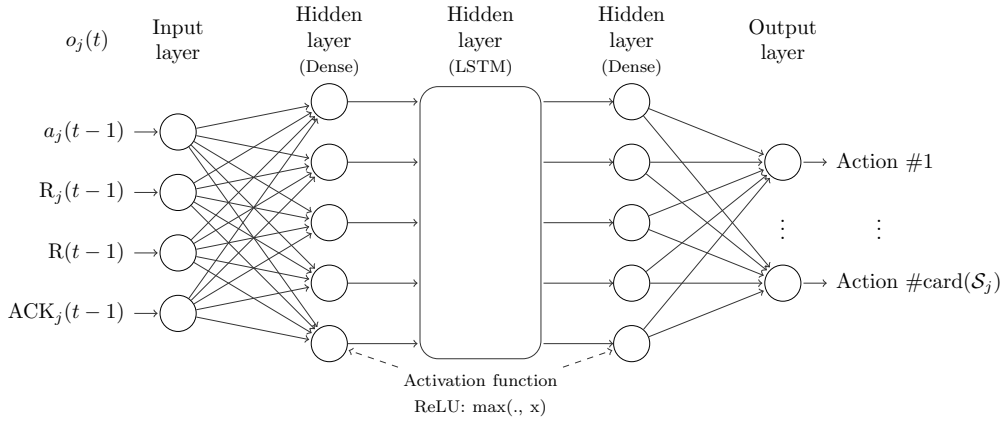

Fig. 3. Illustration of the architecture of the used DRQN.

($\beta \ll \alpha \leq 1$). When the TD error $\delta \geq 0$, the learning rate α is used; otherwise, β is considered. This leads to an optimistic update by according more importance to positive experiences [14]. Accordingly, the weights w_j^b are defined as follow:

$$w_j^b = \begin{cases} \alpha, & \text{if } \delta_j^b(t) \geq 0 \\ \beta, & \text{otherwise} \end{cases}. \tag{8}$$

Finally, it is worth to highlight that the target network weights $\hat{\theta}_j$ are updated less frequently to improve the learning stability [16].

C. Definition of the reward function

We treat the optimization problem as a continuing task with a time horizon T_e. Indeed, there is no explicit or predefined terminal state as the maximum network throughput is unknown at the beginning. In contrast, our algorithm must ensure that the system keeps following the optimal policy over the time horizon $T_e < \infty$ once the maximum throughput is reached. Moreover, since all UEs are requesting connections simultaneously, *collisions* may occur when the number of UEs asking a given SBS i for a connection is greater than the number of available beam N_i. However, in our proposed framework, we want to avoid the collision events, and serve as many UEs as possible through the mmWave links, to increase the network sum-rate. Consequently, when a collision happens, we set the reward for all UEs to zero. As a result, we define the reward function of UE j in (7) as:

$$r_j(t) = \begin{cases} R(t), & \text{if there is no collision} \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

Each UE j then acts following its strategy π_j to maximize the accumulated discounted reward:

$$G_j = \sum_{t=1}^{T_e} \gamma^{t-1} r_j(t), \tag{10}$$

where the discounting factor γ is such that $0 \leq \gamma < 1$. Taking $\gamma = 0$ leads to myopic (instantaneous) network throughput maximization. In case of dynamic scenarios, it is better to consider $\gamma \neq 0$ in order to take into account the dynamic nature of the environment: there is no need to change the current user association at time step t due to a low reward perceived because of the environment dynamics if at the next time step the system will recover its *equilibrium*.

As defined, the reward perceived by the agents is continuous and varies with the environment stochasticity viz. fading, shadowing, interference, and noise. Accordingly, this reward setting can lead to many optimal or quasi-optimal *equilibria*, which is a major issue as it results in agents laboriously trying to converge [13].

TABLE I
SIMULATION PARAMETERS.

	Macro cell [17]	**Small cell**
Parameters	Values	
Carrier frequency, f_c	2.0 GHz	28 GHz
Bandwidth, B	10 MHz	500 MHz
Thermal Noise, N_0	-174 dBm/Hz	-174 dBm/Hz
Noise figure	5 dB	
Shadowing, X	9 dB	12 dB
Transmit power	46 dBm	20 dBm
g_0 (TX/RX)	17 dBi / 0 dBi	0 dBi, Eq. (3)
Cell radius, r		50 m
Beam width, θ	360°	20°
Side lobe gain, ξ		0.01
Inter-cell distance		$1.2 \times r$
Pathloss model	$128.1 + 36.7\log_{10}(d)$	Eq. (2), $d_0 = 5$ m

IV. SIMULATION RESULTS

Here we assess the performance of our proposed HDRQN-based user association in a static scenario (i.e., deterministic environment) via Monte-Carlo simulations. UEs are static and fast fading is not considered in this study. The case of dynamic scenarios (i.e., stochastic environment) is left for future works. Table I summarizes the network parameters used during simulations.

In all tests, three small cells are deployed inside the macro cell. UE and small cell locations follow the 3GPP

recommendations [17]. To learn the user association policy, we use the DRQN described in Fig. 3. This architecture includes 2 multi-layers perceptron (MLP) of 32 hidden units, one LSTM with 64 memory cells followed by another 2 MLP of 32 hidden units. All layers use a rectifier linear unit (ReLU) except the final layer, which has a linear activation function. All simulation results are plotted for $\gamma = 0.9$. For hysteretic learning, we use a base learning rate $\mu = 0.001$ and scale the learning rate into $\alpha = \hat{\alpha}\mu$ and $\beta = \hat{\beta}\mu$ with $\hat{\alpha} = 1$ and $\hat{\beta} \in [0, 1]$. The DRQNs are trained offline over a time horizon $T_e = 5000$ using an ϵ-greedy policy with ϵ annealing from 1 to 0.01. Each time step, each agent uniformly samples a mini batch of size 32 from its CERT memory of size 500. The target network weights are updated every 20 time steps. Note that all hyperparameters values are selected via informal search. At the end of the learning phase, we collect the agent strategies and we compare the performance of our approach with 2 centralized frameworks, which serve as baselines:

1) SNR maximization-based beam allocation (*max-SNR*), which simply associates UEs with the BS providing the maximum SNR.
2) The *heuristic* algorithm proposed in [7] without power and beam width constraints.

The max-SNR method does not take the interference into consideration and accordingly has limited performance especially in case of dense networks. In contrast, the heuristic method iteratively looks for an optimal association taking into account the inter-beam interference; however, at each iteration, the association is based on the SNR, which may prevent from reaching a global optimum.

A. Complexity analysis

A naive algorithm may find the optimal solution of problem (4a)-(4d) through an exhaustive search. For UE j, there are $\text{card}(\mathcal{S}_j)$ possible choices of BSs. For all UEs, there are $\prod_{j\in\mathcal{U}} \text{card}(\mathcal{S}_j)$ possible combinations in which only some of them satisfy the constraint (4c). For each combination, checking if constraint (4c) is satisfied requires $O(\sum_{i\in\mathcal{S}} \text{card}(\mathcal{U}_i))$ iterations. In the worst case, i.e. $\text{card}(\mathcal{S}_j) = N_s+1$, running this naive algorithm will therefore require $O((\sum_{i\in\mathcal{S}} \text{card}(\mathcal{U}_i)) \prod_{j\in\mathcal{U}} \text{card}(\mathcal{S}_j) = O(K(1+N_s)^K)$ iterations. On the one hand, the complexity of both max-SNR and heuristic algorithms during execution is related to sorting the SNR values; however, the need to collect such values is the most notable disadvantage of these centralized approaches. Considering a *quicksort* algorithm, this complexity in the worst case ($\text{card}(\mathcal{S}_j) = N_s + 1$) is around $O(n\log(n))$ for max-SNR and $O(n + n\log(n))$ for the heuristic algorithm[2] where $n = K(1 + N_s)$. On the other hand, the complexity of our algorithm depends on the local DRQNs. Let L_h be the size of hidden layers and L_c the number of cells in LSTM layer. The complexity of UE j DRQN is in the order of $O(4L_h + 2L_h^2 + L_hL_c + 2L_h^2 + L_h(\text{card}(\mathcal{S}_j))) \approx O(4L_h^2 + L_hL_c)$. In terms of signaling overhead, as the reward is the same for all UEs, the complexity of our approach at each iteration is equal to $O(KB)$, where B is the number of bits exchanged in the message sequence chart of Fig. 2. This is lower than the signaling overhead of the algorithm proposed in [4].

[2]One pass to sort the SNR values and another to find the association.

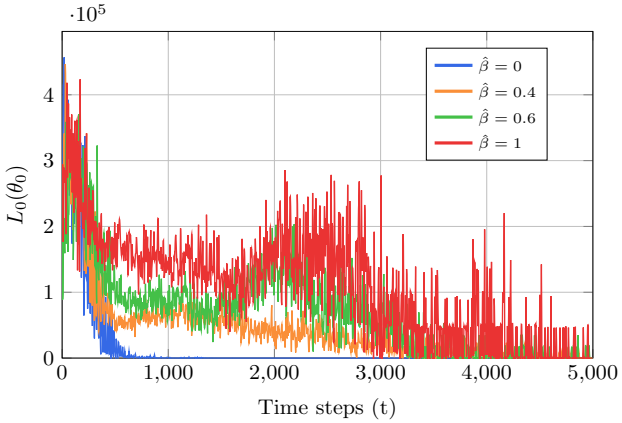

Fig. 4. Loss function for different values of $\hat{\beta}$, $N_s = 3$, $K = 9$.

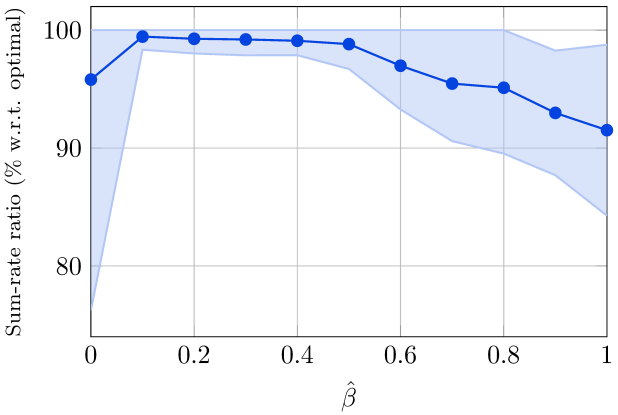

Fig. 5. Sum-rate ratio for different values of $\hat{\beta}$, $N_s = 3$, $K = 9$, averaged over 50 runs.

B. Effect of the hysteretic parameter

Despite the few information available locally to UEs, they successfully learn the BS to which they should request a connection. Fig. 4 shows the evolution of the loss function in (6) with respect to the hysteretic parameter $\hat{\beta}$. Lowering $\hat{\beta}$ helps increase the convergence speed but results in agents being more and more optimistic. In fact, when $\hat{\beta}$ decreases, agents tend to give less and less importance to actions producing negative TD errors. Consequently, they may fail to (autonomously) coordinate, which prevents from reaching the global optimum. Fig. 5 shows the sum-rate ratio with respect to the optimal user association solution for different values of $\hat{\beta}$. According to these results, taking $\hat{\beta} \in [0.1, 0.5]$ is a good compromise between convergence speed and network sum-rate.

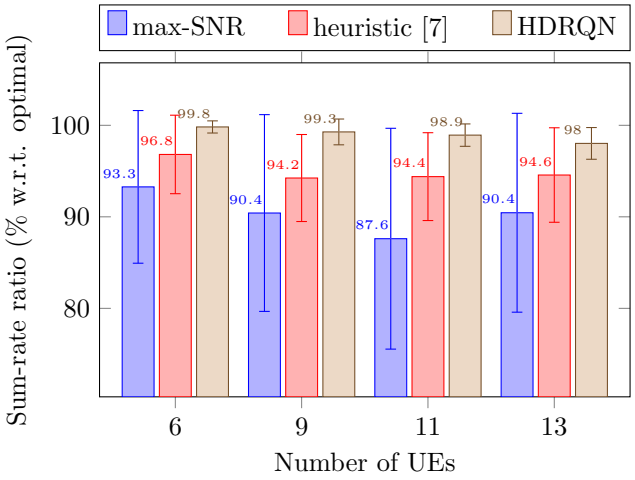

Fig. 6. Performance comparison, $\hat{\beta} = 0.4$, $N_1 = 2$, $N_2 = N_3 = 3$, averaged over 50 runs.

C. Performance comparison

We now run the algorithm on different network configurations where we change the number of active UEs[3]. Results are averaged over 50 runs. At each run, UE positions are randomly reset. Fig. 6 shows that the HDRQN-based association can achieve up to 98% of the optimal performance obtained via exhaustive search and outperforms the 2 baselines. In fact, on average, our proposed solution has a 4% and 9% performance gain over the heuristic and max-SNR algorithms, respectively. When the number of UEs increases, ensuring coordination between agents becomes more and more complex, resulting in slightly decreasing performance with respect to the global optimum. For example, when the number of UEs $K = 6$, our proposed algorithm achieves 99.8% of the optimal performance, while when $K = 13$, only 98% is reached. In addition, it is worth to highlight the very low variance of the sum-rate ratio of the proposed scheme shown in Fig. 6 (the standard deviation in average is only 1.25% of the optimal performance for the proposed scheme, compared to 4.75% and 8.25% for the heuristic and max-SNR algorithms, respectively), which demonstrates its robustness and stability when compared to the two baselines.

V. Conclusion

In this work we present a novel and flexible approach to handle the user association in dense networks with multiple radio access technologies. We first formulate the user association as a sum-rate maximization problem; then, we cast it as a multi-agent reinforcement learning task where agent decisions are based on partial and local observations. Eventually, we define a DRQN architecture and the associated signaling protocol to enable UEs to learn the association policy in a distributed way, with limited complexity. Our simulation results show the feasibility of the proposed algorithm and that it can achieve up to 98% of the performance provided by the optimal solution.

Future work will focus on dynamic scenarios, taking into account the time varying nature of communication channels and considering more realistic antenna models.

[3]The limitation to $K = 13$ is due to the computational complexity to find the optimal solution.

References

[1] T. Bai, R. Vaze, and R. W. Heath, "Analysis of Blockage Effects on Urban Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 5070–5083, Sep. 2014.

[2] S. Singh, R. Mudumbai, and U. Madhow, "Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control," *IEEE/ACM Transactions on Networking*, vol. 19, pp. 1513–1527, Oct 2011.

[3] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 23, pp. 836–850, June 2015.

[4] Y. Liu, X. Fang, M. Xiao, and S. Mumtaz, "Decentralized Beam Pair Selection in Multi-Beam Millimeter-Wave Networks," *IEEE Transactions on Communications*, vol. 66, pp. 2722–2737, June 2018.

[5] Q. Mao, F. Hu, and Q. Hao, "Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 20, pp. 2595–2621, Fourthquarter 2018.

[6] L. Busoniu, R. Babuŝka, and B. D. Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, pp. 156–172, March 2008.

[7] 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 Transactions on Vehicular Technology*, vol. 68, pp. 592–603, Jan 2019.

[8] N. Zhao, Y. Liang, D. Niyato, Y. Pei, and Y. Jiang, "Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2018.

[9] G. Ghatak, A. De Domenico, and M. Coupechoux, "Coverage Analysis and Load Balancing in HetNets With Millimeter Wave Multi-RAT Small Cells," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 3154–3169, May 2018.

[10] M. Gapeyenko, V. Petrov, D. Moltchanov, M. R. Akdeniz, S. Andreev, N. Himayat, and Y. Koucheryavy, "On the Degree of Multi-Connectivity in 5G Millimeter-Wave Cellular Urban Deployments," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 1973–1978, Feb 2019.

[11] T. Bai and R. W. Heath, "Coverage and Rate Analysis for Millimeter-wave Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2015.

[12] O. Naparstek and K. Cohen, "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 18, pp. 310–323, Jan 2019.

[13] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Reward Function and Initial values: Better Choices for Accelerated Goal-directed Reinforcement Learning," in *International Conference on Artificial Neural Networks*, pp. 840–849, Springer, 2006.

[14] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic q-learning: an Algorithm for Decentralized Reinforcement Learning in Cooperative Multi-agent Teams," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 64–69, IEEE, 2007.

[15] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 2681–2690, PMLR, 06–11 Aug 2017.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[17] 3GPP TR 36.872 V12.1.0, "Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects (Release 12), ," Dec 2013.