A Q-Learning-Based Approach for Distributed Beam Scheduling in mmWave Networks

Changing the World's Energy Future

Xiang Zhang, Shamik Sarkar, Sneha Kasera, Mingyue Ji, Arupjyoti Bhuyan

Idaho National Laboratory

A Q-Learning-Based Approach for Distributed Beam Scheduling in mmWave Networks

Xiang Zhang, Shamik Sarkar, Sneha Kasera, Mingyue Ji, Arupjyoti Bhuyan

December 2022

**Idaho National Laboratory
Idaho Falls, Idaho 83415**

http://www.inl.gov

A Q-Learning-Based Approach for Distributed Beam Scheduling in mmWave Networks

Xiang Zhang*, Shamik Sarkar†, Arupjyoti Bhuyan‡, Sneha Kumar Kasera†* and Mingyue Ji*†

Department of Electrical and Computer Engineering, University of Utah*

School of Computing, University of Utah †

Idaho National Laboratory‡

Email: *{xiang.zhang, mingyue.ji}@utah.edu,†shamik.sarkar@utah.edu, †kasera@cs.utah.edu, ‡arupjyoti.bhuyan@inl.gov

Abstract—We consider the problem of distributed downlink beam scheduling and power allocation for millimeter-Wave (mmWave) cellular networks where multiple base stations (BSs) belonging to different service operators share the same unlicensed spectrum with no centralized control or explicit coordination among them. Our goal is to design efficient distributed beam scheduling and power allocation schemes such that the network-level payoff, defined as the weighted sum of the total throughput and a power penalization term, can be maximized. To this end, we propose a distributed scheduling approach to power allocation and adaptation for efficient interference management over the shared spectrum by modeling each BS as an independent Q-learning agent. As a baseline, we compare the proposed approach to the state-of-the-art non-cooperative game-based approach which was previously developed for the same problem. We conduct extensive experiments under various scenarios to verify the effect of multiple factors on the performance of both approaches. Experiment results show that the proposed approach adapts well to different interference conditions by learning from experience and thus can achieve higher payoff than the game-based approach. The proposed approach can also be integrated into our previously developed Lyapunov stochastic optimization framework for the purpose of network utility maximization with optimality guarantee. As a result, the weights in the payoff function can be automatically and optimally determined by the virtual queue values from the sub-problems derived from the Lyapunov optimization framework.

Index Terms—mmWave, distributed scheduling, reinforcement learning, Q-learning, optimality

I. INTRODUCTION

The proliferation of mmWave frequencies in 5G cellular networks has increased wireless bandwidth by orders of magnitude and has also enhanced spectrum availability. Spectrum sharing enables the secondary utilization of additional unlicensed or shared spectrum that is available by allowing concurrent beam-based transmission and has the potential to largely enhance the the system-level throughput performance [1]–[3]. However, highly directional transmission may also present a severe interference condition, which is even worsened by the dense population of access points and user equipment (UEs), if there is no proper coordination of the beams. To handle interference and improve system throughput, two major paradigms – centralized and distributed, are considered in the literature. In general, centralized approaches can be effective but are usually limited by high complexity and limited scalability especially for large networks.

Distributed approaches [4]–[13], on the other hand, are scalable and have the advantage of improving system security by removing any central point of attack as well as any falsification of reports of spectrum use to the central control server. Some existing distributed approaches [4]–[9], including our own [9], have proposed the use of game theory, for beam scheduling. In [9], combined with the Lyapunov optimization framework, we proposed a distributed game-based beam scheduling for mmWave networks with non-cooperative operators.

In this paper, we propose an alternate approach that uses Q-learning for distributed beam scheduling as well as for power allocation for mmWave networks with non-cooperative operators. Our main contributions are two-fold. First, we present a general framework for dynamic spectrum sharing for the purpose of optimizing a network-level payoff function, which is defined as the sum throughput penalized by power consumption. The weights in the payoff function can be tuned to find a desirable trade-off between throughput maximization and power consumption. This formulation can work for various different beam scheduling methods and therefore, provides a unified framework for performance evaluation and comparison of these methods. Second, under the proposed payoff optimization framework, we apply classical Q-learning due to its simplicity and yet promising performance. We propose a learning-based power allocation algorithm by modeling each BS as an independent Q-learning agent that interacts with the radio environment determined by the joint actions of all BSs and channel uncertainty. We compare the proposed approach with our, state-of-the-art, non-cooperative game-based approach [9] and demonstrate that the our learning approach adapts well to different interference conditions and achieves higher network-level payoff. We conduct simulation-based experiments under various interference scenarios and the results show that our approach can achieve 23% to 80% more payoff than the game-based approach under practical network settings, and the performance gain is more prominent in the relatively low SINR regime. In addition, our approach can be integrated seamlessly into a general network utility maximization framework by using the Lyapunov stochastic optimization proposed in [9]. In this case, the weights in the payoff function can be automatically and optimally determined by the virtual queues derived from the Lyapunov optimization. Further experiments show that the proposed approach can also

achieve 7% to 23% more utility than the game-based approach. This performance gain is significant as the utility is defined as an increasing concave function of the average throughput which has a diminishing marginal utility.

Why Reinforcement Learning: In general, reinforcement learning (RL)-based methods have the advantage of being adaptive to different interference conditions by learning from experience, i.e., past interaction with the environment, the quality of each decision made indicated by the corresponding reward. In addition, by actively exploring non-greedy actions, there is a higher chance of finding the optimal actions in the long run. In contrast, the game-based methods are greedy by nature – regardless of the interference, each BS will always choose an action that maximizes its payoff in the current step. This greedy nature prevents the BSs from exploring non-greedy actions or adapting their decisions to different interference conditions. This motivates the use of Q-learning for adaptive interference management in mmWave networks in this work.

A. Related Work

Besides game-theoretic approaches, another line of research [10]–[16] has focused on learning-based methods. Galindo-Serrano *et al.* [10] considered the interference control problem in cognitive radio systems where a set of UEs aim to maximize their own throughput while ensuring that the aggregate interference to the primary licensed UEs does not exceed a threshold. A decentralized Q-learning algorithm was proposed based on partial observation of the interference state. In [12], Kar *et al.* presented and analyzed a distributed RL algorithm for collaborative multi-agent Markov decision processes. In particular, a distributed variant of Q-learning was proposed where each agent sequentially refines its learning parameters based on the local payoff data and the information received from neighboring agents. Despite these classical methods, the use of deep neural networks (DNNs) as function approximators has also gained tremendous attention recently [13]–[16]. Several works [13], [15], [16] are relevant along this direction. Ghadimi *et al.* [13] proposed a deep RL approach for downlink power control and rate adaptation in order to optimize a utility function based on partial observability of the system state where each user is given an equal share of the bandwidth. In [15], a scalable and distributed multi-agent deep RL framework was proposed for transmit power control in cellular networks by assuming that each transmitter can obtain the CSI and QoS information from neighboring transmitters. Most relevant to our work is [16] where a deep RL-based distributed power allocation and UE scheduling algorithm was proposed using a centralized training and distributed execution paradigm. Each access point can exchange instantaneous interference measurement with its neighbors.

There are, however, several drawbacks of the aforementioned works. First, most of these works assume data exchange among BSs in the vicinity. This may not be practical if these BSs belong to different operators. How these approaches perform in such a fully distributed scenario also needs to be
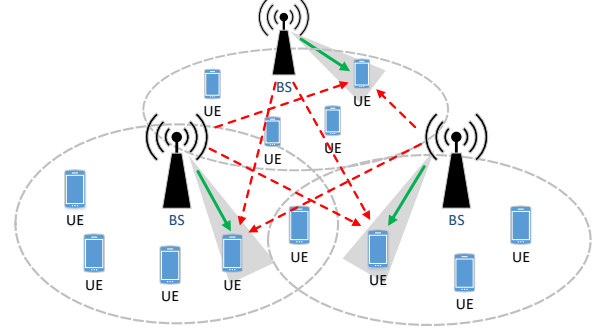

Fig. 1: A cellular network consisting of M BSs and K UEs where each BS is associated with four UEs. The solid green lines represent the data links and the dashed red lines represent the interfering links.

evaluated. Second, the performance of the deep RL-based approaches depend heavily on the size and quality of the training dataset. However, it is not straightforward how these datasets can be obtained in advance. Offline training of the DNNs usually takes a significant amount of time which may not be suitable for delay-critical wireless systems. Third, models trained for a specific network configuration, e.g., number of BS/UEs, topology, fading etc., may not generalize well to other configurations, thus presenting scalability and robustness issues for these approaches.

In this paper, we present a general framework for distributed payoff optimization in non-cooperative mmWave networks and propose a Q-learning-based beam scheduling and power allocation approach using an independent modeling for each agent (i.e., BS) with a simple tabular representation of action-state values. The proposed approach has lower complexity and better scalability than most deep RL-based approaches and is robust to network configuration change. At a similar complexity level, the proposed approach outperforms the game-based approach [9] devised for the same problem. Experiment results demonstrate that the proposed approach achieves a similar performance in the high SINR regime to the game-based approach and beats the game-based approach by a large margin in the relatively low SINR regime.

II. PROBLEM FORMULATION

A. System Description

Consider a cellular network with M BSs $\mathcal{M} \triangleq \{1, 2, ..., M\}$ and K UEs $\mathcal{K} \triangleq \{1, 2, ..., K\}$ as shown in Fig. 1. Each BS belongs to a different service operator and is responsible for serving a set of $|\mathcal{K}_i| = K_i$ UEs within its coverage area. We assume that each UE can only subscribe to one BS and each BS can serve at most one UE at any given time. This means that $\mathcal{K}_i \neq \emptyset, \forall i \in \mathcal{M}$, $\mathcal{K}_i \cap \mathcal{K}_j = \emptyset, \forall i \neq j$, and $\cup_{i \in \mathcal{M}} \mathcal{K}_i = \mathcal{K}$. The BS-UE association $\{\mathcal{K}_i\}_{i \in \mathcal{M}}$ is assumed to be determined by some exogenous mechanism and is fixed during the considered scheduling process. The system

operates synchronously[1] over a shared unlicensed spectrum of bandwidth W Hz. We use a frame structure as shown in Fig. 2. Each time *frame* contains N_f *blocks* and each block contains N_b time *slots* where each slot has a duration of T_s seconds. Therefore, each frame has a duration $T_f = N_f N_b T_s$ seconds and each block has duration $T_b = N_b T_s$ seconds. Beam and
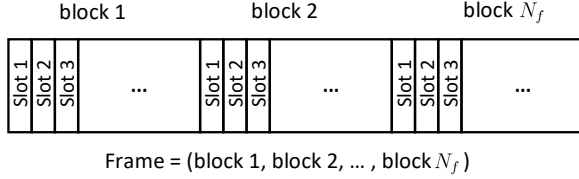

Fig. 2: Frame structure. Each frame contains N_f blocks and each block contains N_b slots.

UE scheduling happens in each block of a frame meaning that the beam and UE selection will stay unchanged in each block but will possibly change over different blocks. BS/UEs are equipped with directional antennas which are characterized by a keyhole-like antenna model that is commonly used in the literature (e.g., by [17], [18] etc). The keyhole model has a constant main-lobe radiation gain $G^{\max}$ and a constant side-lobe gain $G^{\min}$. In particular, the antenna gain $G(\theta)$ in the direction θ is

$$G(\theta) = \begin{cases} G^{\max}, & |\theta| \leq \Theta/2 \\ G^{\min}, & |\theta| > \Theta/2 \end{cases} \quad (1)$$

where Θ is the beamwidth. The antenna also has a total power radiation gain of $E = \Theta G^{\max} + (360° - \Theta)G^{\min}$. We further use $G^{\text{BS}}_{j,i}$ and $G^{\text{UE}}_{j,i}$ to respectively represent the antenna gain of BS_i and UE_j along the direction connecting BS_i and UE_j. The *main to side-lobe gain ratio* (MSR) is defined as $\text{MSR} \triangleq 10 \lg \left(G^{\max}/G^{\min} \right)$ dB. A large MSR means that the antenna has strong radiation in the main-lobe while a small MSR implies energy leakage in the side-lobe. Due to the proximity of locations, the BSs may interfere with the UEs associated with other BSs. Let $\text{UE}_{j_i}(j_i \in \mathcal{K}_i)$ be the UE scheduled by BS_i. Also, for any UE_j, let BS_{i_j} be the BS that UE_j subscribes to ($j \in \mathcal{K}_{i_j}$). The Signal-to-Interference-Noise-Ratio (SINR) at UE_j can be written as

$$\text{SINR}_{j,i_j} = \frac{p_{j,i_j} G^{\text{UE}}_{j,i_j} G^{\text{BS}}_{j,i_j} |h_{j,i_j}|^2 d^{-\eta}_{j,i_j}}{\sum_{\ell \in \mathcal{M}, \ell \neq i} p_{j_\ell, \ell} G^{\text{UE}}_{j,\ell} G^{\text{BS}}_{j,\ell} |h_{j,\ell}|^2 d^{-\eta}_{j,\ell} + \sigma^2}, \quad (2)$$

where $p_{j,i}$ denotes the transmit power of BS_i to UE_j if UE_j is served by BS_i; η is the path-loss factor; $\sigma^2 = N_0 W$ is the power of the random Gaussian noise (N_0 is the noise power spectrum density); $h_{j,i}$ is the small-scale fading between UE_j and BS_i which is assumed to follow the Nakagami-m distribution [19] with probability density

$$f(h; \mu, \Omega) = \frac{2\mu^\mu}{\Gamma(\mu)\Omega^\mu} h^{2\mu-1} \exp\left(-\frac{\mu}{\Omega}h^2 \right), \ h \geq 0, \quad (3)$$

where $\mu \triangleq \frac{\mathbb{E}[h^2]^2}{\text{Var}(h^2)}$, $\Omega \triangleq \mathbb{E}[h^2]$ and Γ is the Gamma function. We assume a block fading channel where the fading coefficients stay unchanged during each frame and are i.i.d. over different frames[2]. We further define the *equivalent channel gain* g_{j,i_j} between UE_j and BS_{i_j} as $g_{j,i_j} \triangleq \text{SINR}_{j,i_j}/p_{j,i_j}$ if UE_j is scheduled and $p_{j,i_j} \neq 0$.

B. Payoff Maximization

Each BS is subject to an instantaneous peak transmit (TX) power constraint in each slot, i.e., $\sum_{j \in \mathcal{K}_i} p_{j,i} \leq p^{\max}_i$. Since it is assumed that at most one UE can be scheduled at a time, we have $p_{j_i,i} \leq p^{\max}_i$ where UE_{j_i} is the scheduled UE by BS_i. Let $\boldsymbol{p} \triangleq \{p_{j_i,i}\}_{i \in \mathcal{M}}$ denote the TX powers of the BSs to their respective scheduled UEs. We consider a general form of *payoff* function (for a unit time duration of one second) for each BS defined as

$$R_i(\boldsymbol{p}) \triangleq \alpha_i W \log(1 + \text{SINR}_{j_i,i}) - \beta_i p_{j_i,i}, \quad (4)$$

i.e., the payoff of BS_i is the throughput of its scheduled UE (weighted by α_i) plus a power penalizing term (weighted by β_i). The weights $\alpha_i, \beta_i \geq 0$ can be tuned manually or determined using some algorithms[3] in order to find a desirable trade-off between throughput and power consumption. In particular, the ratio α_i/β_i determines the relative importance of throughput maximization to power consumption. If α_i/β_i is very large, eq. (4) becomes equivalent to maximizing the throughput $R_i(\boldsymbol{p}) \approx \alpha_i W \log(1 + \text{SINR}_{j_i,i})$. Note that the solution becomes trivial when either α_i or β_i is equal to zero. For any given set of scheduled UEs $\{j_i\}_{i \in \mathcal{M}}$, we aim to find efficient power allocation schemes to maximize the sum payoff $R(\boldsymbol{p})$ of all BSs $R(\boldsymbol{p}) \triangleq \sum_{i \in \mathcal{M}} R_i(\boldsymbol{p})$. Let $\boldsymbol{p}(t)$ denote the power allocation profile in slot t. Our goal is to maximize the long-term average payoff

$$\bar{R} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} R(\boldsymbol{p}(t)). \quad (5)$$

The challenge lies in that this sum payoff maximization problem must be solved in a distributed manner, that is, there is no centralized control or coordination among the BSs as they belong to different operators. It should also be noted that the above formulation is not particular to any specific scheduling method so new scheduling methods can be developed under the same framework and be effectively evaluated by comparing to previous methods.

III. PROPOSED APPROACH

Under the general formulation in Section II-B, we propose to solve the payoff maximization problem (5) using Q-learning by modeling each BS as an independent learning agent that interacts with the radio environment which is governed by the

---

[1]This can be achieved by letting each BS synchronizing with the GPS or other synchronization mechanisms.

[2]Although UE mobility is not considered here, the proposed approach still applies if the UEs have mobility to some extent for which the proposed approach can keep track of.

[3]An example is presented in Section V-A where the weights are determined by the queue values derived from the Lyapunov optimization framework.

collective behavior of all agents and channel uncertainty. By properly defining the state space and rewards, the proposed learning-based beam scheduling and power allocation is able to outperform the game-theoretic (GT) approach [9] – a previously developed iterative power allocation algorithm for the considered mmWave scheduling problem, especially in the interference-limited regime. In the following, we first present a brief background of Q-learning and then proceed to the description of the proposed approach.

A. Q-learning Preliminary

In RL, an agent interacts with the environment by making decisions that may affect the state of the environment in a sequence of discrete time steps. In particular, at time t, based on the observation of the current state $s^{(t)}$ of the environment, the agent takes an *action* $a^{(t)}$ according to a *policy* π as $a^{(t)} \sim \pi(\cdot|s^{(t)})$ with a special case of being deterministic with $a^{(t)} = \pi(s^{(t)})$. After taking the action, the agent receives an immediate reward $r^{(t)}$ which indicates the quality of the chosen action $a^{(t)}$ in state $s^{(t)}$. As a result of the above interaction, the environment transitions to a new state $s^{(t+1)}$. The goal of RL is to maximize the agent's long-term expected return $G^{(t)}$ defined as $G^{(t)} \triangleq \sum_{k=0}^{\infty} \gamma^k r^{(t+k+1)}$, where γ is the discount factor which indicates the importance of future rewards. Model-free RL aims to find a an optimal policy π^* that maximizes the expected return $G^{(t)}$ by learning directly from the agent-environment interactions represented by a set of quadruples $\left\{ (s^{(\ell)}, a^{(\ell)}, r^{(\ell)}, s^{(\ell+1)}) : \ell \leq t \right\}$ called *experience* (up to time t), without any specific knowledge of the underlying transition probabilities of the environment.

Q-learning is a model-free off-policy learning algorithm for estimating the optimal action-state values $q_*(a, s)$ for each action-state pair $(a, s) \in \mathcal{A} \times \mathcal{S}$ ($\mathcal{A}$ and $\mathcal{S}$ denote the action and state space respectively). Let $Q(s, a)$ denote an estimate of $q_*(a, s)$. At time t, the agent chooses its action using the ϵ-greedy action selection method, that is, with a small probability ϵ (also termed as exploration rate), the agent chooses a random action in $\mathcal{A}$; otherwise it chooses a greedy action $a^{(t)} = \arg\max_{a \in \mathcal{A}} Q(a, s^{(t)})$. After the selection, the action-state values are updated according to

$$Q\left(a^{(t)}, s^{(t)}\right) \leftarrow (1 - l_r)Q\left(a^{(t)}, s^{(t)}\right)$$
$$+ l_r \left(r^{(t)} + \gamma \max_{a \in \mathcal{A}} Q\left(a, s^{(t+1)}\right) \right), \quad (6)$$

and $Q(a, s)$ does not update if $(a, s) \neq (a^{(t)}, s^{(t)})$. $l_r \in (0, 1]$ is the learning rate which determines to what extent the new estimate $r^{(t)} + \gamma \max_{a \in \mathcal{A}} Q(s^{(t+1)}, a)$ overrides the old estimate $Q(a^{(t)}, s^{(t)})$. Q-learning usually employs a tabular representation $[Q(a, s)]_{|\mathcal{A}| \times |\mathcal{S}|}$, the Q-table, to store the estimated action-state values. For continuous action or state spaces, function approximators (e.g., linear, DNN) can be used [20], [21]. For a stationary underlying transition model, the Q-learning algorithm converges to with probability one asymptotically if 1) The learning rate $l_r(t)$ at time t satisfies $\sum_{t=1}^{\infty} l_r(t) = \infty, \sum_{t=1}^{\infty} l_r(t)^2 < \infty$, and 2) Every action-

state pair is visited infinitely many times. For optimizing an expected return over a finite horizon T, a constant learning rate l_r can be used.x'f

B.

One key feature of the learning-based methods, specifically Q-learning that will be used in this work, is the ability to adapt by learning from experience and exploring, going beyond the mere greedy nature of the game-based methods. One major challenge in the considered mmWave scheduling problem is how to handle the strong interference due to the lack of centralized coordination of beams. Being purely greedy in this scenario can potentially hurt the overall performance. In particular, if we model each BS as a non-cooperative game player that myopically focuses on maximizing its own payoff (say the throughput) in each slot, then each BS will always choose the maximum power to transmit since it gets maximum throughput from this decision. However, if the beams of different BSs overlap, there will be very strong interference at the scheduled UEs, which in turn yields a small network-level payoff (also see Section IV-B for a detailed analysis). What is even worse is that this situation can happen over and over again as the BSs do not learn from these bad experiences. In contrast, if we model each BS as a Q-learning agent, the case of overlapping beams can still occur. However, the decisions of the BSs can be very different from the game-based methods. First, each BS can explore non-greedy actions using the ϵ-greedy action selection, partly avoiding the maximum TX power dilemma. Second, each BS can also learn from its past experience to improve the performance. If the overlapping beam situation happens and the BS has chosen the maximum power, then it will receive a small reward ($r^{(t)}$ in (6)) due to strong inter-cell interference. This will inform the BS to avoid using maximum power in similar situations in the future and thus improves the long-term throughput performance.

C. Proposed Beam Scheduling and Power Allocation

Due to the adaptation ability of Q-learning as described in the previous section and its simplicity, we focus on applying the classical Q-learning algorithm to the considered mmWave scheduling problem. In particular, we model each non-cooperative BS as an independent learning agent that implements the Q-learning algorithm presented in the previous section in parallel. The key Q-learning components for each agent are defined as follows.

1) *Environment*: Each agent interacts with the physical radio environment governed by the collective behaviors, e.g., UE scheduling, TX powers, beam generation, etc., of the BSs subject to random channel realization.

2) *Action*: The action for BS_i in each slot is the TX power $p_{j_i, i}^{(t)}$. To use the tabular representation of Q-learning, the action and state spaces must be discrete. Therefore, we quantize the TX power range $[0, p_i^{\max}]$ uniformly into P_q discrete levels $\mathcal{P}_q = \{p_i^1, p_i^2, \cdots, p_i^{P_q}\}$ to represent the

action space where

$$p_i^j = (j-1)\frac{p_i^{\max}}{P_q - 1}, \ j \in \{1, \cdots, P_q\}. \quad (7)$$

This means $p_i^1 = 0$ and $p_i^{P_q} = p_i^{\max}$. The same uniform power quantization is used by all BSs.

3) *Observation*: Each BS's observation of the environment is defined as the received (RX) interference (plus noise) at its scheduled UE. Let $I_{j_i,i}$ denote the measured RX interference at UE_{j_i}. Suppose $I_{j_i,i}$ follows a (possibly unknown) distribution $\mathcal{D}_{j_i,i}$ over the range $[I_{j_i,i}^{\min}, I_{j_i,i}^{\max}]$ with $I_{j_i,i}^{\min}$ and $I_{j_i,i}^{\max}$ being the minimum and maximum possible interference respectively. The RX interference also needs to be quantized in order to be represented by a discrete *state*. We propose a *percentile-based* quantization method as follows. We first derive I_q percentiles $\mathcal{I}_q = \{I_1, I_2, \cdots, I_{I_q}\}$ over the distribution $\mathcal{D}_{j_i,i}$. This means that the probability that $I_{j_i,i}$ falls into any interval $(I_j, I_{j+1}]$ is identical and is equal to $1/I_q, \forall j \in \{1, \cdots, I_q - 1\}$. If the measured interference $I_{j_i,i}$ falls into the interval $(I_j, I_{j+1}]$, we say that the observation of BS_i is 'state j'. Therefore, the state space of BS_i can be represented by $\mathcal{S}_i = \{1, 2, \cdots, I_q\}$. The proposed quantization method guarantees that each state will be visited approximately the same number of times in the long run which facilitates convergence. An illustration of the percentile-based quantization method with $I_q = 10$ states is shown in Fig. 3. We assume all BSs use the same number of states. It should be noted that the UE interference distributions are not know by the BSs so they have to be estimated (See the Training phase later in this section), after which the above state quantization can be performed.
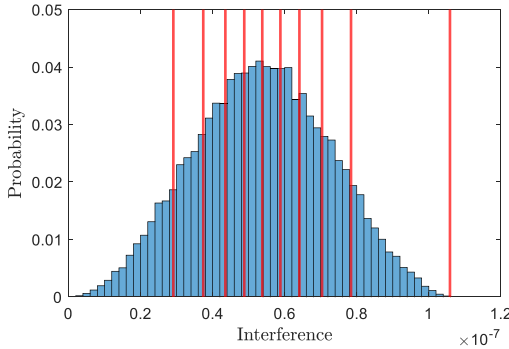

Fig. 3: Percentile-based interference quantization with ten levels based on an empirical interference distribution.

4) *Reward*: The reward of BS_i in slot t is defined as

$$r_i^{(t)} \triangleq \alpha_i \left(T_s W \log \left(1 + \text{SINR}_{j_i,i}^{(t)} \right) \right) - \beta_i \left(T_s p_{j_i,i}^{(t)} \right), \quad (8)$$

where $\text{SINR}_{j_i,i}^{(t)}$ is the SINR at UE_{j_i} in slot t. The goal of BS_i is to maximize the long-term expected (discounted) return

$$G_i^{(t)} = \sum_{k=0}^{\infty} \gamma^k r_i^{(t+k+1)} \quad (9)$$

starting from any time t. It should be noted that when the discount factor γ is close to 1, eq. (9) can be used to approximate (5) after averaging over time.

With the above definitions of the action, observation/state and the reward function, we propose to solve the sum payoff maximization problem (5) by letting each BS 'selfishly' maximize its own average payoff $\bar{R}_i \triangleq \lim_{T \to \infty} \frac{1}{T} R_i(\boldsymbol{p}(t))$. To do this, we model each BS as an independent learning agent implementing the ϵ-greedy action selection method with the goal of optimizing its long-term expected return (9). For a any finite T and $\gamma \approx 1$, optimizing $\bar{R}_i = \frac{1}{T} \sum_{t=1}^{T} R_i(\boldsymbol{p}(t))$ becomes equivalent to optimizing (9). Therefore, we have provided a fully distributed approach using Q-learning in a multi-agent scenario. The proposed beam scheduling and power allocation scheme consists of a training phase followed by an execution phase, which are described as follows.

Training Phase: This phase is to estimate the empirical distribution of the RX interference at each UE so that the interference quantization can be performed. In particular, for the set of scheduled UEs $\{j_i\}_{i \in \mathcal{M}}$, we run T_{train} frames of 'simulated scheduling' in which the TX powers of the BSs are chosen randomly from $\mathcal{P}_q$ in each slot and the wireless channels are subject to change from frame to frame. We record the interference at each scheduled UE_{j_i} in all the training frames and derive an empirical interference distribution $\hat{\mathcal{D}}_{j_i,i}$, which will be used to quantize the RX interference in the execution phase. Note that during the training phase, although the powers are randomly selected, the BS/UEs still achieve some data throughput in each slot. Moreover, this training phase only needs to be done once before the 'real' scheduling begins, so the overhead induced by this phase becomes negligible if we consider the scheduling problem over a large number of frames.

Execution Phase: Beam scheduling and power allocation happen in this phase where the frame structure of Fig. 2 is used. Since we do not consider UE scheduling in this paper, the UEs can be scheduled randomly or in a round robin manner in different blocks. Therefore, we focus on the application of the proposed scheduling approach in one block. Each BS implements the Q-learning algorithm as follows. At the beginning of slot t, based on the current state which is defined as the quantized RX interference at UE_{j_i} in slot $t-1$ (this interference is measured by UE_{j_i} and then feedback to BS_i), BS_i chooses TX power $p_{j_i,i}^{(t)}$ according to the ϵ-greedy action selection method, and then generates a beam towards UE_{j_i} and starts the data transmission. Note that no beams will be generated if $p_{j_i,i}^{(t)} = 0$. After the beam generation, BS_i updates its Q-table according to (6) where the next state $s^{(t+1)}$ is defined as the quantized RX interference at UE_{j_i} in slot t (after the power selection), and the reward $r_i^{(t)}$ is defined in (8). The above process is repeated until the end of the current block. The proposed approach, performed in one block, is summarized in Algorithm 1. We use the notation $[n] \triangleq \{1, 2, \cdots, n\}$ for $n \in \mathbb{N}$.

Remark 1: In Algorithm 1, the Q-tables of the BSs are

Algorithm 1 Proposed Beam Scheduling & Power Allocation: Execution Phase

1: **Input**: $P_q, I_q, N_b, \alpha, \beta, \gamma, \epsilon, l_r$ and $\{j_i\}_{i \in \mathcal{M}}$.

2: **Initialization**: Each BS_i initializes Q-table as $Q_i(a, s) = 1, \forall (a, s)$. Set $t = 1$.

3: *Step 1:* BS_i chooses TX power $p_{j_i,i}^{(t)}$ in slot t according to

$$p_{j_i,i}^{(t)} = \begin{cases} \text{randomly pick from } \mathcal{P}_q, & \text{w.p. } \epsilon \\ p_{\hat{a}}, \hat{a} = \arg \max_{a \in [P_q]} Q_i\big(a, s^{(t)}\big), & \text{w.p. } 1 - \epsilon \end{cases}$$

BS_i generates a beam towards UE_{j_i} if $p_{j_i,i}^{(t)} \neq 0$ and starts data transmission. Continue for one slot.

4: *Step 2:* Each BS_i updates its Q-table as: let $Q_i(a, s) \leftarrow Q_i(a, s)$, if $(a, s) \neq (a^{(t)}, s^{(t)})$; let $Q_i(a, s) \leftarrow (1 - l_r)Q_i(a, s) + l_r \left(r_i^{(t)} + \gamma \max_{a \in [P_q]} Q_i\big(a, s^{(t+1)}\big) \right)$, if $(a, s) = (a^{(t)}, s^{(t)})$.

5: *Step 3:* $t \leftarrow t + 1$. If $t \leq N_b$, go back to *Step 1*, else stop.

6: **Output**: Average reward of all BSs.

---

initialized with all one matrices, i.e., the initial value estimate are set to be $Q_i(a, s) = 1, \forall a, s$. This is termed as the principle of *being optimistic in the face of uncertainty* which is widely used in value-based RL applications.

Remark 2 (Complexity): For each BS, the storage complexity of the proposed algorithm is $\mathcal{O}\left(\frac{K P_q I_q}{M}\right)$ (supposing each BS is associated with the same number of UEs) since each BS has to store a Q-table of size $P_q \times I_q$ for each of its K/M associated UEs. In the execution phase, the implementation complexity per slot is $\mathcal{O}\left(\max\{P_q, I_q\}\right)$ which is due to the UE interference quantization ($\mathcal{O}(I_q)$) and greedy action selection ($\mathcal{O}(P_q)$). The Q-table update has complexity $\mathcal{O}(1)$. It can be seen that both the storage and implementation complexity scale linearly with the number of discrete powers and interference states, and the storage complexity also scales linearly with the number of UEs. This linear scaling is acceptable in general. Our experiments in Section IV show that the typical values of $P_q = 10, I_q = 20$ suffice to achieve the near-optimal (by letting P_q, I_q being arbitrarily large) performance for the considered network in the experiment with four BSs and twelve UEs in total.

IV. EVALUATION

A. Experiment Setup

We consider a cellular network (see Fig. 4) with four BSs each belonging to a different operator. Each BS is associated with three UEs located in its coverage area. Let $l = 20$ meters be the height of the BS antenna. UE antenna height is assumed to be zero. Therefore, the distance from BS_i to UE_j is equal to $d_{j,i} = \sqrt{l^2 + \bar{d}_{j,i}^2}$ where $\bar{d}_{j,i}$ is the planar distance between BS_i site and UE_j. The system has a shared bandwidth of $W = 400$ MHz with a center frequency at 37 GHz. Each BS is
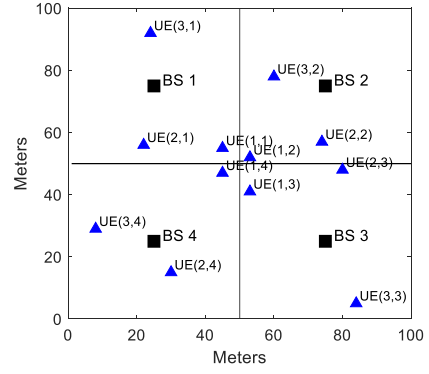

Fig. 4: Locations of the BSs and UEs on a 100×100 meter^2 planar grid. UE (j, i) represents the jth UE of BS_i.

subject to a peak power constraint $p_i^{\max} = 39$ dBm (7.94 Watt). Noise power is calculated according to

$$\sigma^2 \,(\text{dBm}) = 10 \lg(\kappa_B T_0 \times 10^3) + \text{NR}\,(\text{dB}) + 10 \lg W$$

where $\kappa_B = 1.38 \times 10^{-23}$ J/K is the Boltzmann's constant, NR is the UE noise figure and T_0 is the temperature. Taking the typical values of NR $= 1.5$ dB and $T_0 = 290$ K, the total noise power over the 400 MHz bandwidth is equal to $\sigma^2 = -86.46$ dBm. We examine beam scheduling and power allocation in one block with $N_b = 100$ slots. Each slot has a duration of one milli-second. The physical environment and learning parameters are listed as follows:

Parameter	Value
exploration rate ϵ	0.05
discount factor γ	0.9
learning rate l_r	0.1
$p_i^{\max}, \forall i \in \mathcal{M}$	7.94 Watt
noise power σ^2	-86.46 dBm
pass loss η	4
Nakagami fading Ω, μ	$100, 10^4$
block size N_b	100 slots
slot duration T_s	1 millisecond
antenna height l	20 meters

B. Baseline Scheme

Game-Theoretic (GT) Power Allocation: In [9], a non-cooperative game-based power allocation was proposed for distributed interference management in mmWave networks. In particular, each BS was treated as an independent *player* that selfishly attempts to maximize its own payoff, defined in the form of (4). A parallel power adaptation scheme was proposed based on the the concept of *best response* in game theory. In each slot, BS_i updates its power according to

$$p_{j_i,i}^{(t+1)} = \left[\frac{\alpha_i W}{\beta_i} - \frac{1}{g_{j_i,i}^{(t)}} \right]_0^{p_i^{\max}}, \qquad (10)$$

where $g_{j_i,i}^{(t)} \triangleq G_{j_i,i}^{\mathrm{BS}} G_{j_i,i}^{\mathrm{UE}} |h_{j_i,i}|^2 d_{j_i,i}^{-\eta} / (I_{j_i,i}^{(t)} + \sigma^2)$ is the equivalent channel gain between BS_i and UE_{j_i} in slot t. $g_{j_i,i}^{(t)}$ can be obtained by BS_i by letting UE_{j_i} measure the RX interference (plus noise) $I_{j_i,i}^{(t)} + \sigma^2$ and then sending back to BS_i. The Euclidean projection operator $[\cdot]_a^b$ is defined as $[x]_a^b = a$ if $x < a$, $[x]_a^b = b$ if $x > b$ and $[x]_a^b = x$ if $x \in [a,b]$. The above power adaptation is proved to converge to Nash equilibrium under certain conditions.

Drawback of the GT power allocation: The GT power allocation has a disadvantage due to its greedy nature: it may perform poorly in the high interference regime. This is because, for example, for the case of $\beta_i \approx 0$, each BS only aims to maximize its own throughput. The solution to GT is always choosing the maximum power to transmit, regardless of the interference. This may cause severe interference if the scheduled UEs are close to each other or there is beam overlapping (See Fig. 5), and thus dampening the
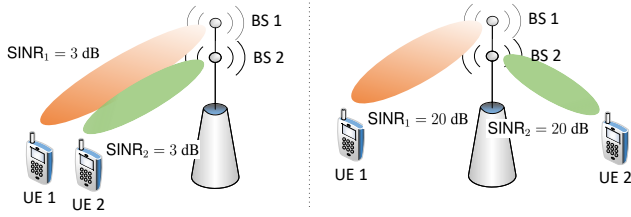

Fig. 5: BS colocation and and closely located UEs (left). There is strong interference due to beam overlapping. GT cannot distinguish between these two cases.

overall performance. However, our proposed Q-learning-based approach can overcome the above disadvantage by adapting to the physical environment (via observation and action-state value update) which is governed by the joint behaviors of all the agents. Each BS may make decisions other than maximum power based on the current interference state and its experience. For example, for the overlapping beam case, if all BSs are transmitting with high powers, being greedy by choosing a large TX power will emit a small reward as all UE are experiencing strong interference. By learning from the small reward, the proposed approach can shift to a lower power to explore new possibilities of higher reward. However, the GT allocation will be greedy and unable to adapt. Another drawback of the GT method is that it operates with continuous power which is infeasible in practice. However, quantization of TX power will inevitably incur performance loss by the adaptation rule of (10). In Section IV-C, we verify the effect of multiple factors that affect the performance of the proposed approach and show that the performance can be significantly enhanced over GT.

C. Results

We compare the proposed approach with the GT power allocation and verify the effect of the reward weights α, β, the number of power levels and interference states and the BS/UE antenna gain and beamwidth. Throughout the experiment, we assume that all UEs have omnidirectional antennas.[4] We fix $\alpha = 1$ for all BSs and and let $\beta = 0$ and $\beta = 0.1W = 4 \times 10^7$ to verify its effect.

1) Effect of P_q and I_q: The BS antenna MSR and beamwidth are chosen to be 20 dB and 30° respectively. The 1st UE of each BS is scheduled. This UE selection represents the behavior of the cell-edge UEs which usually suffer from strong interference from neighboring BSs. This phenomenon is even more prominent in ultra-dense small BS 5G cellular networks. To verify the effect of P_q, we fix $I_q = 10$ and let $P_q \in \{10, 20, 40\}$. Figs. 6a and 6c show the effect of P_q for $\beta = 0$ and $0.1W$ respectively. Each curve represents the average reward achieved up to the current slot, averaged over 50 independent trials each containing a set of i.i.d. channel realizations. For both values of β, it can be seen that the proposed approach outperforms GT. For $\beta = 0$, the proposed approach achieves 23% to 39% more average reward than GT in the 100th slot. For $\beta = 0.1W$, the proposed approach achieves 63% to 87% more average reward than GT. Moreover, the average reward increases as P_q increases because larger P_q provides more choices for power selection. To verify the effect of I_q, we fix $P_q = 10$ and let $I_q \in \{2, 4, 8, 16\}$. Figs. 6b and 6d show the result. For both $\beta = 0$ and $0.1W$, the achieved average reward of the proposed approach increases as I_q increases. For $\beta = 0$, when $I_q = 2$, the proposed approach achieves a similar performance to GT. However, when $I_q = 16$, there is a 33% reward gain compared to GT. For $\beta = 0.1W$, the proposed approach achieves 24% to 80% more reward than GT from $I_q = 2$ to $I_q = 16$. The effect of I_q is expected because when there are more interference states for each agent, the decision making of each agent becomes more flexible and can cater to the specific interference condition according the agent's past experience.

2) Effect of beamwidth and MSR: The effect of beamwidth and MSR are shown in Fig. 7 and Fig. 8. We fix $\beta = 0.1W$. In Fig. 7, the first UE of each BS is scheduled. These UEs represent the cell-edge UEs. We compare the performance of the proposed approach with GT under the BS antenna configurations $(20\,\mathrm{dB}, 30°)$, $(30\,\mathrm{dB}, 20°)$ and $(40\,\mathrm{dB}, 10°)$. For the first two cases with BS beamwidth 30° and 20°, the proposed approach achieves 87% and 134% more reward than GT. GT performs poorly in these cases by being greedy to choose the maximum power because there is beam overlapping which causes very strong interference to the non-target UEs due to high TX powers. This implies that the proposed approach has much better performance than GT in the interference-limited regime. However, when the beamwidth is further reduced to 10°, the proposed approach achieves a similar reward to GT. This is because in this case, BS beams are very sharp so they cause little interference to non-target UEs. It has been shown by [9] that when the interference level is very low, GT achieves near-optimal performance. Therefore, the proposed approach also achieves near-optimal performance in
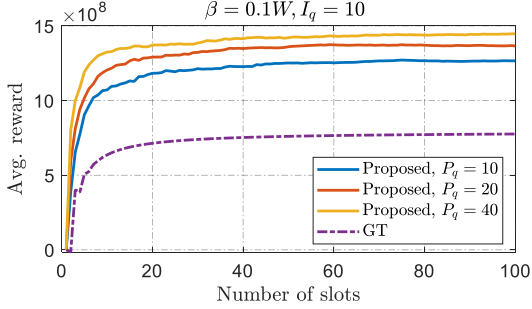
---

[4]Since varying the UE antenna MSR and beamwidth has a similar effect to that of the BS antenna, we use omnidirectional UEs in the experiment.
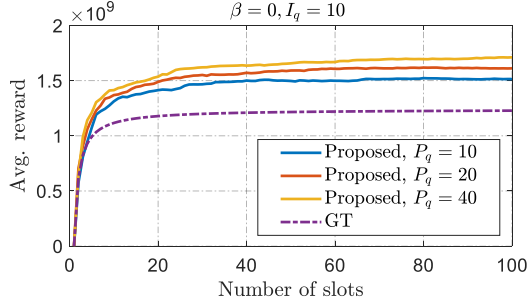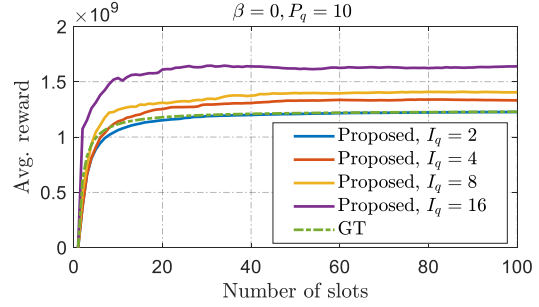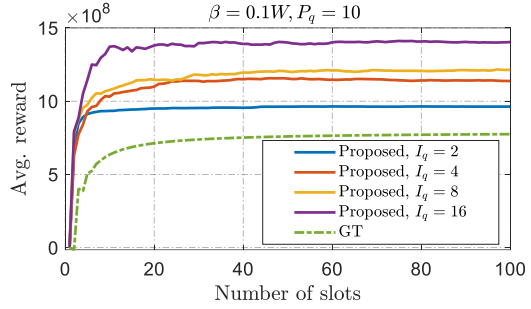
Fig. 6: Effect of P_q and I_q for different β. BSs have MSR of 20 dB and beamwidth $30°$, UEs are omnidirectional.
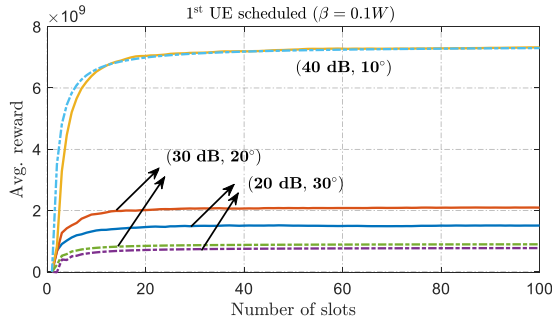
this case. Fig. 8 shows the case when the third UE of each

Fig. 7: Proposed approach (solid lines) vs. game-based approach (dash lines) when the 1st UE of each BS is scheduled.

Fig. 8: Proposed approach (solid lines) vs. game-based approach (dash lines) when the 3rd UE of each BS is scheduled.

BS is scheduled. Due to their separate locations, these UEs receive less interference and represent the cell-center UEs, which usually have high SINR. It can be seen that for any of the considered BS antenna configurations, the proposed approach outperforms GT by a small margin, and the margin diminishes as the beams become sharper (see the extreme case (40 dB, $10°$)). The reason for this competitive performance is that the interference level is relatively low because the scheduled UEs are sparsely distributed. This demonstrates that the proposed approach is at least as good as GT in the high SINR regime.
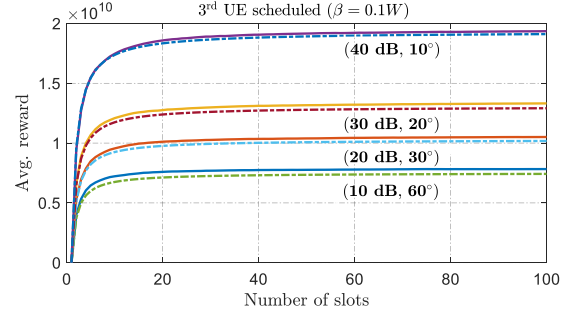
V. EXTENSIONS

A. Incorporation of the Lyapunov Optimization Framework

One interesting aspect of the proposed approach is that the weights α, β can be automatically determined if we apply the Lyapunov optimization framework on top of the proposed power allocation algorithm. More specifically, let us consider the following *utility maximization* problem

$$\max \quad \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{K}_i} U(\bar{X}_{j,i}) \tag{11a}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{K}_i} \bar{p}_{j,i} \leq T_f p_i^{\text{avg}}, \quad \forall i, \tag{11b}$$

$$p_{j_i,i}(k,n) \leq p_i^{\max}, \quad \forall i, k, n, \tag{11c}$$

where $p_{j_i,i}(k,n)$ is the TX power of BS_i in the nth block of the kth frame. Each BS_i is subject to a long-term average and an instantaneous peak power constraint p_i^{avg} and p_i^{max} respectively. $\bar{p}_{j,i}$ represents the average power consumption of BS i to UE j in all frames. $\bar{X}_{j,i}$ denotes the average number of received bits by UE_j in each frame and is referred to as the average throughput in the following. U represents the utility function, e.g., fairness function which is concave and non-decreasing and thus ensures fairness among the BSs. Using the Lyapunov stochastic optimization framework [22], the above problem can be decomposed into two sub-problems to be solved in each frame, together with two virtual queues to enforce the average constraints. In particular, the first sub-problem aims to solve the auxiliary variables $\gamma_{j,i}(k)$:

$$\max \quad \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{K}_i} VU(\gamma_{j,i}(k)) - H_{j,i}(k)\gamma_{j,i}(k) \quad \text{(12a)}$$

$$\text{s.t.} \quad 0 \leq \gamma_{j,i}(k) \leq T_f W \log\left(1 + g_{j,i}^{\text{max}}(k)p_i^{\text{max}}\right), \ \forall i,j,k \quad \text{(12b)}$$

where V is a constant. $g_{j,i}^{\text{max}}(k) \triangleq \max_n g_{j,i}(k,n)$ denotes the maximum equivalent channel gain in the kth frame. $H_{j,i}(k)$ is the UE throughput queue which is updated by

$$H_{j,i}(k+1) = \max\left\{H_{j,i}(k) + \gamma_{j,i}(k) - X_{j,i}(k),\, 0\right\},$$
$$\forall i \in \mathcal{M}, \forall j \in \mathcal{K}_i. \quad \text{(13)}$$

The second sub-problem aims to solve the TX powers $p_{j,i}(k,n)$:

$$\min \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{K}_i} \left(\sum_{n \in [N_f]} \mathbb{E}\left[T_{j,i}^d(k,n)p_{j,i}(k,n)\right] - T_f p_i^{\text{avg}} \right)$$
$$\times Z_i(k) - H_{j,i}(k)\widehat{X}_{j,i}(k) \quad \text{(14a)}$$

$$\text{s.t.} \quad 0 \leq p_{j,i}(k,n) \leq p_i^{\text{max}}, \quad \forall i,k,n \quad \text{(14b)}$$

where

$$\widehat{X}_{j,i}(k) \triangleq \sum_{n=1}^{N_f} \mathbb{E}\left[T_{j,i}^d(k,n)W \log\left(1 + \text{SINR}_{j,i}(k,n)\right)\right]$$

denotes the expected throughput of UE_j in the kth frame. $T_{j,i}^d(k,n)$ denotes the data transmission time for UE j by BS i during block n of frame k. $Z_i(k)$ is the TX power queue which is updated by

$$Z_i(k+1) =$$
$$\max\left\{Z_i(k) + \sum_{j \in \mathcal{K}_i} \sum_{n \in [N_f]} T_{j,i}^d(k,n)p_{j,i}(k,n) - T_f p_i^{\text{avg}},\, 0\right\},$$
$$\forall i \in \mathcal{M}. \quad \text{(15)}$$

Note that the objective of (14a) has the same form as the payoff function (4) if we choose $\alpha_i = H_{j,i}(k)N_b$, $\beta_i = Z_i(k)N_b$. More specifically, given that UE_{j_i} is scheduled, each BS_i has an objective function $H_{j_i,i}(k)\widehat{X}_{j_i,i}(k) - Z_i(k)\mathbb{E}\left[T_{j_i,i}^d(k,n)p_{j_i,i}(k,n)\right]$ (the constant term $T_f p_i^{\text{avg}}$ is omitted as it does not affect the optimal solution) to maximize

in block n, where $\widehat{X}_{j,i_j}(k,n)$ is UE_{j_i}'s throughput in block n. By letting $\mathbb{E}[T_{j_i,i}^d] = T_b$, i.e., the scheduled UE will be receiving data during the entire block, the objective becomes $\alpha_i T_s W \log(1 + \text{SINR}_{j_i,i}(k,n)) - \beta_i T_s p_{j_i,i}(k,n)$. This objective can be optimized by maximizing the sum or average throughput in the N_b slots in block n. In this way, the proposed approach can be used to (approximately) solve the second sub-problem (14) in each block in a *distributed* manner. It can be seen that the reward weights α_i, β_i are optimally determined by the virtual queues derived from the Lyapunov optimization framework. In [9], the GT method (10) was used to solve the second sub-problem. Since we have shown that the proposed approach outperforms GT in a single block, it is expected to also achieve higher utility than GT when the Lyapunov framework is applied. Fig. 9 shows the achieved utility when the utility function $U(x) = x^{3/5}$ is used and under the same experiment setup as in Section IV. BS beamwidth and MSR are chosen as $30°$ and 20 dB while the UEs are ominidirectional. It can be seen that the proposed approach achieves 29%
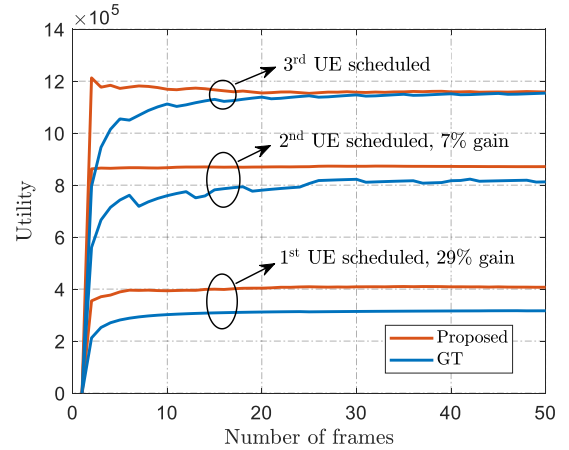

Fig. 9: Proposed approach (red) vs. game-based approach (blue) when the Lyapunov framework is applied.

more utility (at the 50th frame) than GT when the first UE of each BS is scheduled and 7% more when the second UE is scheduled. For the cell-center UEs, i.e., the third UE of each BS, the proposed approach achieves a similar utility as GT but with a faster convergence. The queue values of BS_1 when the first UE is scheduled is shown in Table I. It can be seen that $\beta_1/\alpha_1 = Z_1(k)/H_{1,1}(k) \approx 0, \forall k$. This mimics the behavior of the proposed power allocation algorithm when there is a very small penalty on power consumption.

TABLE I: Virtual queue values corresponding to BS_1.

Frame index k	10	20	30	40	50
$Z_1(k)$	0	0.24	0	0	0
$H_{1,1}(k)/10^9$	3.87	0.14	3.92	1.90	0.11

B. Practical Consideration

As discussed in Section III-C, the proposed approach adopts a per-BS storage complexity of $\mathcal{O}\left(\frac{KP_qI_q}{M}\right)$ and a per-BS per-slot execution complexity of $\mathcal{O}\left(\max\{P_q, I_q\}\right)$. The storage complexity scales linearly with the number of UEs per BS and the execution complexity does not depend on the number of UEs. This demonstrates the scalability of the proposed approach. However, to implement it on real-world cellular networks, there are still several practical considerations. First, in the proposed approach, the interference at the scheduled UE needs to be measured in each slot and then reported back to the associated BS. The measurements that can be configured at the UE for the relevant cellular systems, 5G and Beyond, has to be analyzed for an implementation. Second, it is assumed in the proposed approach that the channels are block-fading and do not change within the duration of each scheduling block. Verification with real world cellular conditions will be appropriate for an implementation of this solution.

VI. Conclusion

In this work, we studied the problem of distributed beam scheduling and power allocation for non-cooperative mmWave networks. We presented a unified framework, with a flexible network payoff function definition, that can be used for systematic performance evaluation and comparison of different scheduling methods. Furthermore, we proposed a Q-learning-based approach using an independent agent modeling where each BS can adaptively control its transmit power for different interference conditions based on its experience and active exploration of non-greedy actions. Experiments showed that the proposed approach outperforms the non-cooperative game-based approach in the sense that they achieve similar performance in the high SINR regime but the proposed approach beats the game-based approach by a large margin in the interference-limited regime. In addition, the proposed approach can be integrated into the Lyapunonv stochastic optimization framework for the purpose of network utility maximization. In this case, the weights in the reward function are automatically and optimally determined by the virtual queues.

Acknowledgement

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] T. Alpcan, T. Basar, R. Srikant, and E. Altman, "Cdma uplink power control as a noncooperative game," in *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)*, vol. 1, 2001, pp. 197–202 vol.1.

[5] J.-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei, "Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3251–3271, 2010.

[6] J. Ning, L. Feng, F. Zhou, M. Yin, P. Yu, W. Li, and X. Qiu, "Interference control based on stackelberg game for d2d underlaying 5g mmwave small cell networks," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[7] S. Sarkar, X. Zhang, A. Bhuyan, M. Ji, and S. K. Kasera, "Uncoordinated spectrum sharing in millimeter wave networks using carrier sensing," *arXiv preprint arXiv:2102.12138*, 2021.

[8] U. O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo, "Near-optimal power control in wireless networks: A potential game approach," in *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–9.

[9] X. Zhang, S. Sarkar, A. Bhuyan, S. K. Kasera, and M. Ji, "A non-cooperative game-based distributed beam scheduling framework for 5G millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.

[10] A. Galindo-Serrano and L. Giupponi, "Distributed q-learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823–1834, 2010.

[11] O. Van Den Biggelaar, J.-M. Dricot, P. De Doncker, and F. Horlin, "Power allocation in cognitive radio networks using distributed machine learning," in *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications-(PIMRC)*. IEEE, 2012, pp. 826–831.

[12] S. Kar, J. M. Moura, and H. V. Poor, "Distributed reinforcement learning in multi-agent networks," in *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2013, pp. 296–299.

[13] E. Ghadimi, F. D. Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–7.

[14] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep q learning approach," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[15] 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.

[16] 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.

[17] 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.

[18] R. Ramanathan, "On the performance of ad hoc networks with beamforming antennas," in *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, 2001, pp. 95–105.

[19] J. C. S. S. Filho and M. D. Yacoub, "Nakagami-m approximation to the sum of m non-identical independent nakagami-m variates," *Electronics Letters*, vol. 40, no. 15, pp. 951–952, 2004.

[20] J. C. Santamaria, R. S. Sutton, and A. Ram, "Experiments with reinforcement learning in problems with continuous state and action spaces," *Adaptive behavior*, vol. 6, no. 2, pp. 163–217, 1997.

[21] H.-M. Gross, V. Stephan, and M. Krabbes, "A neural field approach to topological reinforcement learning in continuous action spaces," in *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, vol. 3. IEEE, 1998, pp. 1992–1997.

[22] M. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.