2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN) | 979-8-3503-4319-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLCN59089.2024.10624763

Joint Power Control and Beamforming for Interference Mitigation in Multi-User Millimeter-Wave Systems

Madan Dahal†, Mojtaba Vaezi† and Wonjae Shin‡
†Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, USA
‡School of Electrical Engineering, Korea University, South Korea
Emails: {mdahal, mvaezi}@villanova.edu, wjshin@korea.ac.kr

***Abstract*—Interference poses a significant challenge in improving cell throughput in multi-cell multi-user networks. Coordinated beamforming and power control have shown promise in mitigating interference and maximizing cell throughput. However, existing techniques often suffer from computational complexity and the overhead of collecting channel state information (CSI) from interfering cells. In this paper, we propose a deep reinforcement learning based approach that addresses these limitations by eliminating the need for explicit CSI knowledge. Through extensive simulations in millimeter-wave networks with various cell configurations, we demonstrate the effectiveness of our interference management technique. The simulation results showcase the ability of our method to learn near-optimal power and beamforming strategies for multi-cell multi-user networks, all without the need for explicit CSI information.**

I. INTRODUCTION

Inter-cell interference is the main barrier to achieving high throughputs and spectral efficiency in today's networks. It also causes *outages* at the cell edges [1]. Interference management is a long-standing problem and has been extensively studied in the literature [1]–[3]. *Interference alignment* [2], [4] is a theoretical breakthrough that offers greater efficiency than time-division multiple access (TDMA). However, its practical implementation is limited due to the requirement of *global* channel state information (CSI), which is not feasible in practical wireless systems. Coordinated multi-point (CoMP) [5] is another well-known solution to addressing inter-cell interference by enabling neighboring base stations (BSs) to share data and CSI for coordinated downlink transmissions and joint processing of received signals in the uplink. However, CoMP relies on a high-speed backhaul network for efficient information exchange between BSs [5].

The lack of knowledge about the optimal strategy for multi-cell networks has led to a pragmatic approach of treating interference as noise. These approaches rely on the *signal-to-interference-plus-noise ratio* (SINR) for network performance, but they require knowledge of the interfering cells' CSI, which is not practical. Multi-cell coordination, e.g., beamforming and transmit power coordination, has been identified as an effective approach to mitigate inter-cell interference and enhance cell throughput in multi-cell networks. However, many existing methods for multi-cell coordination suffer from limitations such as high computational complexity and the need for collecting global CSI, which are impractical in dynamic wireless environments [4], [5]. In [6], [7] joint power control and beamforming are considered but they focused on time-invariant channels with fixed routes.

To address these limitations and explore the potential for discovering better solutions, researchers have increasingly embraced the power of deep learning techniques, specifically *deep reinforcement learning* (DRL) [8]. DRL has been successful in solving various communication problems in different scenarios, including beamforming, power allocation, and interference cancellation [9]–[11]. In reinforcement learning [12] an agent learns to interact with an environment by taking a sequence of actions to maximize a cumulative reward, e.g., the spectral efficiency or any other desired quantity. Existing research, e.g., [10], mostly focus on single-user scenarios, which may not fully capture the dynamics of real-world cellular networks where multiple users are in the same cell, leading to significant multi-user interference.

In this paper, we propose a deep Q-network (DQN)-based algorithm to address inter-cell interference in a multi-cell multi-user network by jointly optimizing power and beamforming. While industry standards [13] typically require the user equipment (UE) to report its CSI, which can be a vector or matrix depending on the number of antenna elements, we reduce the reporting overhead by utilizing the UE's position instead to reduce the reporting overhead associated with CSI and enhance overall cell throughput and coverage [14], [15]. Our goal is to maximize the spectral efficiency of the network, evaluated in terms of achievable sum-rate, by jointly optimizing the transmit power and beamforming vectors at all BSs to maximize the UEs received SINR. Simulation results demonstrate a significant increase in spectral efficiency with the spectral efficiency scaling almost linearly with the number of cells with no explicit CSI.

II. CHANNEL AND SYSTEM MODEL

We consider a downlink cellular network with L ($L \geq 2$) cells. Each cell consists of a BS adopting M antennas in a

M. Dahal and M. Vaezi's work was supported by the U. S. National Science Foundation under Grant CNS-2239524. W. Shin's work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. 2021-0-00467).

uniform linear array (ULA) and is assumed to serve multiple single-antenna users. UEs are randomly and uniformly distributed in the cell serving area. Each BS simultaneously serves U UEs where $U \geq 1$, and each UE can only be served by one BS at a time. The operating frequency band is within the millimeter-wave (mmWave) range, and the channels may be multi-path or line-of-sight.

A. Channel Model

Since mmWave has a short wavelength, it allows for the deployment of BSs equipped with a large number of antennas. This technology is regarded as a main solution to the spectrum shortage caused by the rising bandwidth demand in wireless networks [16]. However, high attenuation in mmWave results in reduced inter-site distances, increasing the possibility of inter-cell interference in these networks.

The mmWave channel can be described with standard multipath models of lower mmWave frequency [17]. Adopting this channel model, the channel from the jth BS to the uth user served by ℓth BS can be written as

$$\mathbf{h}_{\ell,j,u} = \frac{\sqrt{M}}{\rho_{\ell,j,u}} \sum_{n=1}^{N_{\mathrm{p}}} \alpha_{\ell,j,u,n} \mathbf{a}^*(\theta_{\ell,j,u,n}), \tag{1}$$

where $\rho_{\ell,j,u}$ is the path loss from the jth BS to the uth user, N_{p} is the number of paths between the transmitter and receiver, and $\alpha_{\ell,j,u,n}$ and $\theta_{\ell,j,u,n}$ are the complex gain and the angle of departure (AoD), respectively. The steering vector $\mathbf{a}^*(\theta_{\ell,j,u,n})$ depends on the angular directions of the departing plane wave, and for an M-element uniform linear array is given by $\mathbf{a}(\theta) = \left[1, e^{-j2\pi\vartheta}, e^{-j4\pi\vartheta}, \ldots, e^{-j2\pi\vartheta(M-1)}\right]^T$. Here $\vartheta \triangleq \frac{d}{\lambda}\cos(\theta)$ is the normalized spatial angle which is related to the physical angle of departure $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and d and λ, respectively, are the antenna spacing and the wavelength of operation [17].

B. System Model

In mmWave systems, M is often large to account for considerable high-frequency path loss and ensure that the received signal is sufficiently powerful. Moreover, M RF links would be required for a fully digital beamforming, which would be expensive and use considerable amount of power at high frequencies. Given this, analog-only beamforming is popular in mmWave as it only needs one RF chain. Because of this, it is presumed that BSs have an analog-only beamforming design in which the beamforming is implemented with analog phase shifters. Due to hardware limitations on large-scale multiple-antenna systems, the BSs often use pre-defined beamforming codebooks [18] that scan all potential directions for data transmission. To simplify, the weights of each beamforming vector are implemented using constant-modulus phase shifters. Beamforming vectors are selected from the codebook whose each element is given by

$$\mathbf{w} = \frac{1}{\sqrt{M}}\left[e^{j\theta_1}, \ldots, e^{j\theta_M}\right]^T, \tag{2}$$

where the phase shift $\theta_m, m = \{1, 2, \ldots, M\}$, is selected from a finite set $\mathbf{\Phi}$ with 2^r possible discrete values. That is $\mathbf{\Phi} = \left[0, \frac{\pi}{M}, \frac{2\pi}{M}, \ldots, \frac{(M-1)\pi}{M}\right]$ for r-bit quantized phase shifters, uniformly drawn from $[0, \pi]$.

In the downlink transmission, if the transmitted symbol from the ℓth BS to user u is $x_{\ell,u}$, the received signal at a given UE at cell ℓ can be expressed as

$$y_{\ell,u} = \mathbf{h}_{\ell,\ell,u}^H \mathbf{w}_{\ell,u} x_{\ell,u} + \sum_{k \neq u} \mathbf{h}_{\ell,\ell,u}^H \mathbf{w}_{\ell,k} x_{\ell,k} + \sum_{j \neq \ell} \sum_{u=1}^{U} \mathbf{h}_{\ell,j,u}^H \mathbf{w}_{j,u} x_{j,u} + n_{\ell,u}, \tag{3}$$

where $\mathbf{h}_{\ell,j,u} \in \mathbb{C}^{M \times 1}$, $\forall \ell, j \in \{1, \ldots, L\}$, is the channel vector from BS j to the user u in cell ℓ which is described in (1), $\mathbf{w}_{j,u} \in \mathbb{C}^{M \times 1}$ is the beamforming vector for user u at BS j as described in (2), and $n_{\ell,u} \in \mathcal{CN}(0, \sigma^2)$ is the noise at the UE u. Furthermore, $x_{j,u}$ is subject to the average power constraint $\mathbb{E}||x_{j,u}|| = P_{j,u}$, where $P_{j,u}$ is the power of BS j alloted to user u. Then, the SINR at the UE u located at cell ℓ is given by

$$\gamma_{\ell,u} = \frac{P_{\ell,u} |\mathbf{h}_{\ell,\ell,u}^H \mathbf{w}_{\ell,u}|^2}{\sigma^2 + \sum_{k \neq u} P_{\ell,k} |\mathbf{h}_{\ell,\ell,u}^H \mathbf{w}_{\ell,k}|^2 + \sum_{j \neq \ell} \sum_{u=1}^{U} P_{j,u} |\mathbf{h}_{\ell,j,u}^H \mathbf{w}_{j,u}|^2}. \tag{4}$$

III. Problem Formulation

Sum achievable rate, or simply sum-rate, is a common measure of spectral efficiency in cellular networks. Considering this, in this paper our goal is to maximize the network sum-rate which is defined as $\sum_{\ell=1}^{L} \sum_{u=1}^{U} \log_2(1 + \gamma_{\ell,u})$, and is equivalent to $\log_2 \prod \prod (1 + \gamma_{\ell,u})$. Since the logarithm is a monotonic function, to find the arguments that maximize the sum-rate we can solve

$$\max_{P_{\ell,u}, \mathbf{w}_{\ell,u}} \quad \prod_{\ell=1}^{L} \prod_{u=1}^{U} (1 + \gamma_{\ell,u}) \tag{5}$$

$$\text{subject to} \quad P_{\ell,u} \in \mathcal{P}, \quad \forall \ell, \forall u, \tag{6}$$

$$\mathbf{w}_{\ell,u} \in \mathcal{W}, \quad \forall \ell, \forall u, \tag{7}$$

$$\sum_{u} P_{\ell,u} \leq P_{\ell}^{\max}, \quad \forall \ell, \forall u, \tag{8}$$

$$\gamma_{\ell,u} \geq \gamma_{\min}, \tag{9}$$

in which $\mathcal{P}$ is the possible transmit powers, $\mathcal{W}$ is *beamforming codebook* from which $\mathbf{w}_{\ell,u}$ is selected; $P_{\ell}^{\max}$ is a maximum power for any ℓth BS and $\gamma_{\min}$ denotes the the minimum SINR for any user in the cellular network.

Due to the constant modulus constraints, the above optimization problem is non-convex and challenging to solve. It can be solved by exhaustively searching over the space of the Cartesian product of $\mathcal{P} \times \mathcal{W}$. The complexity of such a solution will be $|\mathcal{P}|^{UL}|\mathcal{W}|^{UL}$, which is very high when L and U are large. In the following, we develop an alternative solution based on DRL.

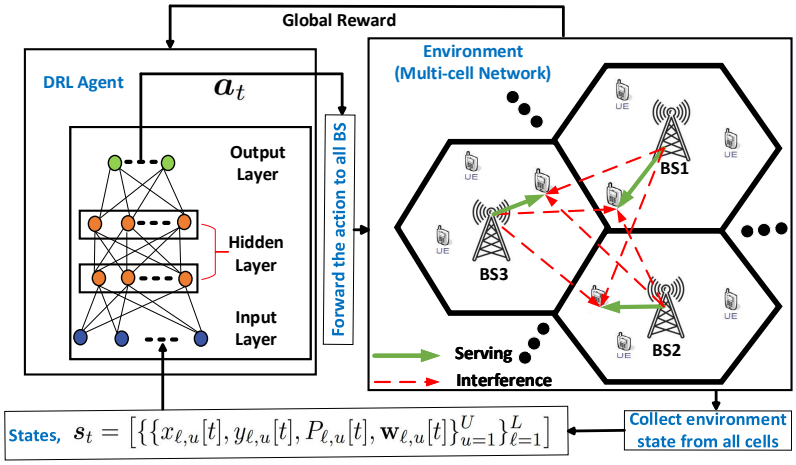


Fig. 1: The downlink multi-cell multi-user network for the proposed DQN algorithm where each BS is communicating with its users in the presence of the interfering BSs.

IV. Preliminaries of Reinforcement Learning

Reinforcement learning (RL) is a machine learning approach that involves learning through trial and error and by rewarding desired behaviors. In our problem, the goal of RL is to make decisions about the beamforming vectors and base station powers in equation (5) to optimize the system performance. RL is based on the concepts of agent, environment, state, action, and reward, where the agent learns to make optimal decisions in an environment to maximize the cumulative reward. The DRL agent would be a central node such as a central radio resource management controller in 4G/5G.

The state $\boldsymbol{s}_t \in \mathcal{S}$ is what an agent observes at a time step t, and is a representation of the environment. It consists of x and y coordinates of the UE u of cell ℓ as $x_{\ell,u}[t]$ and $y_{\ell,u}[t]$, the transmit power of the BS of cell ℓ to UE u as $P_{\ell,u}[t]$, and the beamforming vector of the BS of cell ℓ to UE u as $\mathbf{w}_{\ell,u}[t]$. The observed states by the agent at time t, is given by

$$\boldsymbol{s}_t = \left[\{\{x_{\ell,u}[t], y_{\ell,u}[t], P_{\ell,u}[t], \mathbf{w}_{\ell,u}[t]\}_{u=1}^{U}\}_{\ell=1}^{L}\right]. \tag{10}$$

To address challenges related to real-time UE locations and their accessibility, several approaches like privacy-preserving techniques, exploring non-location-based solutions, utilizing indoor positioning techniques can be considered [14], [19]. Our algorithms need UEs' locations but the location should not necessarily be exact. The agent at time step t produces action $\boldsymbol{a}_t$ that will result in state $\boldsymbol{s}_{t+1}$. In our problem, actions are to change the power and beamforming vector of each BS. The interference coordination and power control for the uth UE served by the ℓth BS at time step t is given by

$$P_{\ell,u}[t] := P_{\ell,u}[t-1] + PC_{\ell,u}[t], \tag{11}$$

in which $PC_{\ell,u}[t]$ is the power control command for user u at BS ℓ which is $+1$dB or -1dB depending on the action related to that command. If $\sum_{u=1}^{U} P_{\ell,u}[t] > P_{\ell}^{\max}$, then $PC_{\ell,u}[t]$ will be pushed to -1dB to obey the total power limit.

Action $\boldsymbol{a}_t \in \mathcal{A}$ has the following form

$$\boldsymbol{a}_t = \{\underbrace{a_{u,1}, \ldots, a_{u,L}}_{\text{power control}}, \underbrace{a_{u,L+1}, \ldots, a_{u,2L}}_{\text{beamforming}}\}_{u=1}^{U}, \tag{12}$$

Each element of the above action is either '0'or '1', Specifically, for any u and ℓ we have

- $a_{u,\ell} = 0$: decrease the transmit power of user u in ℓth BS by 1 dB.
- $a_{u,\ell} = 1$: increase the transmit power of user u in ℓth BS by 1 dB.
- $a_{u,L+\ell} = 0$: step down the beamforming codebook index of user u in ℓth BS.
- $a_{u,L+\ell} = 1$: step up the beamforming codebook index of user u in ℓth BS.

It is seen that, by taking action $\boldsymbol{a}_t$, the agent is changing the beamforming vectors as well as transmit power for each user in the serving and all interfering cells. Thus, this is a collaborative interference management algorithm via coordinated power and beamforming design.

The *state-action value function* $Q_\pi(\mathbf{s}_t, \boldsymbol{a}_t)$ describes the expected reward after taking one specific action following the policy π. More accurately,

$$Q_\pi(\mathbf{s}_t, \boldsymbol{a}_t) = \mathbb{E}\left[r_{t+1} + \alpha Q_\pi(\mathbf{s}_{t+1}, \boldsymbol{a}_{t+1}) | \mathbf{s}_t, \boldsymbol{a}_t\right]. \tag{13}$$

This is also known as the Bellman equation, in which α is a discount factor whose range is $[0, 1]$, $\mathbf{s}_{t+1}$ and $\boldsymbol{a}_{t+1}$ are the new state and action, respectively, r_{t+1} is the reward achieved when moving to the new state. The agent and the environment interact in discrete time steps, and the agent will act based on the state it will receive according to the π policy. The environment will react to the new state and provide a reward as feedback.

The *reward* is an incentive mechanism that tells the agent the consequence of an action. The agent's final objective is to maximize the total cumulative reward. Defining the reward function is a crucial step in evaluating the performance. Since the agent is seeking to increase its reward and our ultimate goal is to maximize sum-rate of the multi-cell multi-user network, an immediate definition of reward would be

$$r_{t+1} = \prod_{\ell=1}^{L} \prod_{u=1}^{U} (1 + \gamma_{\ell,u}), \tag{14}$$

where $\gamma_{\ell,u}$ is SINR received by the UE u at cell ℓ when action $\boldsymbol{a}_t$ is taken by the agent. This resembles the objective of the optimization problem in (5). This tells the immediate effect of taking action $\boldsymbol{a}_t$ in state $\mathbf{s}_t$ at time step t.

V. DRL-based Downlink Interference Control

In this paper, we use a value-based DRL which learns the state or state-action value. The agent acts by choosing the best action in the state. The value function $Q_\pi(\mathbf{s}_t, \boldsymbol{a}_t)$ is obtained using a neural network and is optimized by using a replay buffer (denoted by R).

Let $\boldsymbol{\Theta}_t \in \mathbb{R}^{u \times v}$ represent the weights of neural networks at time steps t, where u is the number of hidden nodes and v is the number of layers. We define $\boldsymbol{\theta}_t \triangleq \text{vec}(\boldsymbol{\Theta}_t) \in \mathbb{R}^{uv}$ and use this as a function approximator. DQN with the initial weight $\boldsymbol{\theta}$ is adjusted at every time step t to reduce the error via the mean-squared error loss function $L_t(\boldsymbol{\theta}_t)$

$$\min_{\boldsymbol{\theta}_t} L_t(\boldsymbol{\theta}_t) \triangleq \mathbb{E}_{\mathbf{s}_t, \boldsymbol{a}_t}\left[(y_t - Q_\pi(\mathbf{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta}_t))^2\right], \tag{15}$$

in which

$$y_t := \mathbb{E}_{\mathbf{s}_t, \boldsymbol{a}_t}\left[r_{t+1} + \alpha \max_{\boldsymbol{a}_{t+1}} Q_\pi(\mathbf{s}_{t+1}, \boldsymbol{a}_{t+1}; \boldsymbol{\theta}_{t-1} | \mathbf{s}_t, \boldsymbol{a}_t)\right]$$

is the estimated function value at time step t given state $\mathbf{s}_t$ and have an action $\boldsymbol{a}_t$. The algorithm tries to reduce this loss in every iteration. The objective of the DQN algorithm is to find a solution that optimizes the state-action value function. The episode (E) is a time frame within which the agent interacts with the environment. Each episode has T time steps.

The algorithm has two phases: *training and testing* phases. During the training phase, the agent is trained offline before it becomes active in the network. In this phase, the weights of the neural network is optimized using the stochastic gradient descent algorithm on the batches of the dataset taken from the replay buffer R. Having a replay buffer allows the agent to use a more diverse mini-batch for performing updates during the training process. It also allows the agent to take larger mini-batch sizes B. Further, by sampling at random from the replay buffer, the updates to the neural network will have low variance since the data entering the optimization method look independent and identically distributed.

At each round of the training process, the agent strikes a balance between exploring the environment and exploiting the knowledge of the best action accumulated through such exploration. We adopt an ϵ-greedy policy [12], where $\epsilon := \max(\epsilon\delta, \epsilon_{\min})$ is the exploration rate, δ is the exploration decay rate, and $\epsilon_{\min}$ is the minimum exploration rate. The exploration rate decays in every episode until it reaches $\epsilon_{\min}$. We exploit if $p > \epsilon$ where p is randomly drawn from Unif$(0, 1)$; we explore otherwise. Mathematically,

$$\mathbf{a}_t = \begin{cases} arg\max_{\boldsymbol{a}_{t+1}} Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_{t+1}; \boldsymbol{\theta}_t), & p > \epsilon, \\ \text{randomly chosen from } \mathcal{A}, & p \leq \epsilon. \end{cases} \tag{16}$$

Based on the selected action $\boldsymbol{a}_t$, the agent computes its reward function according to (14).

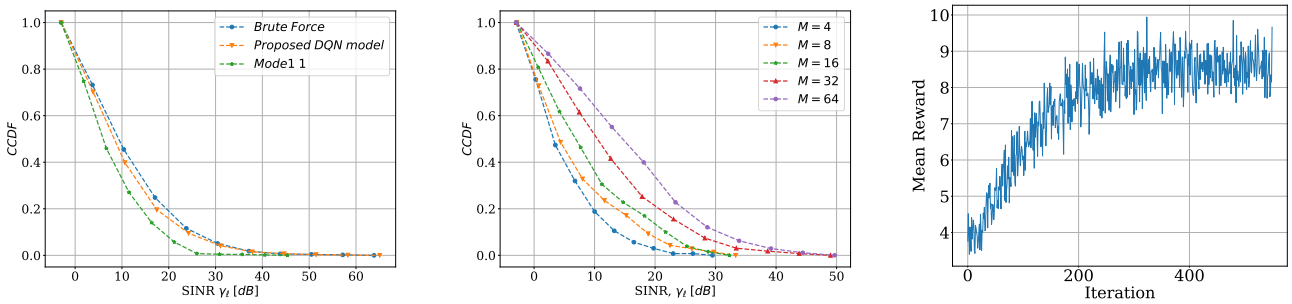
A summary of the training phase is given in Algorithm 1.

Algorithm 1 Training phase of proposed DQN algorithm

```
Input: θ, α, B
Output: Optimized θ
Randomly initialize network Q(s_t, a_t|θ) with weight θ
Initialize time, states, actions, minibatch size B and R
for episode 1 to E do
    Receive initial observation state s_t
    for t=1 to T do
        Select an action based on (16)
        Calculate the reward based on (14)
        Observe the next state s_{t+1}
        Store transition (s_t, a_t, r_{t+1}, s_{t+1}) in R
        Set b = min(B, T(E − 1) + t)
        Sample a random minibatch of size b transitions
        (s_b, a_b, r_{b+1}, s_{b+1}) from R
        Set y_b = [r_{b+1} + α max_{a_{t+1}} Q_π(s_{b+1}, a_{t+1}; θ_t)]
        Perform SGD on (y_b − Q_π(s_b, a_b; θ_t))^2 to find θ*
        Update θ_t = θ* in the DQN
        s_t = s_{t+1}
    end for
end for
```

After the convergence of the training, we use the optimized weights for the evaluation (testing) of the DRL algorithm to assess the quality of the learned policy [20]. The evaluation can be performed during training or after that. In this phase, the agent chooses its actions greedily (no exploration) for each state.

Our method does not need any explicit CSI information, it only need the power measurements from which SINR will be measured, as in [21]. To achieve this, when the serving BS is not transmitting, each UE in that cell will receive and measure interference plus noise ($I + N$) level. Next, when the serving BS is transmitting, the UE will measure signal plus interference plus noise ($S + I + N$) level. The received power (S) of the UE can hence be determined by subtracting two measurements, and the SINR can be approximately obtained by $S/(I + N)$. UEs then fed back SINR to their serving BS. The serving BS receives the immediate reward r_{t+1} which is SINR feedback by user. The serving BS relays the information to central agent which stores these experiences in a replay memory data set $R = \{e_1, \ldots, e_t\}$, where $e_t = \{\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1}\}$ represents the current state, action, reward, and the next state after executing $\boldsymbol{a}_t$ at state $\boldsymbol{s}_t$.



(a) CCDF for $L = 3$ with $M = 16$. (b) CCDF for $L = 3$ with different M. (c) Convergence curve for $L = 2$.

Fig. 2: Performance of the proposed DQN algorithm for $L = 2$ and $L = 3$ cell network.

Our approach utilizes centralized DRL in which all cooperating BSs are connected to a central node via backhaul links. This central node contain an agent which calculates the actions strategy (optimized power allocation and beamforming vector) for the entire network and communicates it to the BSs through this backhaul links. This optimized power and beamforming vector optimize the UE's received SINR.

The overhead incurred from transmitting data over the backhaul to this central location, considering a total of UEs (N_{UE}) in the service area, follows an order of complexity $\mathcal{O}(gLN_{UE})$. Here, the periodicity g represents the number of measurements sent by any specific UE during time step t [22]. The complexity of brute force is $\mathcal{O}((|\mathcal{P}||\mathcal{W}|)^{UL})$ which is very high.

VI. Training Setup and Simulation Results

The training setup, simulation details, performance measures and numerical results are demonstrated in this section.

A. Simulation Details Performance Measures

We have set up the experiments described below and an appropriate performance measure to show the performance improvement of our proposed method.

1) Simulation Setup: We consider an L-cell network with hexagonal geometry each with a cell radius of $150m$ and inter-site distance $D = 225m$. The operation frequency is 28 GHz. Each L contains $U = 2$ UEs. UEs are uniformly distributed within each cell and move at a speed of 2 km/h. In (1) to (4), where needed $d = \frac{\lambda}{2}$, $N_p = 4$ with probability 0.8 and $N_p = 1$ (line of sight channel) with probability 0.2, and radio frame duration $T = 10ms$. The initial position of the UEs, initial power of the BSs, and initial beamforming vectors are selected randomly.

In order to plot the effective SINR, we set the minimum SINR as $\gamma_{\min} = -3$ dB which represents the minimum SINR for any user in the cellular network. If the SINR falls below the minimum value, the episode aborts which means the call is dropped. We follow the DQN structure of [23]. The training parameters of the DRL are listed in Table I.

TABLE I: DQN parameters.

Parameters	Value
α	0.995
Initial ϵ	1.000
$\epsilon_{\min}$	0.1
Learning rate	0.01
u	24
v	2
$P_{\ell}^{\max}$	20 W
δ	0.995
Batch size, B	32

In our experiments, we use Rectified Linear Unit activation and Adam optimizer for network. All the simulation results are obtained with TensorFlow 1.14.0 and python 3.7.

Spectral efficiency (measured by achievable sum-rate) is the main performance evaluation measure. We evaluate the average network sum-rate by

$$R_{\text{sum}} = \frac{1}{E}\sum_{e=1}^{E}\sum_{\ell=1}^{L}\sum_{u=1}^{U}\log_2(1+\gamma_{\ell,u}^{[e]}), \tag{17}$$

where E is the total number of episodes and $\gamma_{\ell,u}^{[e]}$ is the SINR at episode e. Another performance measure is overall network coverage, evaluated by the *complementary cumulative distribution function (CCDF)* of the SINR ($\gamma_{\ell,u}$ for all cells).

B. Results

We first compare the performance of the proposed algorithm with that of the algorithm given in [9] and the brute force method in Fig. 2. In Fig. 2(a), CCDF of $\gamma_{\ell,u}$ for proposed and existing method [9] are plotted for three cell networks. It is noticeable that our DQN algorithm brings further gain in SINR compared to that of the existing method. The performs of existing method [9] is the worst because it trains the network until reaching the minimum value. Due to the limitation of this approach the network is not able to learn more about the environment. For example, while using existing method [9] only 9% of the time the UEs have SINR > 20 dB, this number is about 18% for the proposed algorithm. Further, it can be seen that the network sum-rate for the proposed algorithm is very close to that of the brute force method, which can

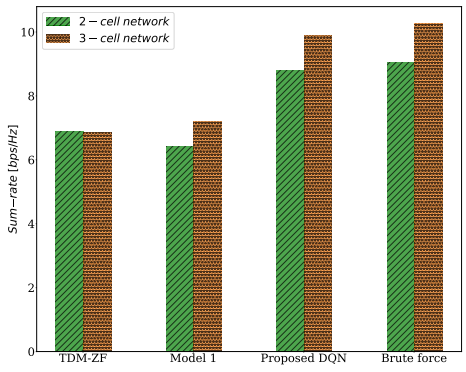


Fig. 3: Network sum-rate for the proposed DQN algorithm versus other methods for $L = 2$ and $L = 3$ with $M = 16$.

be seen as the upper bound. In Fig. 2(b), we see that as M increases, the probability of having higher SINRs increases since it depends on the beamforming array gain which is directly related to the M. In Fig. 2(c), average user rate in the network for proposed method is plotted. It can be observed that after a certain iteration the algorithm is converging.

In Fig. 3, we compare the network sum-rate for the three different algorithms. It is noticeable that our DQN algorithm improves the SINR compared to that of *Model* 1 and *TDM-ZF* (Time-division multiplexing Zero Forcing) [24]. Calculating sum-rate of *TDM-ZF* require global knowledge of the CSI. *TDM-ZF* replaces the problem of interference in an L-cell network with L independent single-cell problems, making it immune to interference due to which the performance is not good. Our method, on the other hand, uses universal frequency and does not need CSI knowledge. Yet, network sum-rate is increased almost linearly with L. That is, interference can be harnessed with practical assumptions. Further, it can be seen that the network sum-rate for the proposed algorithm is very close to that of the brute force method.

VII. Conclusions

A deep reinforcement learning-based interference management in a multi-cell multi-user mmWave network has been proposed in this paper. The main objective is to maximize the network sum-rate without requiring to have access to the explicit knowledge of CSI. We assume that each BS selects its beamforming vector and power command from the finite set. The input features of DQN are the UEs coordinates, BSs power, and beamforming vectors. The output has a sequence of interference management along with power control and beamforming that optimize the objective function. Our proposed algorithm scales very well and achieves nearly the same sum-rate when up-to-date CSI (brute force) is known. Furthermore, the performance of the algorithms improves as the number of cells increases.

References

[1] S. Sun, Q. Gao, Y. Peng, Y. Wang, and L. Song, "Interference management through CoMP in 3GPP LTE-advanced networks," *IEEE Wirel. Commun.*, vol. 20, pp. 59–66, 2013.

[2] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3457–3470, 2008.

[3] O. El Ayach, S. W. Peters, and R. W. Heath, "The practical challenges of interference alignment," *IEEE Wirel. Commun.*, vol. 20, no. 1, pp. 35–42, 2013.

[4] S. A. Jafar, *Interference alignment: A new look at signal dimensions in a communication network*. Now Publishers Inc, 2011.

[5] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," *IEEE Commun. Mag*, vol. 50, no. 2, pp. 148–155, 2012.

[6] J. Choi, "Massive MIMO with joint power control," *IEEE Wirel. Commun. Lett.*, vol. 3, no. 4, pp. 329–332, 2014.

[7] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 9, pp. 6177–6189, 2018.

[8] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[9] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 4, pp. 2125–2140, 2019.

[10] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Wirel. Commun.*, vol. 68, no. 3, pp. 1581–1592, 2020.

[11] M. Dahal and M. Vaezi, "Deep reinforcement learning for interference management in millimeter-wave networks," in *Proc. IEEE 56th Asilomar Conf. Signals Syst. Comput.*, pp. 1064–1069, 2022.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Overall descriptions," Tech. Rep. Jan. 2019.

[14] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, 2014.

[15] F. B. Mismar and B. L. Evans, "Partially blind handovers for mmWave new radio aided by sub-6 GHz LTE signaling," in *IEEE ICC Workshops*, pp. 1–5, 2018.

[16] D. Moltchanov, E. Sopin, V. Begishev, A. Samuylov, Y. Koucheryavy, and K. Samouylov, "A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems," *IEEE Commun. Surveys Tuts.*, 2022.

[17] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 3, pp. 436–453, 2016.

[18] J. Zhang, Y. Huang, Q. Shi, J. Wang, and L. Yang, "Codebook design for beam alignment in millimeter wave communication systems," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4980–4995, 2017.

[19] J. Nikonowicz, A. Mahmood, M. I. Ashraf, E. Björnson, and M. Gidlund, "Indoor positioning trends in 5G-advanced: Challenges and solution towards centimeter-level accuracy," *arXiv:2209.01183*, 2022.

[20] D. L. Poole and A. K. Mackworth, *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.

[21] M. Vaezi, X. Lin, H. Zhang, W. Saad, and H. V. Poor, "Deep reinforcement learning for interference management in UAV-based 3D networks: Potentials and challenges," *IEEE Commun. Mag.*, 2024.

[22] 3GPP, "NR; Physical channels and modulation," Tech. Rep. 38.211, 2018, version 14.2.2.

[23] G.-B. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 274–281, 2003.

[24] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.