A Novel Multi-Agent Deep Reinforcement Learning-enabled Distributed Power Allocation Scheme for mmWave Cellular Networks

Xiang Zhang*, Arupjyoti Bhuyan‡, Sneha Kumar Kasera†* and Mingyue Ji*†
Department of Electrical and Computer Engineering, University of Utah*
Idaho National Laboratory‡
School of Computing, University of Utah†
Email: *{xiang.zhang, mingyue.ji}@utah.edu, ‡arupjyoti.bhuyan@inl.gov, †kasera@cs.utah.edu

2023 IEEE International Conference on Communications Workshops (ICC Workshops) | 979-8-3503-3307-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCWORKSHOPS57953.2023.10283502

Abstract—**We consider the power allocation problem over shared spectrum for millimeter-Wave (mmWave) cellular downlink. Existing approaches usually find sub-optimal solutions by solving a non-convex optimization which leads to scalability issues due to centralized control. Therefore, distributed and adaptive approaches are desirable. Recently, model-free Deep Reinforcement Learning (DRL) has achieved success in such wireless resource management tasks. By modeling the radio environment as a Markov Decision Process (MDP) with the base stations (BSs) being the agents, power allocation can be automated at the agent level with comparable throughput performance to conventional centralized schemes. The multi-agent setting presents new challenges as the radio environment is impacted by the joint actions of the agents and is no longer stationary from any individual agent's perspective. Existing literature bypasses this non-stationarity violation by ignoring it which may cause performance degradation. To tackle this issue, we propose a distributed continuous power allocation scheme based on a modified version of multi-agent Deep Deterministic Policy Gradient (MADDPG) that is tailored for the distributed multiple-agent setting. The proposed scheme employs a centralized-training-distributed-execution framework where Q-functions are trained over subsets of BSs while each BS determines its transmit power based only on its own local observation. It admits constant per-BS communication and computation complexity and is thus scalable to large networks. Numerical evaluation shows that the proposed scheme adapts well to a wide range of interference conditions and can achieve comparable or better performance than several state-of-the-art non-learning approaches.**

I. Introduction

Millimeter-Wave (mmWave) communication is one of the key enabling technologies for the fifth-generation (5G) cellular systems. The proliferation of mmWave frequency bands has increased the link capacity by several orders of magnitude compared to sub-6 GHz wireless systems and is able to support massive connections [1]. To combat propagation loss, directional beamforming is commonly used [2]. It was shown [3] that even at mmWave frequencies, spectrum availability is still limited considering the abundance of mobile and data-intensive services. Therefore, spectrum sharing is necessary for better utilization of unlicensed and shared spectrum. However, the concurrency of highly directional transmissions presents new challenges to spectrum sharing. Without proper coordination, beams could overlap and cause severe interference which hinders the performance. The situation is further exacerbated by the use of small cells with densely populated user equipment (UE).

Recently, deep reinforcement learning (DRL) has achieved notable success in wireless resource management [4]–[17]. Nasir and Guo [4] proposed a deep Q-network (DQN)-based (discrete) power allocation scheme which achieves competitive throughput performance to conventional centralized approaches like weighted minimum mean square error (WMMSE) [18] and fractional programming (FP) [19]. Treating as RL agents, the transmitters improve their decision making by actively interacting with the radio environment and benefit from learning with accumulated experiences. This work was further extended to continuous power control [5] and joint spectrum and power allocation [6]. Some other types of DRL algorithms were applied to the same tasks [7], [9]–[11]. For mmWave networks, Feng *et al.* [12] proposed a DQN-based resource management scheme to learn and predict blockage patterns in a backhaul capacity-limited system. A DQN-based joint spectrum and (discrete) power allocation scheme was proposed in [13]. Elsayed *et al.* [14] studied the clustering problem for mmWave networks with user mobility and proposed a DQN-based clustering scheme. Moreover, Sana *et al.* [15] proposed a deep recurrent Q-network (DRQN)-based handover scheme for dynamic mmWave user association. One common issue with these works is that, *the stationarity assumption of MDP is violated* in the multi-agent setting as the environment seen by each agent is impacted by the unknown behaviors of other agents. This violation is usually ignored in the existing literature.

In this paper, we design power allocation schemes for mmWave cellular downlink by leveraging the multi-agent Deep Deterministic Policy Gradient (MADDPG) algorithm [20]. Each base station (BS) is modeled as an agent that determines its transmit power autonomously in real time. MADDPG addresses the multi-agent environment non-stationarity issue by conditioning the Q-function of individual agents also on other agents' actions which are made available by using a centralized-training-distributed-execution framework. However, conditioning the Q-functions over the

979-8-3503-3307-7/23/$31.00 ©2023 IEEE

actions of a large number of agents necessitates high inter-BS communication overhead and may incur instability in agent training. To make it scalable, we propose a *distributed version* of MADDPG where the Q-function (critic) of each agent is trained over a subset of BSs (i.e., each BS and its neighbors) with a system-level reward. This formulation suppresses the unnecessary information exchange among BSs which barely impact each other's local environment dynamics. It also increases the stability and training efficiency of the deep neural network (DNN)-based actor/critic training by restricting the input of the DNNs to a relatively small size. We then propose a distributed power allocation scheme based on the proposed distributed MADDPG algorithm. Simulations shows that the proposed scheme can achieve comparable or better performance than WMMSE [18] and FP [19].

Several works [4], [5], [7], [13] are most related to ours. We explain the difference as follows. First, our scheme deals with continuous power control while the DQN-based schemes [4], [7], [13] can only handle discrete powers and the effect of quantization has not been properly investigated. Second, we address agent heterogeneity by equipping each BS with a unique actor/critic that accommodates its specific local radio environment. In contrast, in [4], [5], a single global actor is trained using experiences gathered from all BSs which is then copied to each BS for use. For heterogeneous systems like mmWave networks where each BS can face very different beam coverage and interference conditions, a single actor/critic may not be able to fit all agents. Third, the proposed distributed MADDPG only requires information exchange among subsets of BSs. This largely reduces the communication overhead compared to [4], [5] where network-level experience collection is required.

II. Problem Description

Consider a mmWave cellular network with K base stations (BSs) $\{1,\cdots,K\}$, each of which is associated with a number of user equipment (UE). The BSs are equipped with multiple antennas to enable beam-based directional transmissions while the UEs are equipped with a single antenna. The network is a fully synchronized and slotted system operating on a shared spectrum of W Hz. We adopt a block fading model for the downlink channels. In particular, the small-scale fading keeps unchanged during each slot and follows a temporally correlated Nakagami distribution [21] with probability density

$$f(x|m,\Omega)=\frac{2m^m}{\Gamma(m)\Omega^m}x^{2m-1}\exp\left(-\frac{m}{\Omega}x^2\right),\forall x\geq 0 \quad (1)$$

with parameters $\Omega=\mathbb{E}[X^2], m=\Omega^2/\mathrm{Var}(X^2)$ and Γ denotes the Gamma function. The fading coefficients $\{h^{(t)},\forall t\}$ are generated in a way such that 1) $h^{(t)},\forall t$ follows a Nakagami distribution with the same parameters m,Ω, and 2) the squared channels between any two consecutive time slots have a correlation coefficient of $\rho=\frac{\mathrm{Cov}(|h^{(t)}|^2,|h^{(t+1)}|^2)}{(\mathrm{Var}(|h^{(t)}|^2)\mathrm{Var}(|h^{(t+1)}|^2))^{1/2}},\forall t$. Let $h_{u_j i}^{(t)}$ denote the fading coefficient from BS i to UE u_j which is the scheduled UE of BS j. As we do not consider UE scheduling in this work, we write $h_{u_j i}^{(t)}$ as $h_{ji}^{(t)}$ for brevity. The equivalent channel gain $g_{ji}^{(t)}$ can then be written as $g_{ji}^{(t)}=\mathrm{PL}(d_{ji})G_{ji}|h_{ji}^{(t)}|^2$ in which $\mathrm{PL}(d_{ji})$ denotes the path loss and d_{ji} is the distance between BS i and UE u_j. G_{ji} denotes the antenna gain of BS i towards UE u_j. We use the dual-slope path loss model that is widely adopted for mmWave channel modeling [22]–[24]. In particular, the path loss at distance d is equal to

$$\mathrm{PL}(d)=\begin{cases}1/d^{\alpha_0}, & d\leq d_C\\ d_C^{\alpha_1-\alpha_0}/d^{\alpha_1}, & d>d_C\end{cases} \quad (2)$$

where d_C is the critical distance, and $\alpha_0,\alpha_1(\alpha_1\geq\alpha_0>0)$ are the near- and far-field path loss exponents. Moreover, the BSs use a keyhole-like sectorized antenna model [25] which has a constant mainlobe radiation gain of $G^{\max}$ and a constant sidelobe gain $G^{\min}$. In particular, let Δ be the beamwidth, then the antenna gain in the direction of θ is equal to $G^{\max}$ if $|\theta|\leq\Delta$ and $G^{\min}$ otherwise. The main-to-sidelobe ratio (MSR) is defined as $\mathrm{MSR}\triangleq 10\lg(G^{\max}/G^{\min})$ dB.

Let $\boldsymbol{p}^{(t)}\triangleq[p_1^{(t)},\cdots,p_K^{(t)}]$ denote the power allocation profile of all BSs at time t, then the received signal-to-interference-plus-noise ratio (SINR) at the scheduled UE of BS i is equal to

$$\mathrm{SINR}_i^{(t)}=\frac{g_{ii}^{(t)}p_i^{(t)}}{\sum_{j\neq i}g_{ij}^{(t)}p_j^{(t)}+\sigma^2} \quad (3)$$

where $\sigma^2=n_0W$ is the total noise with n_0 denoting the noise power spectrum density. The normalized throughput (bps/Hz) of BS i can be written as $C_i^{(t)}=\log_2\left(1+\mathrm{SINR}_i^{(t)}\right)$. The goal is to maximize the total throughput

$$C_{\mathrm{sum}}^{(t)}=\sum_i C_i^{(t)},\ \forall t \quad (4)$$

subject to instantaneous power constraints $p_i^{(t)}\leq p_i^{\max},\forall i$. In particular, a power allocation $\boldsymbol{p}^{(t)}$ has to be computed at the beginning of each slot such that $C_{\mathrm{sum}}^{(t)}$ can be maximized. Although this problem has been extensively studied, finding the optimal solution is still challenging due to its non-convex nature. For mmWave networks, power control becomes more complicated as the beam-based transmissions has to be properly coordinated to reduce interference. Unlike conventional approaches using centralized control, we aim to leverage data-driven learning approaches to design distributed and scalable power allocation schemes in which BSs choose powers based on their local measurements and limited information exchange with other BSs.

III. Proposed Approach

We present the proposed multi-agent DRL-based power allocation in this section. We first derive a *distributed* version of MADDPG based on which a power allocation scheme is then developed. An overview of DRL is provided before proceeding to the description of the main approach.

A. DRL Overview

In RL, an agent aims to optimize an expected return through repeated interactions with the environment over time. In each interaction, the agent receives a reward signal from the environment as an indicator of the quality of the action taken. The agent learns in a trial-and-error manner by gradually refining its decision making using the received reward signals. More specifically, in a discrete-time Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, R, T)$, given some state $s_t \in \mathcal{S}$ at time t, the agent takes an action $a_t \in \mathcal{A}$ with probability $\mu(a_t|s_t)$ according to a policy μ satisfying $\int_a \mu(a|s_t)da = 1$. Impacted by a_t, the environment transitions (governed by the transition function T) to a new state s_{t+1} and the agent receives a scalar reward $r_t = R(s_t, a_t, s_{t+1})$, which indicates how good the taken action a_t is. The set of transition quadruples $\{(s_t, a_t, r_t, s_{t+1}), \forall t\}$ is referred to as an *experience*. The return G_t is defined as the cumulative future rewards, i.e., $G_t \triangleq \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}$ with $\gamma \in (0, 1]$ being the discount factor that adjusts the relative importance of the near and far future rewards. The state-action value function (or Q-function) Q^μ under a specific policy μ is defined as the expected return starting from any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, i.e.,

$$Q^\mu(s, a) \triangleq \mathbb{E}\left[G_t | s_t = s, a_t = a\right], \tag{5}$$

where the expectation is taken over both the policy μ and the transition dynamics T. Model-free RL optimizes the agent's expected return without knowing or explicitly learning the transition dynamics and has seen significant developments in recent years due to the use of neural function approximators.

Deep Deterministic Policy Gradient (DDPG) [26] is a DRL algorithm which focuses on deterministic policies that map each state s to a specific action $a = \mu(s)$. It uses an actor-critic architecture in which two separate DNNs θ^μ and θ^Q are used to represent the policy $\mu(s|\theta^\mu)$ (called *actor*) and the Q-function $Q(s, a|\theta^Q)$ (called *critic*) respectively. Lowe *et al.* [20] extended DDPG to the multi-agent domain and proposed the multi-agent DDPG (MADDPG) algorithm. It addresses the non-stationarity issue due to multi-agent participation by conditioning the Q-function of each agent also on the actions of other agents (referred to as a *centralized critic*), i.e., agent i is defined with a Q-function $Q_i(s, a_1, \cdots, a_N)$ that has all agents' actions as input. The intuition behind this modification is that, given a set of fixed actions of all other agents, the environment perceived by each agent becomes stationary regardless of what policies are used by other agents [27].

Under the multi-agent setting, each agent i gets a local observation o_i instead of the true global state. Actions are chosen based on each agent's local observation, i.e., $a_i = \mu_i(o_i|\theta_i^\mu)$. The training of the centralized critic $Q_i(\boldsymbol{s}, \boldsymbol{a}|\theta_i^Q)$ requires the knowledge of the global state $\boldsymbol{s} \triangleq (o_i)_{i=1}^K$ and the joint actions $\boldsymbol{a} \triangleq (a_i)_{i=1}^K$. This is enabled by using a centralized-training-distributed-execution framework in which the actor/critics are trained periodically with network-level experiences while the actions are determined based solely on each agent's local observation. In particular, a fixed-size experience replay buffer $\mathcal{D}$ is used to store the past experiences collected from all agents in a first-in-first-out (FIFO) manner. Mini-batches of experiences are then sampled from $\mathcal{D}$ to train the actors and critics using SGD. More specifically, given a mini-batch $\mathcal{B} = \{(\boldsymbol{s}^j, \boldsymbol{a}^j, \boldsymbol{r}^j, \boldsymbol{s}'^j)\}_j$ with $\boldsymbol{r}^j = (r_i^j)_{i=1}^K$ being the rewards of all agents, the critic of agent i is trained by minimizing the loss

$$L(\theta_i^Q) = \frac{1}{|\mathcal{B}|} \sum_{j \in \{1, \cdots, |\mathcal{B}|\}} \Big(y_i^j - Q_i\big(\boldsymbol{s}^j, (a_k^j)_{k=1}^K | \theta_i^Q\big)\Big)^2. \tag{6}$$

The regression target (over the j^{th} sample in $\mathcal{B}$) $y_i^j = r_i^j + Q_i'(\boldsymbol{s}'^j, (a_k')_{k=1}^K | \theta_i^{Q'})\big|_{a_k' = \mu_k'(o_k'^j | \theta_k^{\mu'}), \forall k}$ is generated by two *target networks* $Q_i'(\cdot|\theta_i^{Q'})$ and $\mu_i'(\cdot|\theta_i^{\mu'})$ in order to stabilize the training process. The actor of agent i is trained by minimizing

$$L(\theta_i^\mu) = -\frac{1}{|\mathcal{B}|} \sum_{j \in \{1, \cdots, |\mathcal{B}|\}} Q_i\big(\boldsymbol{s}^j, (a_k^j)_{k \neq i}, a_i | \theta_i^Q\big)\big|_{a_i = \mu_i(o_i^j | \theta_i^\mu)}. \tag{7}$$

Finally, the weights of the target networks are updated according to $\theta_i^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta_i^{Q'}, \theta_i^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta_i^{\mu'}, \forall i$ for some small number $\tau \in (0, 1)$. In MADDPG, exploration is achieved by adding a random noise N_t to the actor output, i.e., $a_i^{(t)} = \mu_i(o_i^{(t)} | \theta_i^\mu) + N_t$.

B. Proposed Power Allocation Scheme

We adopt a distributed version of MADDPG as the framework for the proposed power allocation scheme. In particular, each BS is treated as a DRL agent equipped with an actor network for determining transmit powers and a critic network for evaluating the Q-functions. It is assumed that BSs do not know each other's transmit powers, but they can obtain interference measurements through feedback from the scheduled UEs as well as some additional measurements through information exchange with neighboring BSs.

One major drawback of the original MADDPG [20] algorithm is that the Q-function $Q_i(\boldsymbol{s}, \boldsymbol{a})$ of each agent takes the global state $\boldsymbol{s} = (o_i)_{i=1}^K$ and the joint action $\boldsymbol{a} = (a_i)_{i=1}^K$ as input. This can be problematic for systems with a large number of agents as the input dimension of the critic network can be huge which may result in slow convergence and instability in DNN training. To address this issue, we modify MADDPG to allow *decentralized critic training* which means the critics of agents are trained over subsets of BSs (see Fig. 1). In particular, we restrict the training of each agent i to within its neighbor set $\mathcal{N}_i \subseteq \{1, \cdots, K\}$, which is defined as a set of agents whose actions have significant impact on agent i's local environment. Agent i instead learns a localized critic

$$Q_i\Big((o_j)_{j \in \widehat{\mathcal{N}}_i}, (a_j)_{j \in \widehat{\mathcal{N}}_i} | \theta_i^Q\Big), \tag{8}$$

where $\widehat{\mathcal{N}}_i \triangleq \mathcal{N}_i \cup \{i\}$. Q_i and μ_i are then trained similarly according to (6) and (7). (8) means that each agent only needs to gather information from its neighbors and network-level information exchange (which introduces delays) can be

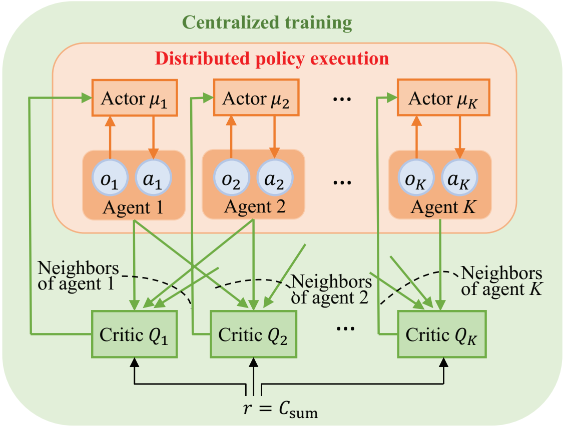


Fig. 1: Proposed distributed MADDPG framework. Each agent's critic is trained on its neighbor set.

reduced.

The definition of localized critic (8) is motivated by the following observation. In mmWave networks, BSs that are far apart from each other or with properly placed beams will contribute little interference to each other and have negligible impact on the local environment transition dynamics of each other. This is because mmWave frequency itself suffers from rapid attenuation due to propagation characteristics, and highly directional beams suppress interference to undesired directions. Therefore, it is unnecessary to use a centralized critic that needs to be trained over the entire set of BSs. More generally, our formulation provides a flexible trade-off between inter-BS information exchange and how accurately the overall radio environment can be perceived by each agent. Two special cases can be considered. When no information exchange is allowed among agents, i.e., $\mathcal{N}_i = \emptyset, \forall i$, (8) becomes $Q_i(o_i, a_i|\theta_i^Q)$, which is the case of independent learning [28]. In contrast, if information exchange is allowed among arbitrary agents, i.e., $\mathcal{N}_i = \{1, \cdots, K\} \backslash \{i\}, \forall i$, then (8) becomes $Q_i\big((o_j)_{j=1}^K, (a_j)_{j=1}^K | \theta_i^Q\big)$, which is equivalent to the original MADDPG critic. We describe how to design the neighbor set in Section IV-A. In addition, a separate replay buffer $\mathcal{D}_i$ can be defined for each agent i which stores the experiences in the form of

$$\big((o_j)_{j\in\widehat{\mathcal{N}}_i}, (a_j)_{j\in\widehat{\mathcal{N}}_i}, (r_j)_{j\in\widehat{\mathcal{N}}_i}, (o'_j)_{j\in\widehat{\mathcal{N}}_i}\big), \ \forall i \qquad (9)$$

We next describe the design of actions, local observations and rewards in the proposed scheme.

Action. Each BS needs to determine the transmit power $p_i^{(t)} \in [0, p_i^{\max}]$ to its scheduled UE at the beginning of each slot. Since the Tanh activation is used for the output layer of the actor networks, the actor output $a_i^{(t)} = \mu_i(o_i^{(t)}|\theta_i^\mu)$ falls into $[-1, 1]$. To achieve exploration, a random noise N_t is added to $a_i^{(t)}$, which is then clipped to within the range $[-w, w]$ for some $w \in (0, 1)$[1]. Therefore, the actor output is mapped to the powers according to $p_i^{(t)} = \frac{a_i^{(t)}+w}{2w} p_i^{\max}$.

Local Observation. The local observation represents each agent's (partial) perception of the radio environment. It should capture the environment features that are relevant to the agent's decision making. To control complexity and make the proposed scheme scalable, we limit the information exchange to neighboring BSs. In particular, the observation $o_i^{(t)}$ of agent i is defined as

$$o_i^{(t)} \triangleq \Big\{ p_i^{(t-1)}, g_{ii}^{(t-1)}, g_{ii}^{(t)}, I_i^{(t-1)}, \widehat{I}_i^{(t)}, C_i^{(t-1)}, \frac{C_i^{(t-1)}}{\sum_{j\in\widehat{\mathcal{N}}_i} C_j^{(t-1)}}, \\ g_{ij}^{(t-1)} p_j^{(t-1)}, g_{ij}^{(t)} p_j^{(t-1)}, C_j^{(t-1)}, \forall j \in \mathcal{N}_i \Big\}. \qquad (10)$$

which includes several interference and channel measurements at BS i and throughput obtained from neighboring BSs. In particular, $p_i^{(t-1)}$ is the power of BS i in the previous slot, $g_{ii}^{(t-1)} = \text{PL}(d_{ii}) G_{ii} |h_{ii}^{(t-1)}|^2$ is the direct channel between BS i and its scheduled UE in slot $t-1$[2], $g_{ii}^{(t)} = \text{PL}(d_{ii}) G_{ii} |h_{ii}^{(t)}|^2$ is the direct channel in slot t. It is assumed that the channel changes from $h_{ii}^{(t-1)}$ to $h_{ii}^{(t)}$ at the very beginning of slot t right before the new powers $p_i^{(t)}, \forall i$ are determined, $g_{ii}^{(t)}$ can also be used by BS i. $I_i^{(t-1)} = \sum_{j\neq i} g_{ij}^{(t-1)} p_j^{(t-1)} + \sigma^2$ is the total interference (plus noise) measured at BS i in slot $t-1$, $\widehat{I}_i^{(t)} = \sum_{j\neq i} g_{ij}^{(t)} p_j^{(t-1)} + \sigma^2$ is the total interference measured at the beginning of slot t where the channels have changed but the powers have not been updated. $C_i^{(t-1)}$ is the throughput of BS i at slot $t-1$, and $C_i^{(t-1)} / \sum_{j\in\widehat{\mathcal{N}}_i} C_j^{(t-1)}$ represents the relative importance of BS i in terms of throughput contribution among its neighbors. Moreover, $g_{ij}^{(t-1)} p_j^{(t-1)}$ and $g_{ij}^{(t)} p_j^{(t-1)}$ are the measured interference from BS $j \in \mathcal{N}_i$ in slot $t-1$ and the beginning of slot t respectively. Finally, $C_j^{(t-1)}$ represents the throughput achieved by BS j in the previous slot. Note that $C_j^{(t-1)}$ has to be delivered to BS i (from BS j) despite all other interference measurements can be directly obtained at BS i. We include one previous slot in order for the agents to better keep track of the time-varying channels.

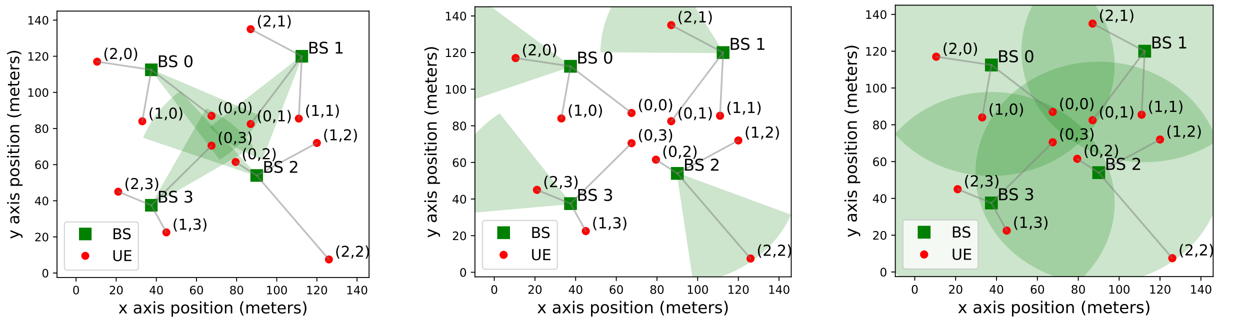
Reward. Unlike [4], [5] which rely on a heuristic reward design, we use a centralized reward, i.e., $r_i^{(t)} = \sum_{j=1}^K C_j^{(t)}, \forall i$ which is intuitive and computationally efficient.

With the above definitions, we summarize the proposed power allocation scheme as follows. At the beginning of slot t, each BS i conducts the interference measurements and exchanges throughput of the previous slot with its neighbors in order to construct its local observation $o_i^{(t)}$. BS i then chooses $a_i^{(t)} = \mu_i(o_i^{(t)}|\theta_i^\mu) + N_t$ which is then mapped to the actual power $p_i^{(t)} = \frac{a_i^{(t)}+\omega}{2\omega} p_{\max}$. The chosen power is used for one slot until the next slot begins and the new observation $o_i^{(t+1)}$ can be obtained. The experience of each agent i $\big((o_j)_{j\in\widehat{\mathcal{N}}_i}, (a_j)_{j\in\widehat{\mathcal{N}}_i}, (r_j)_{j\in\widehat{\mathcal{N}}_i}, (o'_j)_{j\in\widehat{\mathcal{N}}_i}\big)$ is then pushed to the corresponding replay buffer $\mathcal{D}_i$. For every T_{train} slots, the actors and critics are trained using mini-batch SGD according to (6) and (7).

Complexity. The input sizes of the actor and critic networks

[1] Since hyperbolic tangent function requires an infinitely large input to achieve the the output values ±1, clipping to $[-w, w]$ where $w < 1$ increases numerical stability of DNN training.

[2] The direct channels can be estimated via pilot training.



(a) Config. 1: overlapped beams. (b) Config. 2: wide (weak) beams. (c) Config. 3: omnidirectional (no beams).

Fig. 2: Network with various beam configurations. Gray solid lines indicate BS-UE association.

are $|o_i^{(t)}| = 3N + 7$ and $(N+1)|o_i^{(t)}| = (N+1)(3N+7)$ respectively (See (8)) where $N = |\mathcal{N}_i|$ is the neighbor size. It can be seen that these input sizes do not scale with K if N is fixed. For our implementation in Section IV with $N = 6$, each actor/critic network contains ~30k/55k parameters. Due to the local observation design of (10), each BS needs to send its throughput in the previous slot to its neighbors. This incurs a communication overhead $\mathcal{O}(N)$ per BS.

IV. Numerical Evaluation

A. Simulation Setup

Consider a network with 4 BSs under various beam and UE configurations as shown in Fig. 2. Each BS is associated with three UEs where UE (j, i) denotes the j^{th} UE of BS i. Antenna beams are aligned with the scheduled UEs. The purpose of this simulation is to verify the performance of the proposed scheme under controlled interference situations which will be elaborated in Section IV-B. The proposed scheme is then evaluated on a more general network topology (Fig. 3) to demonstrate its scalability.

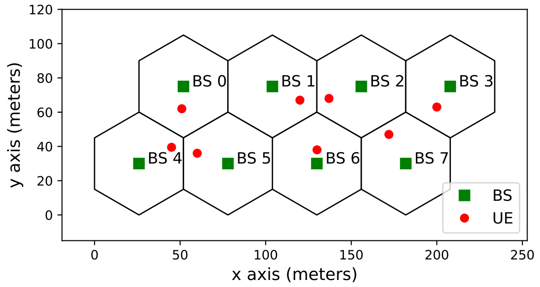


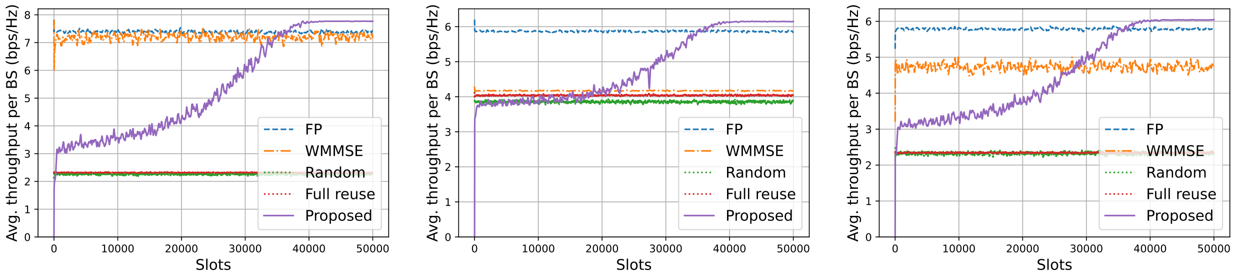
Fig. 3: Network with 8 BSs.

The parameters used in the simulation are summarized in Table I. The total noise power is calculated according to $\sigma^2\,(\text{dBm}) = 10\lg(\kappa_B T_0 \times 10^3) + \text{NR}\,(\text{dB}) + 10\lg W$ with κ_B, NR and T_0 being Boltzmann's constant, receiver noise figure and temperature respectively. Taking the typical values of $\text{NR} = 1.5$ dB, $T_0 = 290$ K, we have $\sigma^2 = -86.46$ dBm. The proposed scheme is implemented with PyTorch. Each actor/critic is represented by a fully-connected DNN with 5 layers including 3 hidden layers each containing $200, 100$ and 50 neurons respectively. Each actor network has one output port with Tanh activation clipped to the range $[-0.95, +0.95]$. Each critic network also has one output port activated with ReLU. We restrict the neighbor set of each BS to be $|\mathcal{N}_i| \leq 6, \forall i$ as in a hexagon cell grid, each BS can have at most six other neighboring BSs. If the number of neighbors for a BS is less than six, $6 - |\mathcal{N}_i|$ dummy agents are created to serve as virtual neighbors (with zero transmit power) to ensure a fixed input size of the DNNs. To ensure adequate exploration, the action noise $\{N_t, \forall t\}$ is chosen as i.d.d. Gaussian noise with a decreasing variance, i.e., $N_t \sim \mathcal{N}(0, \sigma_t^2)$ where $\sigma_{t+1} = \max\{(1-\varepsilon)\sigma_t, \sigma_{\min}\}$ with $\varepsilon = 10^{-4}$, $\sigma_0 = 1$ and $\sigma_{\min} = 0.01$. The actor/critics are trained every $T_{\text{train}} = 5$ slots. We have two phases, the training phase (50000 slots) where the actor/critics are trained periodically while interacting with the environment, and the testing phase (1000 slots) in which the trained actors are used to select powers without learning. We also found that normalizing the inputs of the actor/critic networks is crucial to stabilizing the training. Hence, each interference measurement I_i at BS i (see (10)) is scaled to $[0, 1]$ via $I_i' = \frac{I_i - I_i^{\min}}{I_i^{\max} - I_i^{\min}}$ where $I_i^{\max}$ and $I_i^{\min}$ are the maximum and minimum possible interference at BS i. We take $I_i^{\min} = 0$ and $I_i^{\max}$ can be estimated by letting all BSs transmit with maximum powers and observing the interference.

TABLE I: Simulation parameters.

Parameters	Value
$p_i^{\max}, \forall i$	39 dBm
Bandwidth W	400 MHz
Path loss	$\alpha_0 = 2, \alpha_1 = 4, d_c = 26$ m
Fading	$m = 50, \Omega = 1, \rho = 0.5$
DNN optimizer	`Adam`
Learning rates	10^{-4} (actor), 10^{-3} (critic)
Replay buffer size	5×10^5
Batch size	128
γ, τ	$0.9, 5 \times 10^{-3}$

The proposed scheme is compared to two state-of-the-art power allocation schemes WMMSE [18] and FP [19]. These are both centralized algorithms with superior performance



(a) Config. 1 performance.

(b) Config. 2 performance.

(c) Config. 3 performance.

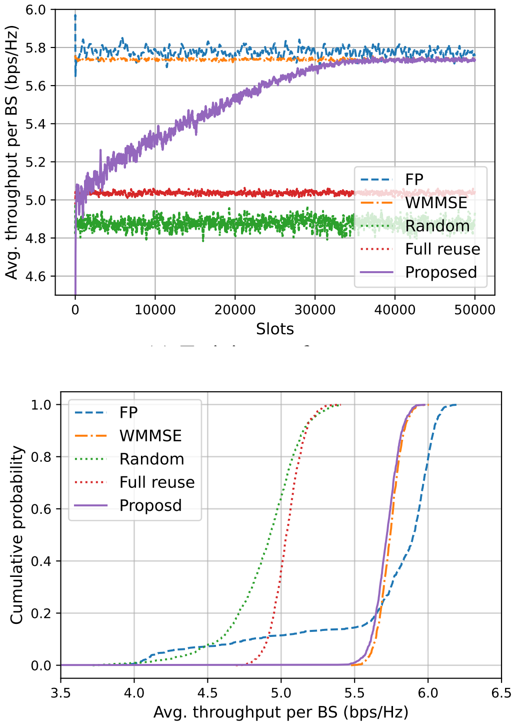
Fig. 4: Throughput performance during training under different configurations.

which is hard to beat in general. For example, several learning-based methods [4], [29] have been shown to approximate their performance. However, the performance of WMMSE/FP depends on heuristic parameter initialization whose impact is unclear. The closed-form FP algorithm [19, Alg. 3] is used in our simulation. For both schemes, we assume that the required CSI can be obtained with no delay at the beginning of each slot. We then run 2000 iterations to obtain a stationary power allocation. We also have two additional baselines which are full reuse (maximum power) and random power allocation.

B. Simulation Results

The simulation results are presented in this section. Each result is an average of three independent simulations run over the same channel realizations to neutralize the performance variance due to random explorations. Each data point on the throughput curve is an average of the previous 100 slots.

We first discuss the results (Fig. 4) corresponding to the smaller four-BS network with various beam configurations. In this case, each BS only has three neighbors, so local observations of all BSs are used to train the critics. Fig. 2a represents a typical scenario where the set of scheduled UEs $\{(0,i),\forall i\}$ are located at the cell edge and experience strong interference due to beam overlapping. The antenna MSR and beamwidth are set to be 10 dB and 30°. From Fig. 4a, it can be seen that the proposed scheme converges in roughly 40000 slots and achieves slightly higher throughput than FP and WMMSE. Fig. 2b represents the case of sparsely distributed wide beams with relatively low power concentration (3 dB-60° beams with scheduled UEs $\{(2,i),\forall i\}$). Fig. 4b shows that the proposed scheme achieves superior performance than the baselines, especially for WMMSE which only has slightly better performance than full reuse. Finally, the proposed scheme is evaluated on a configuration with no beams (BS antennas are omnidirectional, see Fig. 2c with scheduled UEs $\{(1,i),\forall i\}$. This configuration may not be practical for mmWave but it shows the extendability of our scheme to general wireless networks besides mmWave. The corresponding result shows that our scheme achieves higher throughput than FP and outperforms WMMSE by a significant margin. In summary, the above results reveals that our scheme is capable of adapting to different interference regimes as implied by the various beam/UE configurations.



(b) Empirical CDF of testing.

Fig. 5: Performance over the general topology.

Fig. 5 shows the result corresponding to the larger network of Fig. 3. The antenna MSR and beamwidth are configured as 20 dB and 45°. Each BS shares local observations with at most four of its neighbors. Fig. 5a depicts the throughput performance during training. The proposed scheme converges in roughly 32000 slots and achieves similar performance to WMMSE and is slightly lower than FP. This demonstrates the scalability of the proposed scheme. The learned policy is then tested for 1000 slots over another randomly generated channel realization with a smaller correlation of $\rho = 0.1$. Note that the exploration noise is removed during testing. The empirical cumulative density function (CDF) of the average throughput during testing is shown in Fig. 5b. Although the channels are less correlated and harder to be tracked, the learned policy still maintains a very close performance to WMMSE. FP achieves higher throughput but has a less concentrated distribution than the proposed scheme. This demonstrates the generalization ability of the proposed scheme.

V. Conclusion

In this work, we presented a novel distributed power allocation scheme for mmWave downlink using multi-agent deep reinforcement learning. The proposed scheme utilizes a centralized-training-distributed-execution framework to enable autonomous power allocation based on local measurements of the base stations. The non-stationarity issue due to multi-agent participation is addressed by conditioning the Q-function of each agent on the actions of its neighboring agents. Simulation showed that the proposed scheme achieves similar or better throughput performance than the state-of-the-art non-learning-based approaches including FP and WMMSE. It was also demonstrated that the proposed scheme adapts well to a wide range of beam and user configurations.

Acknowledgement

This work was supported through NSF grants NSF SWIFT 2229562, NSF U.S.-Ireland R&D Partnership 2153875 and INL Laboratory Directed Research & Development (LDRD) Program under DOE Idaho Operations Office Contract DE-AC07-05ID14517.

References

[1] A. Ghosh, “The 5G mmwave radio revolution,” *Microw. J*, vol. 59, no. 9, pp. 22–36, 2016.

[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, no. 5, pp. 1513–1527, 2011.

[3] F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor, E. Erkip, C. Fischione, M. Kountouris, P. Popovski, and M. Zorzi, “Spectrum pooling in mmwave networks: Opportunities, challenges, and enablers,” *IEEE Communications Magazine*, vol. 54, no. 11, pp. 33–39, 2016.

[4] Y. S. Nasir and D. Guo, “Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.

[5] ——, “Deep actor-critic learning for distributed power control in wireless mobile networks,” in *2020 54th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2020, pp. 398–402.

[6] ——, “Deep reinforcement learning for joint spectrum and power allocation in cellular networks,” in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[7] L. Zhang and Y.-C. Liang, “Deep reinforcement learning for multi-agent power control in heterogeneous networks,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2551–2564, 2020.

[8] F. Meng, P. Chen, L. Wu, and J. Cheng, “Power allocation in multi-user cellular networks: Deep reinforcement learning approaches,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6255–6267, 2020.

[9] Y. Chen, Y. Li, D. Xu, and L. Xiao, “DQN-based power control for iot transmission against jamming,” in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–5.

[10] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, “Resource management in wireless networks via multi-agent deep reinforcement learning,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3507–3523, 2021.

[11] H. Kabir, M.-L. Tham, and Y. C. Chang, “Twin delayed DDPG based dynamic power allocation for internet of robotic things,” in *2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2022, pp. 1–6.

[12] M. Feng and S. Mao, “Dealing with limited backhaul capacity in millimeter-wave systems: A deep reinforcement learning approach,” *IEEE Communications Magazine*, vol. 57, no. 3, pp. 50–55, 2019.

[13] J. Gao, C. Zhong, X. Chen, H. Lin, and Z. Zhang, “Deep reinforcement learning for joint beamwidth and power optimization in mmwave systems,” *IEEE Communications Letters*, vol. 24, no. 10, pp. 2201–2205, 2020.

[14] M. Elsayed and M. Erol-Kantarci, “Radio resource and beam management in 5G mmwave using clustering and deep reinforcement learning,” in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[15] M. Sana, A. De Domenico, E. C. Strinati, and A. Clemente, “Multi-agent deep reinforcement learning for distributed handover management in dense mmwave networks,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8976–8980.

[16] A. A. Khan and R. S. Adve, “Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8410–8426, 2020.

[17] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, “Edge learning for b5g networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9–39, 2023.

[18] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[19] K. Shen and W. Yu, “Fractional programming for communication systems-part I: Power control and beamforming,” *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[20] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.

[21] Z. Shi, S. Ma, G. Yang, K.-W. Tam, and M. Xia, “Asymptotic outage analysis of HARQ-IR over time-correlated nakagami-m fading channels,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6119–6134, 2017.

[22] X. Zhang and J. G. Andrews, “Downlink cellular network analysis with multi-slope path loss models,” *IEEE Transactions on Communications*, vol. 63, no. 5, pp. 1881–1894, 2015.

[23] Y. Chang, S. Baek, S. Hur, Y. Mok, and Y. Lee, “A novel dual-slope mm-wave channel model based on 3d ray-tracing in urban environments,” in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*. IEEE, 2014, pp. 222–226.

[24] S. Hur, S. Baek, B. Kim, Y. Chang, A. F. Molisch, T. S. Rappaport, K. Haneda, and J. Park, “Proposal on millimeter-wave channel modeling for 5G cellular system,” *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 454–469, 2016.

[25] M. Kiese, C. Hartmann, J. Lamberty, and R. Vilzmann, “On connectivity limits in ad hoc networks with beamforming antennas,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1–15, 2009.

[26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.

[27] R. E. Wang, M. Everett, and J. P. How, “R-MADDPG for partially observable environments and limited communication,” *arXiv preprint arXiv:2002.06684*, 2020.

[28] M. Tan, “Multi-agent reinforcement learning: Independent vs. cooperative agents,” in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.

[29] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, “Learning to optimize: Training deep neural networks for interference management,” *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438–5453, 2018.