INL/CON-23-71768-Revision-0

Distributed Power Allocation for 6-GHz Unlicensed Spectrum Sharing via Multi-agent Deep Reinforcement Learning

April 2023

Changing the World's Energy Future

Xiang Zhang, Sneha Kasera, Mingyue Ji, Arupjyoti Bhuyan

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/CON-23-71768-Revision-0

Distributed Power Allocation for 6-GHz Unlicensed Spectrum Sharing via Multi-agent Deep Reinforcement Learning

Xiang Zhang, Sneha Kasera, Mingyue Ji, Arupjyoti Bhuyan

April 2023

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Distributed Power Allocation for 6-GHz Unlicensed Spectrum Sharing via Multi-agent Deep Reinforcement Learning

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

Abstract—**We consider the problem of spectrum sharing by multiple cellular operators. We propose a novel deep Reinforcement Learning (DRL)-based distributed power allocation scheme which utilizes the multi-agent Deep Deterministic Policy Gradient (MA-DDPG) algorithm. In particular, we model the base stations (BSs) that belong to multiple operators sharing the same band, as DRL agents that simultaneously determine the transmit powers to their scheduled user equipment (UE) in a synchronized manner. The power decision of each BS is based on its own observation of the radio (RF) environment, which consists of interference measurements reported from the UEs it serves, and a limited amount of information obtained from other BSs. One advantage of the proposed scheme is that it addresses the single-agent non-stationarity problem of RL in the multi-agent scenario by incorporating the actions and observations of other BSs into each BS's own critic which helps it to gain a more accurate perception of the overall RF environment. A centralized-training-distributed-execution framework is used to train the policies where the critics are trained over the joint actions and observations of all BSs while the actor of each BS only takes the local observation as input in order to produce the transmit power. Simulation with the 6 GHz Unlicensed National Information Infrastructure (U-NII)-5 band shows that the proposed power allocation scheme can achieve better throughput performance than several state-of-the-art approaches.**

I. Introduction

Spectrum sharing [1] enables efficient utilization of additional unlicensed or shared spectrum by allowing multiple service operators to operate on the same frequency bands, which has been shown to significantly enhance the network-level throughput performance [2], [3]. For example, the Federal Communications Commission (FCC) has defined the Unlicensed National Information Infrastructure (U-NII) sub-6 GHz radio band for sharing with existing primary band allocations [4]. In spectrum sharing, the transmit powers of the base stations (BSs) must be properly controlled in order to achieve effective interference mitigation and throughput optimization. This is usually a challenging task due to the nonconvex nature of the optimization objective.

Deep reinforcement learning (DRL) has achieved notable success in wireless resource management related tasks in recent years [5]–[20]. For example, Nasir and Guo [5] proposed a novel distributed power allocation scheme based on deep Q-network (DQN) which can achieve competitive throughput performance comparing to state-of-the-art non-learning approaches. This reveals the potential of DRL-based approaches in wireless network optimization [21]. By modeling the power allocation task as a Markov Decision Process (MDP), the BSs are treated as RL agents which actively exploit the radio environment and learn through trial and error. This work was later extended to continuous power control [6], [7] using more advanced DRL algorithms that can handle continuous actions. In addition, Gao *et al.* [14] proposed a DQN-based joint spectrum and (discrete) power allocation scheme. The application of DRL algorithms have also been investigated in various other tasks [8], [10]–[12], [16], [18]–[20]. Feng *et al.* [13] proposed a wireless resource management scheme in which DQN is used to learn and predict the blockage patterns. Elsayed *et al.* [15] applied DQN to millimeter-Wave (mmWave) networks for efficient user clustering. Moreover, Sana *et al.* [16] studied the dynamic user association problem in mmWave networks and proposed a novel handover scheme by leveraging deep recurrent Q-network (DRQN). Xue *et al.* [18] proposed an autonomous mmWave beam control scheme based on the double DQN (DDQN) algorithm. More recently, DRL has been applied to intelligent reflecting surface (IRS)-aided wireless communication systems [19], [20], [22].

For power allocation over shared spectrum with multiple BSs, the issue of *environment nonstationarity* has not been properly addressed in the existing literature. More specifically, in multi-agent RL systems, the environment perceived by each agent is affected by the actions of other agents. For system scalability and privacy considerations, sharing of actions among agents is usually prohibited. As a result, each agent's perceived environment will no longer be stationary as it does know other agents' actions, which is a violation of the basic MDP assumptions. Most existing literature bypasses this problem by ignoring it. However, the impact on the algorithm performance remains unclear.

In this paper, we propose a novel distributed power allocation scheme by utilizing multi-agent Deep Deterministic Policy Gradient (MA-DDPG) [23]. In particular, each BS is treated as a DRL agent that needs to determine the transmit powers from slot to slot based on its own perception of the radio environment. MA-DDPG addresses the multi-agent nonstationarity issue by defining a new Q-function for each agent which also takes other agents' actions as input in addition to that agent's own action. The rationale behind this definition is that the environment seen by each agent will be stationary if all other agents' actions are fixed, regardless of their actual policies. In order to train the actor and critic networks of the BSs, we utilize the centralized-training-distributed-execution framework where a replay buffer is used to store experiences collected from all BSs and the actor/critic networks are trained using mini-batch SGD. Meanwhile, the transmit powers are determined in a distributed manner because each BS makes its power decision based solely on its local observation, which also leads to a scalable design. We compare the proposed scheme with two state-of-the-art non-learning approaches including WMMSE [24] and FP [25] both of which require full CSI across the network. Simulation result shows that the proposed scheme can achieve similar or even better performance than the baselines.

II. System Model & Problem Description

We consider the downlink power allocation problem for a network consisting of 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 enforce sector-based transmission and each UE is equipped with a single antenna. The system is slotted, fully synchronized and operates on a shared unlicensed spectrum around 6 GHz, i.e., the Unlicensed National Information Infrastructure (U-NII)-5 band with 500 MHz frequency range. In each slot, each BS schedules at most one of its associated UEs to transmit to and multiple BSs can transmit at the same time. The UEs considered for scheduling by each BS remain the same for the analysis presented in this paper.

The block fading model is used with unchanged channels within each slot and follows a temporal correlated Nakagami distribution [26] from slot to slot. More specifically, the small-scale fading coefficients $\{h^{(t)}, \forall t\}$ have two properties. First, all $h^{(t)}, \forall t$ are identically distributed following a Nakagami distribution with probability density

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

where $\Omega \triangleq \mathbb{E}[H^2], m \triangleq \frac{\Omega^2}{\text{Var}(H^2)}$ and Γ denotes the Gamma function. Second, the squared channels between any two consecutive time slots have a correlation coefficient of $\rho = \frac{\text{Cov}(|h^{(t)}|^2, |h^{(t+1)}|^2)}{\sqrt{\text{Var}(|h^{(t)}|^2)\text{Var}(|h^{(t+1)}|^2)}}, \forall t$. Let $h_{ji}^{(t)}$ denote the channel coefficient from BS i to BS j's scheduled UE in slot t, then the *equivalent channel gain* of that channel can be written as $g_{ji}^{(t)} \triangleq \text{PL}(d_{ji}) G_{ji}^{\text{Tx}} G_{ji}^{\text{Rx}} |h_{ji}^{(t)}|^2$ in which $\text{PL}(d) \triangleq 1/d^2$ denotes the path loss with d being the distance, and G_{ji}^{Tx}, G_{ji}^{Rx} denote the corresponding BS and UE antenna gain respectively. The BSs employ a sectorized transmission model [27] in which the BS antennas have a constant mainlobe radiation gain $G^{\max}$ and a constant sidelobe gain $G^{\min}$. The beamwidth is usually chosen as $120°$. The main-to-sidelobe ratio (MSR) is defined as MSR (dB) $\triangleq 10\log_{10}\left(G^{\max}/G^{\min}\right)$.

Let $\boldsymbol{p}^{(t)} \triangleq [p_1^{(t)}, \cdots, p_K^{(t)}]$ denote the transmit powers of the BSs at time t subject to $p_k^{(t)} \leq p_k^{\max}, \forall k$, then the signal-to-interference-plus-noise ratio (SINR) at BS i's scheduled UE is equal to

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

where σ^2 is the total noise power over the shared bandwidth. The throughput of BS i in slot t is equal to $C_i^{(t)} \triangleq W \log_2(1 + \text{SINR}_i^{(t)})$ where W is the total bandwidth. We aim to maximize the expected (over random channels) total throughput of the network

$$\mathbb{E}\left(C_{\text{sum}}^{(t)}\right) = \mathbb{E}\left(\sum_{i=1}^{K} C_i^{(t)}\right) \qquad (3)$$

subject to the power constraints $p_i^{(t)} \leq p_i^{\max}, \forall i$. This problem is nonconvex and finding the optimal solutions can be challenging in general.

III. Proposed Approach

A. Overview of RL

In reinforcement learning (RL), an agent aims to maximize the expected reward through repeated interactions with the environment over time. This process is usually mathematically modeled as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, R, T)$ where $\mathcal{S}, \mathcal{A}, R$ and T denotes the state space, action space, reward function and the state transition kernel respectively. Given some initial state $s_t \in \mathcal{S}$, 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 T) to a new state s_{t+1} and the agent receives a reward $r_t = R(s_t, a_t, s_{t+1})$ as an indication of how good a_t is. The collection of transition quadruples $\{(s_t, a_t, r_t, s_{t+1}), \forall t\}$ is called *experience*. The *return* G_t is defined as the cumulative future rewards $G_t \triangleq \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}$ with $\gamma \in (0, 1]$ being the discount factor. The Q-function Q^μ under a specific policy μ is defined as the expected return starting from any state-action pair (s, a), i.e., $Q^\mu(s, a) \triangleq \mathbb{E}\left[G_t | s_t = s, a_t = a\right]$.

Deep reinforcement learning (DRL) uses deep neural networks (DNN) to represent the policy (called *actor*) and Q-function (*critic*) in order to take advantage of DNN's representation capability. Deep Deterministic Policy Gradient (DDPG) [28] 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 with parameters (i.e., weights and biases) θ^μ and θ^Q are used to represent the policy $\mu(s|\theta^\mu)$ and the Q-function $Q(s, a|\theta^Q)$. Multi-agent DDPG (MA-DDPG) [23] is an adaptation of DDPG to the multiple-agent domain to combat the nonstationarity issue as mentioned in Section I.

MA-DDPG uses a centralized-training-distributed-execution framework in which the actors and critics are trained periodically with network-level experiences while the action of each agent is determined based solely on that agent's local observation. In particular, let $\boldsymbol{a} \triangleq (a_i)_{i=1}^K$ and $\boldsymbol{s} \triangleq (o_i)_{i=1}^K$ denote the joint actions and observations of all agents where o_i denotes the local observation of agent i. The learning process is described as follows. The critic and actor of agent i is represented by two DNNs $Q_i(\boldsymbol{s}, \boldsymbol{a}|\theta_i^Q)$ and $\mu_i(o_i|\theta_i^\mu)$. In order to make the learning more data-efficient, an *experience replay buffer* $\mathcal{D}$ is used to store the past experiences of all agents in a sliding-window manner. Mini-batches of experiences are then sampled repeatedly from $\mathcal{D}$ to train the actors and critics using SGD. More specifically, given a mini-batch of samples $\mathcal{B} = \{(\boldsymbol{s}^j, \boldsymbol{a}^j, \boldsymbol{r}^j, \boldsymbol{s}'^j)\}_j$ where $\boldsymbol{r}^j = (r_i^j)_{i=1}^K$ are the rewards of the agents, the critic network θ_i^Q of agent i is trained by minimizing the loss

$$L(\theta_i^Q) = \frac{1}{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \Big(y_i^j - Q_i\big(\boldsymbol{s}^j, (a_k^j)_{k=1}^K | \theta_i^Q\big)\Big)^2. \qquad (4)$$

$y_i^j \triangleq r_i^j + Q_i'(\boldsymbol{s}'^j, (a_k')_{k=1}^K | \theta_i^{Q'})\big|_{a_k' = \mu_k'(o_k'^j|\theta_k^{\mu'}), \forall k}$ is the regression target and is generated by two *target networks* $Q_i'(\cdot|\theta_i^{Q'})$ and $\mu_i'(\cdot|\theta_i^{\mu'})$. Note that $\boldsymbol{s}'^j \triangleq (o_k'^j)_{k=1}^K$ denotes the joint observations following the sample $\boldsymbol{s}^j$ in time. Also note that in (4) the Q-function of agent i has an input including the local observations and actions of all agents which is made possible by the use of the replay buffer. The actor network θ_i^μ of agent i is trained by minimizing

$$L(\theta_i^\mu) = -\frac{1}{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} Q_i\big(\boldsymbol{s}^j, (a_k^j)_{k\neq i}, a_i | \theta_i^Q\big)\big|_{a_i = \mu_i(o_i^j|\theta_i^\mu)} \qquad (5)$$

which is equivalent to maximizing the Q-function. Note that in the input of the Q-function in (5), the action of agent i is generated by the actor θ_i^μ while the actions of other agents are sampled from the buffer. Finally, the DNN parameters of the target networks are updated by

$$\begin{aligned} \theta_i^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta_i^{Q'}, \\ \theta_i^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta_i^{\mu'}, \ \forall i \end{aligned} \qquad (6)$$

for some small number $\tau \in (0,1)$. In MA-DDPG, exploration is achieved by adding a random noise n_t to the actor output, i.e., $a_i^{(t)} = \mu_i(o_i^{(t)}|\theta_i^\mu) + n_t$ which is clipped to within a proper range to be consistent with the action space definition.

B. Proposed Power Allocation Scheme

Each BS i is treated as a DRL agent that is equipped with an actor for action selection and a critic for evaluating the Q-function. In particular, the actor μ_i is represented by a DNN θ_i^μ and produces action $a_i = \mu_i(o_i|\theta_i^\mu)$, with o_i being the observation of agent i. The critic Q_i is also represented by a DNN θ_i^Q. It is assumed that each BS does not know the actions of other BSs. The MA-DDPG algorithm with a centralized-training-distributed-execution framework is adopted as shown in Fig. 1. The system is fully synchronized where all BSs choose their powers simultaneously at the beginning of each slot based on their local observations. The system then transitions to a new state and the experiences of all agents are pushed to a centralized replay buffer $\mathcal{D}$ in the form of $(o_i, a_i, r_i, o_i'), i = 1, \cdots, K$. The replay buffer operates in a FIFO manner so the oldest experiences are ejected when the buffer is full. In each slot, the actor and critic networks are trained with mini-batches of data sampled from $\mathcal{D}$ according to (4), (5). In addition, the parameters of the target networks are updated softly according to (6). The action, observation and reward are defined as follows.

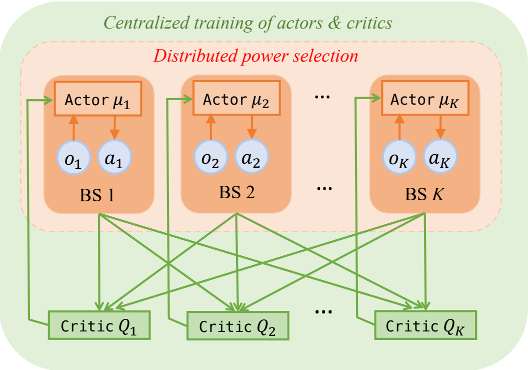

Fig. 1: Proposed MA-DDPG-based power allocation. Each BS k is equipped with an actor μ_k and a critic Q_k.

Action. Each BS needs to determine the transmit power $p_i^{(t)} \in [0, p_i^{\max}]$ to its scheduled UE. Since the tanh() activation is used at 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 by $p_i^{(t)} = \frac{a_i^{(t)} + w}{2w} p_i^{\max}$.

Observation. Becasue the observation reflects each agent's perception of the radio environment, it has to be properly defined such that the relevant features of the the environment can be captured. The observation of agent i (in slot t) is defined as $o_i^{(t)} \triangleq o_{i,1}^{(t)} \cup o_{i,2}^{(t)}$ where

$$\begin{aligned} o_{i,1}^{(t)} &\triangleq \left\{ p_i^{(t-1)}, g_{ii}^{(t-1)}, g_{ii}^{(t)}, I_i^{(t-1)}, \widehat{I}_i^{(t)}, C_i^{(t-1)}, \tfrac{C_i^{(t-1)}}{\sum_{k=1}^K C_k^{(t-1)}} \right\}, \\ o_{i,2}^{(t)} &\triangleq \left\{ g_{ij}^{(t-1)} p_j^{(t-1)}, g_{ij}^{(t)} p_j^{(t-1)}, C_j^{(t-1)}, j = 1, ..., K \right\}. \end{aligned} \qquad (7)$$

The first part $o_{i,1}^{(t)}$ contains several general measurements local to BS i. In particular, $p_i^{(t-1)}$ is BS i's power in the previous slot, $g_{ii}^{(t-1)} = \mathrm{PL}(d_{ii}) G_{ii}^{\mathrm{Tx}} G_{ii}^{\mathrm{Rx}} |h_{ii}^{(t-1)}|^2$ is the direct channel gain in slot $t-1$ which can be estimated via pilot training, $g_{ii}^{(t)} = \mathrm{PL}(d_{ii}) G_{ii}^{\mathrm{Tx}} G_{ii}^{\mathrm{Rx}} |h_{ii}^{(t)}|^2$ is the direct channel gain in slot t. Since the channel changes from $h_{ii}^{(t-1)}$ to $h_{ii}^{(t)}$ at the very be-

[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.

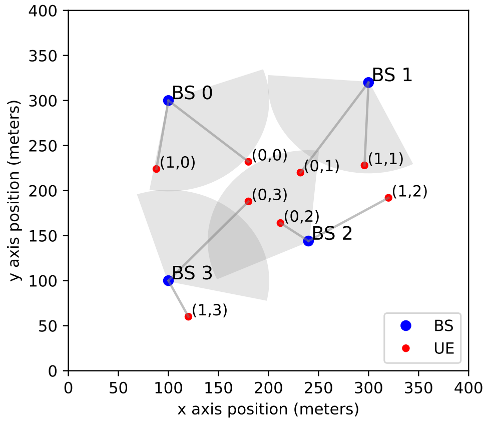

Fig. 2: Network with 4 BSs. Each BS i is associated with two UEs denoted by $(0, i)$ and $(1, i)$.

ginning of slot t before the new power $p_i^{(t)}$ is determined, $g_{ii}^{(t)}$ can also be obtained by BS i. $I_i^{(t-1)} = \sum_{j\neq i} g_{ij}^{(t-1)} p_j^{(t-1)} + \sigma^2$ is the total received interference at BS i in slot $t-1$ and $\widehat{I}_i^{(t)} = \sum_{j\neq i} g_{ij}^{(t)} p_j^{(t-1)} + \sigma^2$ is the interference measured at the beginning of slot t where the channels have changed but the powers have not changed. $C_i^{(t-1)}$ is the throughput of BS i in slot $t-1$, and $C_i^{(t-1)} / \sum_j C_j^{(t-1)}$ represents the contribution of BS i to the total throughput. The second part $o_{i,2}^{(t)}$ contains measurements of received power and throughput of other BSs. In particular, $g_{ij}^{(t-1)} p_j^{(t-1)}$ and $g_{ij}^{(t)} p_j^{(t-1)}$ are the received power of BS $j (\neq i)$ measured at BS i in slot $t-1$ and the beginning of slot t respectively. $C_j^{(t-1)}$ is the throughput of 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 by BS i. We include one previous slot in order for the agents to better keep track of the temporal correlated channels.

State transition. The measured interference in each BS's observation (7) is determined by the random small-scale fading and the BS transmit powers. Given the current state (i.e., joint observations of all BSs) in slot t, once the powers are chosen, the system will transition to a random new state.

Reward. The reward of BS i (in slot t) is defined as the total throughput of the network, i.e., $r_i^{(t)} = \sum_j C_j^{(t)}$. This definition is intuitive and avoids complicated reward design at the cost of a slight communication overhead.

IV. Simulation

A. *Setup*

First consider a network with 4 BSs as shown in Fig. 2 where each BS is associated with two UEs. The transmit beams (with 120° coverage) are aligned with the scheduled UEs (See Fig. 2 where the 0^{th} UE of each BS is scheduled). The maximum power is chosen as $p_i^{\max} = 30$ dBm (1 Watt) for all BSs as per the FCC regulation. The BS antenna gain (MSR) is chosen as 20 dB. The total noise is calculated according to $\sigma^2\,(\text{dBm}) = 10\lg(\kappa_B T_0 \times 10^3) + \text{NR}\,(\text{dB}) + 10\lg W$ where W is the shared bandwidth. κ_B, NR and T_0 denote 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 $\sigma^2 = -85.49$ dBm. The Nakagami fading parameters are chosen as $m = 50, \Omega = 1$ and $\rho = 0.1$. Each actor and critic is represented by a fully-connected feedforward DNN with five layers including the input and output layers. The three hidden layers contain $256, 256$ and 64 neurons respectively with ReLU activation. Each actor network has one output port with Tanh activation clipped to within the range $[-0.96, +0.96] (w = 0.96)$.

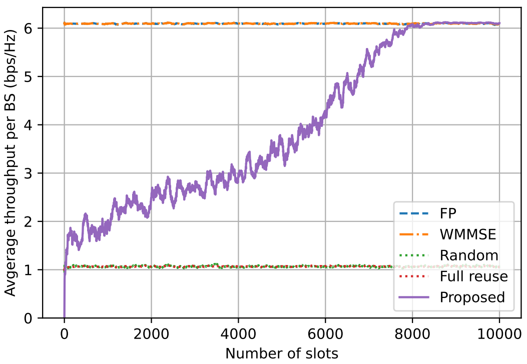

Fig. 3: Throughput performance when the 0^{th} UEs are scheduled.

The proposed scheme is implemented with PyTorch. The Adam optimizer is used with learning rates 10^{-4} and 10^{-3} for the actor and critic networks respectively. The action noise $\{n_t, \forall t\}$ is chosen as i.i.d. Gaussian noise with a decreasing variance, i.e., $n_t \sim \mathcal{N}(0, \sigma_t^2)$ where $\sigma_{t+1} = \max\{(1 - 10^{-4})\sigma_t, 0.001\}$, $\sigma_0 = 1$. This guarantees adequate exploration in the early stage of learning. The replay buffer is implemented as a FIFO queue with size $|\mathcal{D}| = 5 \times 10^5, \forall i$. Batch size is chosen as $|\mathcal{B}| = 128$. Other parameters are chosen as $\gamma = 0.9, \tau = 0.005$. The simulation consists of two phases, the training phase each containing 10^4 slots, and the testing phase each containing 2000 slots. During the testing phase, the learned policy is tested over another set of channel realizations that is different from (but is identically distributed) the training phase.

The proposed scheme is compared with two baseline power allocation schemes WMMSE [24] and FP [25]. These are both centralized algorithms which require the knowledge of all cross channels. 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 that is used for the entire slot.

B. *Results*

Fig. 3 shows the throughput performance in the training phase when the 0^{th} UEs are scheduled in the 4-BS network. The throughput in each slot is an average with the previous 100 slots. It can be seen that the proposed scheme converges in 8000 slots and achieves very close performance to WMMSE

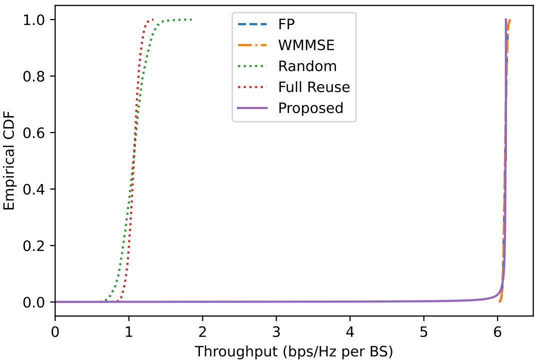

Fig. 4: Test CDF when the 0^{th} UEs are scheduled.

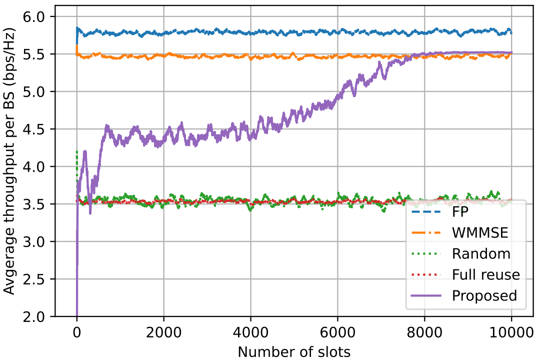

Fig. 5: Throughput performance when the 1^{st} UEs are scheduled.

and FP. The empirical CDF of the testing phase is shown in Fig. 4. Although the channels are different, the proposed scheme still obtains close performance to WMMSE and FP. The throughput performance in the training and testing phase when the 1^{st} UEs are scheduled are shown in Fig. 5 and Fig. 6 respectively. In this case, the proposed scheme achieves better performance than WMMSE and is slightly worse than FP.

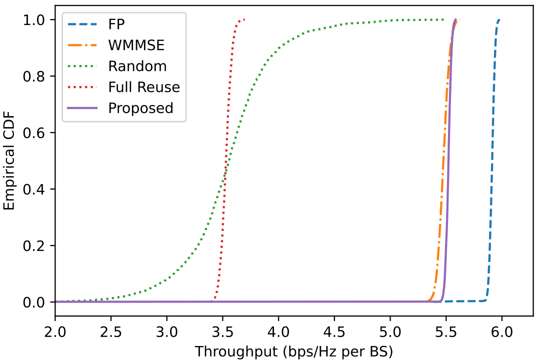

Fig. 6: Test CDF when the 1^{st} UEs are scheduled.

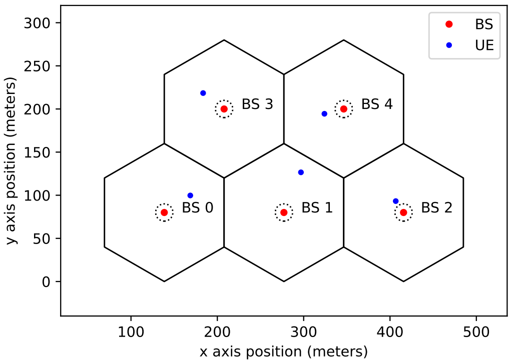

Fig. 7: Network with 5 BSs. Each BS is associated with one UE located within its hexagonal cell.

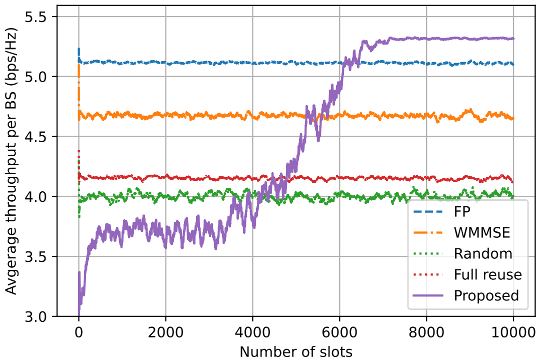

Fig. 8: Throughput performance over 5-BS network

Note that unlike centralized WMMSE and FP, our proposed scheme is distributed and does not require the full knowledge of all channels. From Fig. 4 and Fig. 6, it can be seen that the proposed scheme maintains a close performance to the training phase. This demonstrates its robustness against channel change as it is trained over distributions but not specific channel realizations like WMMSE and FP.

We further evaluate the proposed scheme over another network with 5 BSs as shown in Fig. 7. Each BS has one scheduled UE located within its hexagon cell. The UE positions are generated randomly but their distance should be no less than 10 meters to the BS. Fig. 8 shows the training performance. It can be seen that the proposed scheme outperforms the baseline schemes by a significant margin, i.e., 16% more than WMMSE and 5% more than FP. The learned policy is also tested over two different channel realizations and the average throughput is shown in Fig. 9. It can be seen that the proposed scheme also achieves higher throughput than the baselines in the testing phase.

V. Conclusion

In this work, we presented a novel distributed power allocation scheme based on multi-agent Deep Deterministic

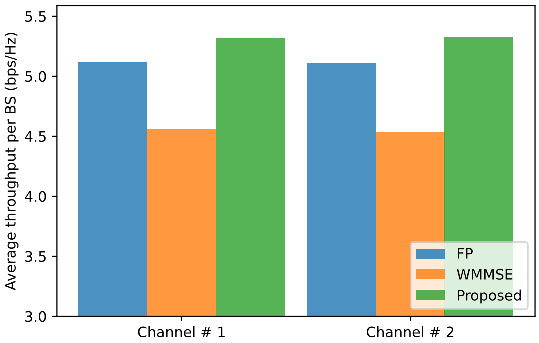

Fig. 9: Comparison of average throughput in the testing phase.

Policy Gradient. The proposed scheme utilizes the centralized-training-distributed-execution framework where the actors and critics are trained periodically using the accumulated data from the experience replay buffer, while each actor determines the transmit power based solely on its local information. By a careful design of the observation space, the proposed scheme has been shown to outperform state-of-the-art approaches including WMMSE and FP.

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] 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.

[2] A. K. Gupta, J. G. Andrews, and R. W. Heath, "On the feasibility of sharing spectrum licenses in mmwave cellular systems," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3981–3995, 2016.

[3] E. A. Jorswieck, L. Badia, T. Fahldieck, E. Karipidis, and J. Luo, "Spectrum sharing improves the network efficiency for cellular operators," *IEEE Communications Magazine*, vol. 52, no. 3, pp. 129–136, 2014.

[4] Wikipedia contributors, "Unlicensed national information infrastructure — Wikipedia, the free encyclopedia," 2022, [Online; accessed 1-March-2023]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Unlicensed_National_Information_Infrastructure&oldid=1129539949

[5] 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.

[6] ——, "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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Q. Xue, Y.-J. Liu, Y. Sun, J. Wang, L. Yan, G. Feng, and S. Ma, "Beam management in ultra-dense mmwave network via federated reinforcement learning: An intelligent and secure approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 1, pp. 185–197, 2023.

[19] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for miso communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 745–749, 2020.

[20] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, 2020.

[21] K. Yang, C. Shen, and T. Liu, "Deep reinforcement learning based wireless network optimization: A comparative study," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2020, pp. 1248–1253.

[22] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2021.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.