Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination

Faris B. Mismar, *Senior Member, IEEE*, Brian L. Evans, *Fellow, IEEE*, and Ahmed Alkhateeb, *Member, IEEE*

***Abstract*— The fifth generation of wireless communications (5G) promises massive increases in traffic volume and data rates, as well as improved reliability in voice calls. Jointly optimizing beamforming, power control, and interference coordination in a 5G wireless network to enhance the communication performance to end users poses a significant challenge. In this paper, we formulate the joint design of beamforming, power control, and interference coordination as a non-convex optimization problem to maximize the signal to interference plus noise ratio (SINR) and solve this problem using deep reinforcement learning. By using the greedy nature of deep Q-learning to estimate future rewards of actions and using the reported coordinates of the users served by the network, we propose an algorithm for voice bearers and data bearers in sub-6 GHz and millimeter wave (mmWave) frequency bands, respectively. The algorithm improves the performance measured by SINR and sum-rate capacity. In realistic cellular environments, the simulation results show that our algorithm outperforms the link adaptation industry standards for sub-6 GHz voice bearers. For data bearers in the mmWave frequency band, our algorithm approaches the maximum sum rate capacity, but with less than 4% of the required run time.**

***Index Terms*— Reinforcement learning (RL), deep learning, beamforming, power control, millimeter wave (mmWave).**

I. Introduction

THE massive growth in traffic volume and data rate continues to evolve with the introduction of fifth generation of wireless communications (5G). Also evolving is enhanced voice call quality with better reliability and improved codecs. Future wireless networks are therefore expected to meet this massive demand for both the data rates and the enhanced voice quality. In an attempt to learn the implied characteristics of inter-cellular interference and inter-beam interference, we propose an online learning based algorithm based on a reinforcement learning (RL) framework. We use this framework to derive a near-optimal policy to maximize the end-user signal to interference plus noise (SINR) and sum-rate capacity. The importance of reinforcement learning in power control has been demonstrated in [1]–[3]. Power control in voice bearers makes them more robust against wireless impairments, such as fading. It also enhances the usability of the network and increases the cellular capacity. For data bearers, beamforming, power control, and interference coordination, can improve the robustness of these data bearers, improve the data rates received by the end-users, and avoid retransmissions.

A major question here is whether there exists a method that (1) can jointly solve for the power control, interference coordination, and beamforming, (2) achieve the upper bound on SINR, and (3) avoids the exhaustive search in the action space for both bearer types. The aim of this paper is to propose an algorithm for this joint solution by utilizing the ability of reinforcement learning to explore the solution space by learning from interaction. This algorithm applies to both voice and data bearers alike. Furthermore, we study the overhead introduced as a result of passing information to a central location, which computes the solution through online learning.

A. Related Work

Performing power control and beamforming in both uplink and downlink was studied in [4]–[7]. A jointly optimal transmit power and beamforming vector was solved for in [7] to maximize the SINR using optimization, but without regards for scattering or shadowing, which are critical phenomena in millimeter wave (mmWave) propagation.

The industry standards adopted the method of almost blank subframe (ABS) to resolve the co-channel inter-cell interference problem in LTE where two base stations (BSs) interfere with one another [8]. While ABS works well in fixed beam antenna patterns, the dynamic nature of beamforming reduces the usefulness of ABS [9].

An online learning algorithm for link adaptation in multiple-input multiple-output (MIMO) bearers was studied in [2]. The algorithm computational complexity was comparable to existing online learning approaches, but with minimal spatial overhead. Interference avoidance in a heterogeneous network was studied in [3]. A Q-learning framework for the coexistence of

Manuscript received June 29, 2019; revised September 22, 2019, November 8, 2019, and December 11, 2019; accepted December 16, 2019. Date of publication December 23, 2019; date of current version March 18, 2020. This article was presented at the 2018 Asilomar Conference on Signals, Systems, and Computers [1]. The associate editor coordinating the review of this article and approving it for publication was M. Bennis. *(Corresponding author: Faris B. Mismar.)*

Faris B. Mismar and Brian L. Evans are with the Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: faris.mismar@utexas.edu; bevans@ece.utexas.edu).

Ahmed Alkhateeb is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: alkhateeb@asu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2961332

0090-6778 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

both macro and femto BSs was proposed. The feasibility of decentralized self-organization of these BSs was established where the femtocells interference towards the macro BSs was mitigated. The use of a Q-learning framework was also proposed in [1]. The framework focused on packetized voice power control in a multi-cell indoors environment. It exploits the use of semi-persistent scheduling, which establishes a virtual sense of a dedicated channel. This channel enabled the power control of the downlink to ensure enhanced voice clarity compared to industry standards, which are based on fixed power allocation.

Joint power control in massive MIMO was introduced in [4]. This approach led to a reduced overhead due to a limited exchange of channel state information between the BSs participating in the joint power control. The joint power control scheme led to enhanced performance measured by the SINR. In the uplink direction, power control with beamforming was studied in [5]. An optimization problem was formulated to maximize the achievable sum rate of the two users while ensuring a minimal rate constraint for each user. Using reinforcement learning to solve the problem for the uplink is computationally expensive and can cause a faster depletion of the user equipment (UE) battery. We on the other hand focus on the downlink and on interference cancellation alongside power control and beamforming.

Over the last two years, the use of deep learning in wireless communications was studied in [6], [10]–[16]. The specific use of deep reinforcement learning to perform power control for mmWave was studied in [6]. This approach was proposed as an alternative to beamforming in improving the non-line of sight (NLOS) transmission performance. The power allocation problem to maximize the sum-rate of UEs under the constraints of transmission power and quality targets was solved using deep reinforcement learning. In this solution, a convolutional neural network was used to estimate the Q-function of the deep reinforcement learning problem. In [10], a policy that maximizes the successful transmissions in a dynamic correlated multichannel access environment was obtained using deep Q-learning. The use of deep convolutional neural networks was proposed in [11] to enhance the automatic recognition of modulation in cognitive radios at low SINRs.

In [15], deep neural networks were leveraged to predict mmWave beams with low training overhead using the omni-directional received signals collected from neighboring base stations. In [16], the authors generalized [15] by mapping the channel knowledge at a small number of antennas to an SINR-optimal beamforming vector for a larger array, even if this array was at a different frequency at a neighboring BS.

The use of adversarial reinforcement learning in beamforming for data bearers was proposed in [17], where an algorithm to derive antenna diagrams with near-optimal SINR performance was devised. There was no reference to power control or interference coordination. Voice bearers in the sub-6 GHz frequency band was studied in [18] but only in a single co-located BS environment, in contrast with our paper where we study voice in a multi-access network with multiple BSs. Joint beamforming and interference coordination at mmWave was performed in [19] using deep neural networks, which require knowledge of the channel to make decisions. The performance of deep neural networks on beamforming was studied in [20] but without the use of reinforcement learning. Table I shows how our work compares with earlier work.

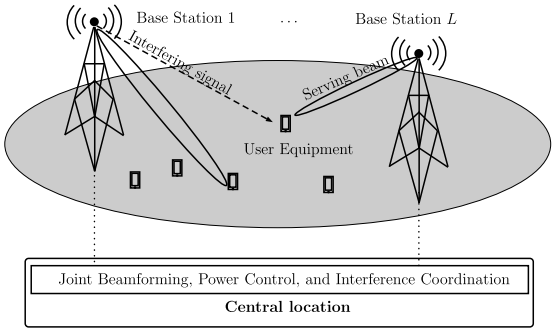

Fig. 1. Performing joint beamforming and power control on the signal from the serving base station while coordinating interference from the other BS. The decisions are computed at a central location, which can be one of the L BSs. The measurements from the UEs are relayed to the central location over the backhaul.

B. Motivation

In this paper, we provide an answer to the question whether a method exists that can perform the joint beamforming, power control, and interference coordination by introducing a different approach to power control in wireless networks. In such a setting, it is not only the transmit power of the serving BS that is controlled as in standard implementations, but also the transmit powers of the interfering base stations from a central location as shown in Fig. 1. As a result of this apparent conflict, a *race condition* emerges, where the serving BS of a given user is an interfering BS of another user. The reason why we choose deep reinforcement learning (DRL) is as follows:

1) The proposed solution does not require the knowledge of the channels in order to find the SINR-optimal beamforming vector. This is in contrast with the upper bound SINR performance, which finds the optimal beamforming vector by searching across all the beams in a codebook that maximizes the SINR (and this requires perfect knowledge of the channel).
2) The proposed solution minimizes the involvement of the UE in sending feedback to the BS. In particular, the UE sends back its received SINR along with its coordinates, while the agent handles the power control and interference coordination commands to the involved BSs. Industry specifications [8] require that the UE reports its channel state information which is either a vector of length equal to the number of antenna elements or a matrix of dimension equal to the number of antenna elements in each direction. In our case, we achieve a reduction in the reporting overhead by using the UE coordinates instead.
3) The implementation complexity of upper bound SINR performance message passing for joint beamforming, power control, and interference coordination commands when multiple BSs are involved is prohibitive.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

TABLE I
LITERATURE COMPARISON

Reference	Bearer	Band	Objective	Procedure*	Algorithm
[4]	data	unspecified	downlink SINR	PC	convex optimization
[5]	data	mmWave	uplink sum-rate	BF, PC	convex optimization
[6]	data	mmWave	dowlink SINR, sum-rate	PC	deep reinforcement learning
[12]	data	unspecified	uplink power, sum-rate	PC	deep neural networks
[13]	data	unspecified	downlink throughput	PC	deep neural networks
[14]	data	unspecified	SINR, spectral efficiency	PC	convolutional neural network
[15]	data	mmWave	downlink achievable rate	BF	deep neural networks
[16]	data	mmWave and sub-6	downlink spectral efficiency	BF	deep neural networks
[17]	data	unspecified	downlink sum-rate	BF	deep adversarial reinforcement learning
[18]	voice	sub-6	downlink SINR	PC	tabular reinforcement learning
[19]	data	mmWave	downlink sum-rate	BF, IC	deep neural networks
[20]	data	unspecified	downlink SINR	BF	deep neural networks
Proposed	voice and data	mmWave and sub-6	downlink SINR	BF, PC, IC	deep reinforcement learning

* PC is power control, IC is interference coordination, and BF is beamforming.

4) Having explicit power control and interference coordination (PCIC) commands sent by the UE to the serving and interfering BSs requires a modification to the current industry standards [21]. These standards today only require the serving BS to send power control commands to the UE for the uplink direction.

C. Contributions

In finding a different approach to power control in wireless networks, this paper makes the following specific contributions:

- Formulate the joint beamforming, power control, and interference coordination problem in the downlink direction as an optimization problem that optimizes the users' received SINR.
- Resolve the race condition between the involved base stations in sub-exponential times in the number of antennas. The race condition is handled by a central location (similar to coordinated multipoint [22]) based on the user reported downlink SINR and coordinates.
- Show how to create a deep reinforcement learning based solution where multiple actions can be taken at once using a binary encoding of the relevant actions performed by the BS, which we define in Section VIII-A.

II. NETWORK, SYSTEM, AND CHANNEL MODELS

In this section, we describe the adopted network, system, and channel models.

A. Network Model

We consider an orthogonal frequency division multiplexing (OFDM) multi-access downlink cellular network of L BSs. This network is comprised of a serving BS and at least one interfering BS. We adopt a downlink scenario, where a BS is transmitting to one UE. The BSs have an intersite distance of R and the UEs are randomly scattered in their service area. The association between the users and their serving BS is based on the distance between them. A user is served by one BS maximum. The cell radius is $r > R/2$ to allow overlapping of coverage. Voice bearers run on sub-6 GHz frequency bands while the data bearers use mmWave frequency band. We employ analog beamforming for the data bearers to compensate for the high propagation loss due to the higher center frequency.

B. System Model

Considering the network model in Section II-A, and adopting a multi-antenna setup where each BS employs a uniform linear array (ULA) of M antennas and the UEs have single antennas, the received signal at the UE from the ℓ-th BS can be written as

$$y_\ell = \mathbf{h}_{\ell,\ell}^* \mathbf{f}_\ell x_\ell + \sum_{b \neq \ell} \mathbf{h}_{\ell,b}^* \mathbf{f}_b x_b + n_\ell \tag{1}$$

where $x_\ell, x_b \in \mathbb{C}$ are the transmitted signals from the ℓ-th and b-th BSs, and they satisfy the power constraint $\mathbb{E}[|x_\ell|^2] = P_{\mathrm{TX},\ell}$ (similarly for b). The $M \times 1$ vectors $\mathbf{f}_\ell, \mathbf{f}_b \in \mathbb{C}^{M\times 1}$ denote the adopted downlink beamforming vectors at the ℓ-th and b-th BSs, while the $M \times 1$ vectors $\mathbf{h}_{\ell,\ell}, \mathbf{h}_{\ell,b} \in \mathbb{C}^{M\times 1}$ are the channel vectors connecting the user at the ℓ-th BS with the ℓ-th and b-th BSs, respectively. Finally, $n_\ell \sim \mathrm{Normal}(0, \sigma_n^2)$ is the received noise at the user sampled from a complex Normal distribution with zero-mean and variance σ_n^2. The first term in (1) represents the desired received signal, while the second term represents the interference received at the user due to the transmission from the other BSs.

1) Beamforming Vectors: Given the hardware constraints on the mmWave transceivers, we assume that the BSs use analog-only beamforming vectors, where the beamforming weights of every beamforming vector $\mathbf{f}_\ell, \ell = 1, 2, \ldots, L$ are implemented using constant-modulus phase shifters, i.e., $[\mathbf{f}_\ell]_m = e^{j\theta_m}$. Further, we assume that every beamforming vector is selected from a beamsteering-based beamforming codebook $\mathcal{F}$ of cardinality $|\mathcal{F}| := N_{\mathrm{CB}}$, with the n-th element in this codebook defined as

$$\begin{aligned}\mathbf{f}_n &:= \mathbf{a}(\theta_n) \\ &= \frac{1}{\sqrt{M}} \left[1, e^{jkd\cos(\theta_n)}, \ldots, e^{jkd(M-1)\cos(\theta_n)}\right]^\top,\end{aligned} \tag{2}$$

where d and k denote the antenna spacing and the wavenumber, while θ_n represents the steering angle. Finally, $\mathbf{a}(\theta_n)$

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

is the array steering vector in the direction of θ_n. The value of θ_n is obtained by dividing the the antenna angular space between 0 and π radians by the number of antennas M.

2) Power Control and Interference Coordination: Every BS ℓ is assumed to have a transmit power $P_{\text{TX},\ell} \in \mathcal{P}$, where $\mathcal{P}$ is the set of candidate transmit powers. We define the set of the transmit powers as the power offset above (or below) the BS transmit power. Our choice of the transmit power set $\mathcal{P}$ is provided in Section VIII-A. This choice of $\mathcal{P}$ follows [21].

Power control and interference coordination take place over a semi-dedicated channel. For voice, this is facilitated through the semi-persistent scheduling, which creates a virtual sense of a dedicated channel as we have mentioned in Section I. For data bearers, the use of beamforming provides a dedicated beam for a given UE, through which power control and interference coordination takes place.

C. Channel Model

In this paper, we adopt a narrow-band geometric channel model, which is widely considered for analyzing and designing mmWave systems [23]–[25]. With this geometric model, the downlink channel from a BS b to the user in BS ℓ can be written as

$$\mathbf{h}_{\ell,b} = \frac{\sqrt{M}}{\rho_{\ell,b}} \sum_{p=1}^{N^p_{\ell,b}} \alpha^p_{\ell,b} \mathbf{a}^* \left(\theta^p_{\ell,b}\right), \tag{3}$$

where $\alpha^p_{\ell,b}$ and $\theta^p_{\ell,b}$ are the complex path gain and angle of departure (AoD) of the p-th path, and $\mathbf{a}(\theta^p_{\ell,b})$ is the array response vector associated with the AoD, $\theta^p_{\ell,b}$. Note that $N^p_{\ell,b}$ which denotes the number of channel paths is normally a small number in mmWave channels compared to sub-6 GHz channels [26], [27], which captures the sparsity of the channels in the angular domain. Finally, $\rho_{\ell,b}$, represents the path-loss between BS b and the user served in the area of BS ℓ. Note that the channel model in (3) accounts of both the LOS and NLOS cases. For the LOS case, we assume that $N^p_{\ell,b} = 1$.

We define $P_{\text{UE}}[t]$ as the received downlink power as measured by the UE over a set of physical resource blocks (PRBs) at a given time t as

$$P^{\ell,b}_{\text{UE}}[t] = P_{\text{TX},b}[t] \left|\mathbf{h}^*_{\ell,b}[t]\mathbf{f}_b[t]\right|^2 \tag{4}$$

where $P_{\text{TX},b}$ is the PRB transmit power from BS b. Next, we compute the received SINR for the UE served in BS ℓ at time step t as follows:

$$\gamma^\ell[t] = \frac{P_{\text{TX},\ell}[t]|\mathbf{h}^*_{\ell,\ell}[t]\mathbf{f}_\ell[t]|^2}{\sigma_n^2 + \sum_{b\neq\ell} P_{\text{TX},b}[t]|\mathbf{h}^*_{\ell,b}[t]\mathbf{f}_b[t]|^2}. \tag{5}$$

This is the received SINR that we will optimize in our paper in Sections V and VI.

III. Problem Formulation

Our objective is to jointly optimize the beamforming vectors and the transmit power at the L BSs to maximize the achievable sum rate of the users. We formulate the joint beamforming, power control, and interference coordination optimization problem as

$$\begin{aligned} \underset{\substack{P_{\text{TX},j}[t],\ \forall j \\ \mathbf{f}_j[t],\ \forall j}}{\text{maximize}} \quad & \sum_{j\in\{1,2,\ldots,L\}} \gamma^j[t] \\ \text{subject to} \quad & P_{\text{TX},j}[t] \in \mathcal{P}, \quad \forall j, \\ & \mathbf{f}_j[t] \in \mathcal{F}, \quad \forall j, \\ & \gamma^j[t] \geq \gamma_{\text{target}}. \end{aligned} \tag{6}$$

where γ_{target} denotes the target SINR of the downlink transmission (all SINR quantities are in dB). $\mathcal{P}$ and $\mathcal{F}$ are the sets of candidate transmit powers and beamforming codebook, respectively as stated earlier. This problem is a non-convex optimization problem due to the non-convexity of the first two constraints. The ℓ-th BS attempts to solve this problem to find optimal $P_{\text{TX},\ell}$ and $\mathbf{f}_\ell$ for the UE served by it at time t. We solve this optimization problem at a central location by searching over the space of the Cartesian product of $\mathcal{P}\times\mathcal{F}$. The optimal solution to this problem is found through an exhaustive search over this space (i.e., by brute force). The complexity of this search is known to be exponential in the number of BSs. We discuss this and the overhead of the communication to a central location in Section VI.

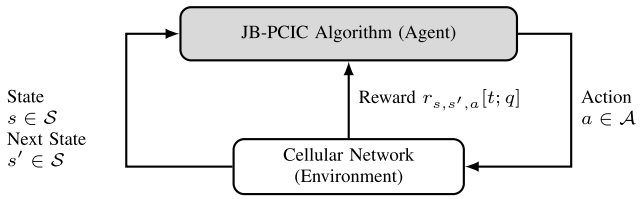

Fig. 2. The agent-environment interaction in reinforcement learning.

Next, we provide a brief overview on deep reinforcement learning in Section IV before delving into the proposed algorithm in Sections V and VI.

IV. A Primer on Deep Reinforcement Learning

In this section, we describe deep reinforcement learning, which is a special type of reinforcement learning [28]. Reinforcement learning is a machine learning technique that enables an *agent* to discover what action it should take to maximize its expected future *reward* in an interactive *environment*. The interaction between the agent and the environment is shown in Fig. 2. In particular, DRL exploits the ability of deep neural networks to learn better representations than handcrafted features and act as a universal approximator of functions.

A. Reinforcement Learning Elements

Reinforcement learning has several elements [29]. These elements interact together, and are as follows:

- *Observations*: Observations are continuous measures of the properties of the environment and are written as a p-ary vector $\mathbf{O} \in \mathbb{R}^p$, where p is the number of properties observed.
- *States*: The state $s_t \in \mathcal{S}$ is the discretization of the observations at time step t. Often, states are also used to mean observations.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

- *Actions*: An action $a_t \in \mathcal{A}$ is one of the valid choices that the agent can make at time step t. The action changes the state of the environment from the current state s to the target state s'.
- *Policy*: A policy $\pi(\cdot)$ is a mapping between the state of the environment and the action to be taken by the agent. We define our stochastic policy $\pi(a \,|\, s) : \mathcal{S} \times \mathcal{A} \to [0, 1]$.
- *Rewards*: The reward signal $r_{s,s',a}[t; q]$ is obtained after the agent takes an action a when it is in state s at time step t and moves to the next state s'. The parameter $q \in \{0, 1\}$ is the bearer selector, which is a binary parameter to differentiate voice bearers from data bearers.
- *State-action value function*: The state-action value function under a given policy π is denoted $Q_\pi(s, a)$. It is the expected discounted reward when starting in state s and selecting an action a under the policy π.

These elements work together and their relationship is governed by the objective to maximize the future discounted reward for every action chosen by the agent, which causes the environment to transition to a new state. The policy dictates the relationship between the agent and the state. The value of the expected discounted reward is learned through the training phase.

If $Q_\pi(s, a)$ is updated every time step, then it is expected to converge to the optimal state-action value function $Q^\star_\pi(s, a)$ as $t \to +\infty$ [29]. However, this may not be easily achieved. Therefore, we use a function approximator instead aligned with [28]. We define a neural network with its weights at time step t as $\boldsymbol{\Theta}_t \in \mathbb{R}^{u \times v}$ as in Fig. 3. Also, if we define $\boldsymbol{\theta}_t := \mathrm{vec}\,(\boldsymbol{\Theta}_t) \in \mathbb{R}^{uv}$, we thus build a function approximator $Q_\pi(s, a; \boldsymbol{\theta}_t) \approx Q^\star_\pi(s, a)$. This function approximator is neural network based and is known as the Deep Q-Network (DQN) [28]. Activation functions, which are non-linear functions that compute the hidden layer values, are an important component of neural networks. A common choice of the activation function is the sigmoid function $\sigma\colon x \mapsto 1/(1 + e^{-x})$ [30]. This DQN is trained through adjusting $\boldsymbol{\theta}$ at every time step t to reduce the mean-squared error loss $L_t(\boldsymbol{\theta}_t)$:

$$\underset{\boldsymbol{\theta}_t}{\text{minimize}}\; L_t(\boldsymbol{\theta}_t) := \mathbb{E}_{s,a}\left[(y_t - Q_\pi(s, a; \boldsymbol{\theta}_t))^2\right] \tag{7}$$

where $y_t := \mathbb{E}_{s'}[r_{s,s',a} + \gamma \max_{a'} Q_\pi(s', a'; \boldsymbol{\theta}_{t-1}) \,|\, s_t, a_t]$ is the estimated function value at time step t when the current state and action are s and a respectively. The process of interacting with the environment and the DQN to obtain a prediction and compare it with the true answer and suffer a loss $L_t(\cdot)$ is often referred to as "online learning." In online learning, the UEs feedback their data to the serving BS, which in turns relays it to the central location for DQN training. This data represent the state of our network environment $\mathcal{S}$, as we explain further in Section VIII.

B. DQN Dimension

we set the dimension of the input layer in the DQN to be equal to the number of states $|\mathcal{S}|$. The dimension of the output layer is equal to the number of actions $|\mathcal{A}|$. For the hidden layer dimension, we choose a small depth since the depth has the greatest impact on the computational complexity.

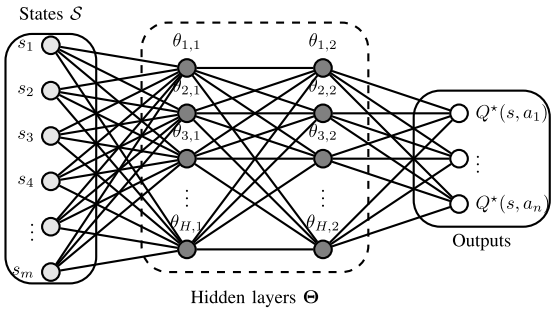

Fig. 3. Structure of the deep Q-network used for the implementation of the algorithms with two hidden layers each of dimension H. Here, $(u, v) = (H, 2)$, $|\mathcal{S}| = m$, and $|\mathcal{A}| = n$.

The dimension of the width follows [31] as we show further in Section VIII-A.

C. Deep Reinforcement Training Phase

In the training phase of the DQN, the weights $\boldsymbol{\theta}_t$ in the DQN are updated after every iteration in time t using the stochastic gradient descent (SGD) algorithm on a minibatch of data. SGD starts with a random initial value of $\boldsymbol{\theta}$ and performs an iterative process to update $\boldsymbol{\theta}$ using a step size $\eta > 0$ as follows:

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t - \eta \nabla L_t(\boldsymbol{\theta}_t). \tag{8}$$

The training of the DQN is facilitated by "experience replay" [32]. The experience replay buffer $\mathcal{D}$ stores the experiences at each time step t. An experience e_t is defined as $e_t := (s_t, a_t, r_{s,s',a}[t; q], s'_t)$. We draw samples of experience at random from this buffer and perform minibatch training on the DQN. This approach offers advantages of stability and avoidance of local minimum convergence [28]. The use of experience replay also justifies the use of off-policy learning algorithms, since the current parameters of the DQN are different from those used to generate the sample from $\mathcal{D}$.

We define the state-action value function estimated by the DQN $Q^\star_\pi(s, a)$ as

$$Q^\star_\pi(s_t, a_t) := \mathbb{E}_{s'}\left[r_{s,s',a} + \gamma \max_{a'} Q^\star_\pi(s', a') \,\middle|\, s_t, a_t\right], \tag{9}$$

which is known as the Bellman equation. Here, $\gamma\colon 0 < \gamma < 1$ is the discount factor and determines the importance of the predicted future rewards. The next state is s' and the next action is a'. Our goal using DQN is to find a solution to maximize the state-action function $Q^\star_\pi(s_t, a_t)$.

Often compared with deep Q-learning is the tabular version of Q-learning [29]. Despite the finite size of the states and action space, tabular Q-learning is slow to converge is because its convergence requires the state-action pairs to be sampled infinitely often [29], [33]. Further, tabular RL requires a non-trivial initialization of the $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ table to avoid longer convergence times [18]. However, deep Q-learning convergence is not guaranteed when using a non-linear approximator such as the DQN [28]. We discuss tabular Q-learning in Section V.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

D. Policy Selection

In general, Q-learning is an off-policy reinforcement learning algorithm. An off-policy algorithm means that a near-optimal policy can be found even when actions are selected according to an arbitrary exploratory policy [29]. Due to this, we choose a near-greedy action selection policy. This policy has two modes:

1) *exploration*: the agent tries different actions at random at every time step t to discover an effective action a_t.
2) *exploitation*: the agent chooses an action at time step t that maximizes the state-action value function $Q_\pi(s, a; \boldsymbol{\theta}_t)$ based on the previous experience.

In this policy, the agent performs exploration with a probability ϵ and exploitation with probability of $1 - \epsilon$, where ϵ: $0 < \epsilon < 1$ is a hyperparameter that adjusts the trade-off between exploration and exploitation. This trade-off is why this policy is also called the ϵ-greedy action selection policy. This policy is known to have a linear regret in t (regret is the opportunity loss of one time step) [34].

At each time step t, the UEs move at speed v and the agent performs a certain action a_t from its current state s_t. The agent receives a reward $r_{s,s',a}[t; q]$ and moves to a target state $s' := s_{t+1}$. We call the period of time in which an interaction between the agent and the environment takes place an *episode*. One episode has a duration of T time steps. An episode is said to have *converged* if within T time steps the target objective was fulfilled.

In our DQN implementation, we particularly keep track of the UE coordinates. When UE coordinates are reported back to the network and used to make informed decisions, the performance of the network improves [35]. Therefore, UE coordinates need to be part of the DRL state space $\mathcal{S}$.

V. Deep Reinforcement Learning in Voice Power Control and Interference Coordination

In this section, we describe our proposed voice power control and interference coordination reinforcement learning algorithm as well as the baseline solutions which we compare our solution against. First, we describe the fixed power allocation algorithm, which is the industry standard algorithm today, and then the implementation of the proposed algorithm using tabular and deep implementations of Q-learning. Finally, we explain the brute force algorithm.

A. Fixed Power Allocation

We introduce the fixed power allocation (FPA) power control as a baseline algorithm that sets the transmit signal power at a specific value. No interference coordination is implemented in FPA. Total transmit power is simply divided equally among all the PRBs and is therefore constant:

$$P_{\text{TX},b}[t] := P_{\text{BS}}^{\max} - 10 \log N_{\text{PRB}} + 10 \log N_{\text{PRB},b}[t] \quad \text{(dBm)} \tag{10}$$

where N_{PRB} is the total number of physical resource blocks in the BS and $N_{\text{PRB},b}$ is the number of available PRBs to the UE in the b-th BS at the time step t.

FPA with adaptive modulation and coding is the industry standard algorithm [21]. In this standard algorithm, the BS fixes its transmit power and only changes the modulation and code schemes of the transmission. This change is known as the "link adaptation." Link adaptation takes place based on the reports sent by the UE back to the BS (i.e., the SINR and received power). Since the BS transmit power is fixed, the link adaptation takes place based on either periodic or aperiodic measurement feedback from the voice UE to the serving BS. This results in an improved effective SINR and a reduction in the voice packet error rate. There is no measurement sent to the interfering BS based on FPA.

B. Tabular RL

We use a tabular setting of Q-learning (or "vanilla" Q-learning) to implement the algorithm for voice communication. In a tabular setting, the state-action value function $Q_\pi(s_t, a_t)$ is represented by a table $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$. There is no neural network involvement and the Q-learning update analog of (9) is defined as:

$$Q_\pi(s_t, a_t) := (1 - \alpha) Q_\pi(s_t, a_t) + \alpha \left(r_{s,s',a} + \gamma \max_{a'} Q_\pi(s', a') \right) \tag{11}$$

where $Q_\pi(s_t, a_t) := [\mathbf{Q}]_{s_t, a_t}$. Here, $\alpha > 0$ is the learning rate of the Q-learning update and defines how aggressive the experience update is with respect to the prior experience. Computationally, the tabular setting suits problems with small state spaces, and maintaining a table $\mathbf{Q}$ is possible.

C. Proposed Algorithm

We propose Algorithm 1 which is a DRL-based approach. This algorithm performs both power control and interference coordination without the UE sending explicit power control or interference coordination commands. This use of the DQN may provide a lower computational overhead compared to the tabular Q-learning depending on the number of states and the depth of the DQN [18]. The main steps of Algorithm 1 are as follows:

- Select an optimization action at a time step t.
- Select a joint beamforming, power control, and interference coordination action.
- Assess the impact on effective SINR $\gamma_{\text{eff}}^{\ell}[t]$.
- Reward the action taken based on the impact on $\gamma_{\text{eff}}^{\ell}[t]$ and its distance from γ_{target} or γ_{min}.
- Train the DQN based on the outcomes.

Power control for the serving BS b is described as

$$P_{\text{TX},b}[t] = \min(P_{\text{BS}}^{\max}, P_{\text{TX},b}[t-1] + \text{PC}_b[t]). \tag{12}$$

We add one more condition for the interference coordination on the interfering BS ℓ as

$$P_{\text{TX},\ell}[t] = \min(P_{\text{BS}}^{\max}, P_{\text{TX},\ell}[t-1] + \text{IC}_\ell[t]) \tag{13}$$

where the role of the BS (serving vs. interfering) can change based on the UE being served. IC and PC commands are actually the same, but the role of the BS makes one an interferer (which needs coordination) and the other a server

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

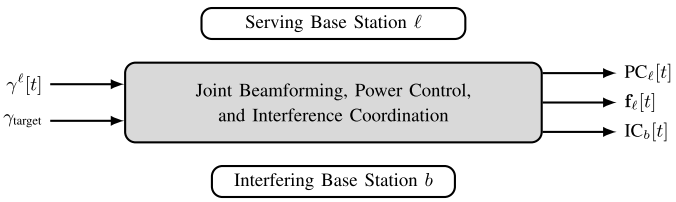

Fig. 4. Downlink joint beamforming, power control, and interference coordination module.

(which needs power control). We model the PCIC algorithm using deep Q-learning as shown in Algorithm 1. Our proposed algorithm solves (6).

Different from [1], we use the effective SINR $\gamma^{\ell}_{\text{eff}}[t]$ (i.e., the SINR including coding gain) for all three voice algorithms where the adaptive code rate β is chosen based on the SINR $\gamma^{\ell}[t]$. We use an adaptive multirate (AMR) codec and quadrature phase shift keying modulation for voice. We choose to fix the modulation since voice bearers do not typically require high data rates [1]. This effective SINR $\gamma^{\ell}_{\text{eff}}[t]$ is the quantity we optimize in Algorithm 1.

For FPA, the run-time complexity is $\mathcal{O}(1)$. For tabular Q-learning PCIC, the run-time complexity is $\mathcal{O}(|\mathcal{S}^{\text{voice}}||\mathcal{A}^{\text{voice}}|)$ [18], where $\mathcal{S}^{\text{voice}}, \mathcal{A}^{\text{voice}}$ are the state and action sets for voice bearers.

Since one of the L BSs also serves as a central location to the surrounding BSs in our proposed algorithm, the overhead due to transmission over the backhaul to this central location for a total of N_{UE} UEs in the service area is in $\mathcal{O}(gLN_{\text{UE}})$, where the periodicity g is the number of measurements sent by any given UE during time step t [36].

D. Brute Force

The brute force PCIC algorithm uses an exhaustive search in the Euclidean space $\mathcal{P}$ per BS to optimize the SINR. This algorithm solves (6) and is the upper limit of the performance for jointly optimizing the SINR for the voice bearers in our problem.

VI. Deep Reinforcement Learning in mmWave Beamforming Power Control and Interference Coordination

In this section, we present our proposed algorithms and quantify the changes in the SINR as a result of the movement of the UEs and optimization actions of the RL-based algorithm.

A. Proposed Algorithm

We propose a DRL-based algorithm where the beamforming vectors and transmit powers at the base stations are jointly controlled to maximize the objective function in (6). The use of a string of bits as an action register enables us to jointly perform several actions concurrently.

First, selecting the beamforming vector is performed as follows. The agent steps up or down the beamforming codebook using circular increments $(n+1)$ or decrements $(n-1)$

$$n \mapsto \mathbf{f}_n[t] \colon n := (n \pm 1) \bmod M \tag{14}$$

Algorithm 1: Deep Reinforcement Learning in Joint Beamforming and PCIC (JB-PCIC)

Input: The downlink received SINR measured by the UEs.
Output: Sequence of beamforming, power control, and interference coordination commands to solve (6).

Initialize time, states, actions, and replay buffer $\mathcal{D}$.
repeat
  repeat
    $t := t + 1$
    Observe current state s_t.
    $\epsilon := \max(\epsilon \cdot d, \epsilon_{\text{min}})$
    Sample $r \sim \text{Uniform}(0, 1)$
    if $r \leq \epsilon$ **then**
      Select an action $a_t \in \mathcal{A}$ at random.
    else
      Select an action $a_t = \arg\max_{a'} Q_{\pi}(s_t, a'; \boldsymbol{\theta}_t)$.
    end
    Compute $\gamma^{\ell}_{\text{eff}}[t]$ and $r_{s,s',a}[t;q]$ from (17).
    if $\gamma^{\ell}_{\text{eff}}[t] < \gamma_{\text{min}}$ **then**
      $r_{s,s',a}[t;q] := r_{\text{min}}$
      Abort episode.
    end
    Observe next state s'.
    Store experience $e[t] \triangleq (s_t, a_t, r_{s,s',a}, s')$ in $\mathcal{D}$.
    Minibatch sample from $\mathcal{D}$ for experience $e_j \triangleq (s_j, a_j, r_j, s_{j+1})$.
    Set $y_j := r_j + \gamma \max_{a'} Q_{\pi}(s_{j+1}, a'; \boldsymbol{\theta}_t)$
    Perform SGD on $(y_j - Q_{\pi}(s_j, a_j; \boldsymbol{\theta}_t))^2$ to find $\boldsymbol{\theta}^{\star}$
    Update $\boldsymbol{\theta}_t := \boldsymbol{\theta}^{\star}$ in the DQN and record loss L_t
    $s_t := s'$
  until $t \geq T$
until convergence or aborted
if $\gamma^{\ell}_{\text{eff}}[t] \geq \gamma_{\text{target}}$ **then** $r_{s,s',a}[t;q] := r_{s,s',a}[t;q] + r_{\text{max}}$

for BSs b and ℓ independently. We monitor the change in γ^{ℓ} as a result of the change in the beamforming vector. We use a code gain of unity in computing $\gamma^{\ell}_{\text{eff}}$ for the data bearers (i.e., $\gamma^{\ell}_{\text{eff}} = \gamma^{\ell}$).

When the beamforming vectors are selected for a given UE, the agent also performs power control of that beam by changing the transmit power of the BS to this UE (or the interference coordination of other BSs). The selection of the transmit power is governed by (12) and (13), both of which define the set $\mathcal{P}$.

For proposed algorithm, the run time of the deep reinforcement learning is significantly faster than the upper bound algorithm for all antenna sizes M as we show in Section VIII-C. Also, the reporting of the UE coordinates (i.e., longitude and latitude) to the BS instead of the channel state information reduces the reporting overhead from M complex-valued elements to the two real-valued coordinates and its received SINR only. If we assume that the reporting overhead for M complex-valued elements is $2M$, then for reporting the UE coordinates, we achieve an overhead reduction gain of $1 - 1/M$.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

We call our algorithm the *joint beamforming, power control, and interference coordination* (JB-PCIC) algorithm.

B. Brute Force

The brute force beamforming and PCIC algorithm uses an exhaustive search in the Euclidean space $\mathcal{P} \times \mathcal{F}$ per BS to optimize the SINR. As in the voice bearers brute force algorithm, this is also the upper limit in the performance for jointly optimizing the SINR in our problem. While the size of $\mathcal{P}$ can be selected independently of the number of the antennas in the ULA M, the size of $\mathcal{F}$ is directly related to M. Similar to the brute force algorithm for voice bearers, this algorithm solves (6) and may perform well for small M and small number of BSs L for data bearers. However, we observe that with large M the search time becomes prohibitive. This is because the run time for this algorithm in $\mathcal{O}((|\mathcal{P}||\mathcal{F}|)^L) = \mathcal{O}(M^L)$, which is much larger than the run time for the proposed algorithm, as we show in Section VIII-C.

VII. Performance Measures

In this section we introduce the performance measures we use to benchmark our algorithms.

A. Convergence

We define convergence ζ in terms of the episode at which the target SINR is fulfilled over the entire duration of T for all UEs in the network. We expect that as the number of antennas in the ULA M increase, the convergence time ζ will also increase. In voice, convergence as a function of M is not applicable, since we only use single antennas. For several random seeds, we take the aggregated percentile convergence episode.

B. Run time

While calculating the upper bound of the brute force algorithm run-time complexity is possible, obtaining a similar expression for the proposed deep Q-learning algorithm may be challenging due to lack of convergence and stability guarantees [28]. Therefore, we obtain the run time from simulation per antenna size M.

C. Coverage

We build a complement cumulative distribution function (CCDF) of $\gamma_{\text{eff}}^{\ell}$ following [38] by running the simulation many times and changing the random seed, effectively changing the way the users are dropped in the network.

D. Sum-Rate capacity

Using the effective SINRs, we compute the average sum-rate capacity as

$$C = \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in\{\ell,b\}} \log_2(1+\gamma_{\text{eff}}^{j}[t]) \tag{15}$$

which is an indication of the data rate served by the network. We then obtain the maximum sum-rate capacity resulting from computing (15) over many episodes.

VIII. Simulation Results

In this section, we evaluate the performance of our RL-based proposed solutions in terms of the performance measures in Section VII. First, we describe the adopted setup in Section VIII-A before delving into the simulation results in Sections VIII-B and VIII-C.

A. Setup

We adopt the network, signal, and channel models in Section II. The users in the urban cellular environment are uniformly distributed in its coverage area. The users are moving at a speed v with both log-normal shadow fading and small-scale fading. The cell radius is r and the inter-site distance $R = 1.5r$. For the voice bearer, we set the adaptive code rate β between 1:3 to 1:1 based on reported SINR and use an AMR voice codec bitrate of 23.85 kbps and a voice activity factor $\nu = 0.8$. The users experience a probability of line of sight of p_{LOS}. The rest of the parameters are shown in Table III. We set the target effective SINRs as:

$$\begin{aligned}\gamma_{\text{target}}^{\text{voice}} &:= 3\,\text{dB},\\ \gamma_{\text{target}}^{\text{bf}} &:= \gamma_0^{\text{bf}} + 10\log M\,\text{dB}\end{aligned} \tag{16}$$

where γ_0^{bf} is a constant threshold (i.e., not dependent on the antenna size). We set the minimum SINR at -3 dB below which the episode is declared aborted and the session is unable to continue (i.e., dropped).

The hyperparameters required to tune the RL-based model are shown in Table II. We refer to our source code [39] for further implementation details. Further, we run Algorithm 1 on the cellular network with its parameters in Table III. The simulated states $\mathcal{S}$ are setup as:

$$\begin{aligned}(s_t^0, s_t^1) &:= \text{UE}_\ell(x[t], y[t]), \quad (s_t^2, s_t^3) := \text{UE}_b(x[t], y[t]),\\ s_t^4 &:= P_{\text{TX},\ell}[t], \quad s_t^5 := P_{\text{TX},b}[t],\\ s_t^6 &:= \mathbf{f}_n^{\ell}[t], \quad s_t^7 := \mathbf{f}_n^{b}[t],\end{aligned}$$

where (x, y) are the Cartesian coordinates (i.e., longitude and latitude) of the given UE.

To derive the actions $\mathcal{A}$, we exploit the fact that $\mathcal{F}$ and $\mathcal{P}$ each has a cardinality that is a power of two. This enables us to construct the binary encoding of the actions using a register $\mathbf{a}$ as shown in Fig. 5. With bitwise-AND, masks, and shifting, the joint beamforming, power control, and interference coordination commands can be derived. We choose the following code:

1) When $q = 0$:
   - $\mathbf{a}_{[0,1]} = 00$: decrease the BS b transmit power by 3 dB.
   - $\mathbf{a}_{[0,1]} = 01$: decrease the BS b transmit power by 1 dB.
   - $\mathbf{a}_{[0,1]} = 10$: increase the BS b transmit power by 1 dB.
   - $\mathbf{a}_{[0,1]} = 11$: increase the BS b transmit power by 3 dB.
   - $\mathbf{a}_{[2,3]} = 00$: decrease the BS ℓ transmit power by 3 dB.
   - $\mathbf{a}_{[2,3]} = 01$: decrease the BS ℓ transmit power by 1 dB.
   - $\mathbf{a}_{[2,3]} = 10$: increase the BS ℓ transmit power by 1 dB.
   - $\mathbf{a}_{[2,3]} = 11$: increase the BS ℓ transmit power by 3 dB.
2) When $q = 1$:
   - $\mathbf{a}_{[0]} = 0$: decrease the transmit power of BS b by 1 dB.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

TABLE II
REINFORCEMENT LEARNING HYPERPARAMETERS

Parameter	Value	Parameter	Value
Discount factor γ	0.995	Exploration rate decay d	0.9995
Initial exploration rate ϵ	1.000	Minimum exploration rate $(\epsilon_{\text{min}}^{\text{voice}}, \epsilon_{\text{min}}^{\text{bf}})$	(0.15,0.10)
Number of states $\lvert\mathcal{S}\rvert$	8	Number of actions $\lvert\mathcal{A}\rvert$	16
Deep Q-Network width H	24	Deep Q-Network depth	2

TABLE III
JOINT BEAMFORMING POWER CONTROL ALGORITHM – RADIO ENVIRONMENT PARAMETERS

Parameter	Value	Parameter	Value
Base station (BS) maximum transmit power $P_{\text{BS}}^{\max}$	46 dBm	Downlink frequency band	(2100 MHz, 28 GHz)
Cellular geometry	circular	Cell radius r	(350, 150) m
Propagation model (voice, bf)	(COST231, [37])	User equipment (UE) antenna gain	0 dBi
Antenna gain $(G_{\text{TX}}^{\text{voice}}, G_{\text{TX}}^{\text{bf}})$	(11, 3) dBi	Inter-site distance R	(525, 225) m
Max. number of UEs per BS N	10	Number of multipaths N_p	(15, 4)
Probability of LOS $p_{\text{LOS}}^{\text{voice}}, p_{\text{LOS}}^{\text{bf}}$	(0.9, 0.8)	UE average movement speed v	(5, 2) km/h
Number of transmit antennas $M^{\text{voice}}, M^{\text{bf}}$	$(1,\{4, 8, 16, 32, 64\})$	Radio frame duration $T^{\text{voice}}, T^{\text{bf}}$	(20, 10) ms

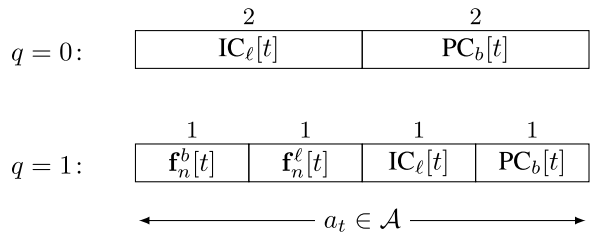

Fig. 5. Binary encoding of joint beamforming, power control, and interference coordination actions using a string of bits in a register $\mathbf{a}$ for different bearer types ($q = 0$ for voice bearers and $q = 1$ for data bearers).

- $\mathbf{a}_{[0]} = 1$: increase the transmit power of BS b by 1 dB.
- $\mathbf{a}_{[1]} = 0$: decrease the transmit power of BS ℓ by 1 dB.
- $\mathbf{a}_{[1]} = 1$: increase the transmit power of BS ℓ by 1 dB.
- $\mathbf{a}_{[2]} = 0$: step down the beamforming codebook index of BS ℓ.
- $\mathbf{a}_{[2]} = 1$: step up the beamforming codebook index of BS ℓ.
- $\mathbf{a}_{[3]} = 0$: step down the beamforming codebook index of BS b.
- $\mathbf{a}_{[3]} = 1$: step up the beamforming codebook index of BS b.

Here, we can infer that $\mathcal{P} = \{\pm 1, \pm 3\}$ dB offset from the transmit power. The choice of these values is motivated by 1) aligning with industry standards [21] which choose integers for power increments and 2) maintaining the non-convexity of the problem formulation (6) by keeping the constraints discrete. The actions to increase and decrease BS transmit powers are implemented as in (12) and (13). We introduce 3-dB power steps for voice only to compensate for not using beamforming, which is aligned with the industry standards of not having beamforming for packetized voice bearers [21].

The reward we use in our proposed algorithms is divided into two tiers: 1) based on the relevance of the action taken and 2) based on whether the target SINR has been met or the SINR falls below the minimum. We start by defining a function $\text{p}(\cdot)$ which returns one of the elements of $\mathcal{P}$ based on the chosen code. Here, $\text{p}(00) = -3, \text{p}(01) = -1, \text{p}(10) = 1, \text{p}(11) = 3$. Next, we write the received SINR due to the aforementioned encoded actions in $\mathbf{a}$ as $\gamma^b_{\mathbf{a}_{[0]},\mathbf{a}_{[3]}}$ and $\gamma^\ell_{\mathbf{a}_{[1]},\mathbf{a}_{[2]}}$ for BSs b and ℓ, respectively.

We further write the joint reward for both voice and data bearers as follows:

$$r_{s,s',a}[t;q] := \Big(\text{p}(\mathbf{a}_{[0,1]}[t]) - \text{p}(\mathbf{a}_{[2,3]}[t])\Big)(1-q) + \left(\gamma^b_{\mathbf{a}_{[0]}[t],\mathbf{a}_{[3]}[t]} + \gamma^\ell_{\mathbf{a}_{[1]}[t],\mathbf{a}_{[2]}[t]}\right) q \quad (17)$$

where $q = 0$ for voice bearers and 1 for data bearers. We reward the agent the most per time step when a joint power control and beamforming action is taken for data bearers and when a joint power control and interference coordination takes place for a voice bearer. We abort the episode if any of the constraints in (6) becomes inactive. At this stage, the RL agent receives a reward $r_{s,s',a}[t;q] := r_{\text{min}}$. Either a penalty r_{min} or a maximum reward r_{max} is added based on whether the minimum γ_{min} has been violated or γ_{target} has been achieved as shown in Algorithm 1. Here, it is also clear that for data bearers the agent is rewarded more for searching in the beamforming codebook than attempting to power up or down. However, for voice bearers, we reward the agent more if it chooses to power control the serving BS b than if it chooses to control the interference from the other BS ℓ.

In our simulations, we use a minibatch sample size of $N_{\text{mb}} = 32$ training examples. With $\lvert\mathcal{A}\rvert = 16$, the width of the DQN can be found using [31] to be $H = \sqrt{(\lvert\mathcal{A}\rvert + 2)N_{\text{mb}}} = 24$. We refer to our code [39] for details.

B. Outcomes

1) Convergence: we study the normalized convergence under (16) where $\gamma_0^{\text{bf}} = 5$ dB. Every time step in an episode is equal to one radio subframe, the duration of which is 1 ms [36]. During this time the UE is likely to be using a sub-optimal selection of beam obtained from a prior iteration. This would cause the UE throughput to degrade by a factor as we show in Section VIII-C. As the size of the ULA M increases, the number of episodes required converge increases with minimal effect of the constant threshold γ_0^{bf} since $M \gg \gamma_0^{\text{bf}}$. This is justified since the number of attempts to traverse the

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

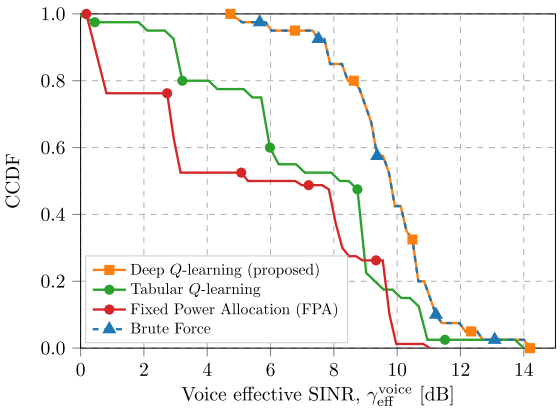

Fig. 6. Coverage CCDF plot of $\gamma_{\text{eff}}^{\text{voice}}$ for three different voice power control and interference coordination algorithms.

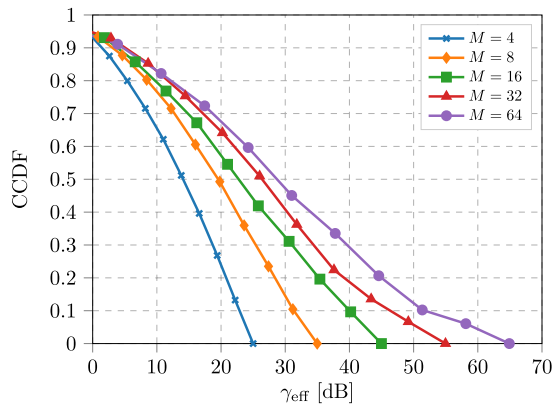

Fig. 7. Coverage CCDF plot of the effective SINR γ_{eff} for the proposed deep Q-learning algorithm vs. the number of antennas M.

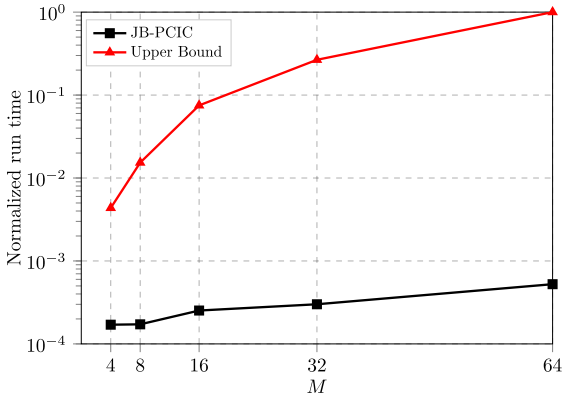

Fig. 8. The normalized run times for the proposed deep Q-learning algorithm as a function of the number of antennas M.

beamforming codebook increases almost linearly with the increase of M.

2) Run time: we study the normalized run time and observe that as the number of antennas M increase, so does the run-time complexity for the proposed algorithm. This is justified due to the increase in the number of beams required for the algorithm to search through to increase the joint SINR.
3) Coverage: for voice bearers we observe that the coverage as defined by the SINR CCDF improves everywhere. For data bearers, the coverage improves where the SINR monotonically increases with the increase in M which is expected because the beamforming array gain increases with an increase in M.
4) Sum-rate: the sum-rate capacity increases logarithmically as a result of the increase of M, which is justified using (5) and (15).

C. Figures

Fig. 6 shows the CCDF of the effective SINR γ_{eff} for the voice PCIC algorithms all for the same episode. This episode generates the highest reward. Here we see that the FPA algorithm has the worst performance especially at the cell edge (i.e., low effective SINR regime), which is expected since FPA has no power control or interference coordination. The tabular implementation of our proposed algorithm has better performance compared with the FPA. This is since power control and interference coordination are introduced to the base stations, though not as effectively, which explains why close to $\gamma_{\text{eff}} = 9$ dB tabular Q-learning PCIC underperforms FPA. Further, we observe that deep Q-learning outperforms the tabular Q-learning implementation of the PCIC algorithm, since deep Q-learning has resulted in a higher reward compared to tabular Q-learning. This is because deep Q-learning has converged at a better solution (identical to the solution obtained through brute force), unlike the tabular Q-learning the convergence of which may have been impeded by the choice of a initialization of the state-action value function. However, as the effective SINR γ_{eff} approaches 13 dB, the users are close to the BS center and therefore all power control algorithms perform almost similarly thereafter.

We show the coverage CCDF in Fig. 7. As M increases, so does the probability of achieving a given effective SINR, since the effective SINR depends on the beamforming array gain which is a function of M as stated earlier. The improvement in the run time is shown in Fig. 8. The brute force algorithm has a significantly larger run time compared to the proposed algorithm. The run time increases as the number of antennas M increase, though much steeper in the brute force algorithm, due to the exponential nature of the run-time complexity. At $M = 4$, only 4% of the run time of the brute force algorithm was needed for our proposed algorithm. In Fig. 9, at smaller ULA sizes M, the impact of the constant threshold γ_0^{bf} becomes dominant and it takes almost similar times to converge for values of M. This is likely to be due to the wider beams in the grid of beams, which are able to cover the UEs moving at speeds v. However, for the large antenna size regime, as the size of the ULA M increases, the number of episodes required converge increases with minimal effect of γ_0^{bf} as we explained earlier. This is due to the longer time required for the agent to search through a grid of beams of size $|\mathcal{F}|$, which are typically narrower at large M. This causes the agent to spend longer time to meet the target SINR. This time or delay is linear in M as we expect based on Section VI since $|\mathcal{F}|$ is linear in M. This delay can have a negative impact on the throughput and voice frames of the data and voice bearers respectively. If we assume the data bearer transmits b bits over a total duration of T^{bf} for beamformed

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

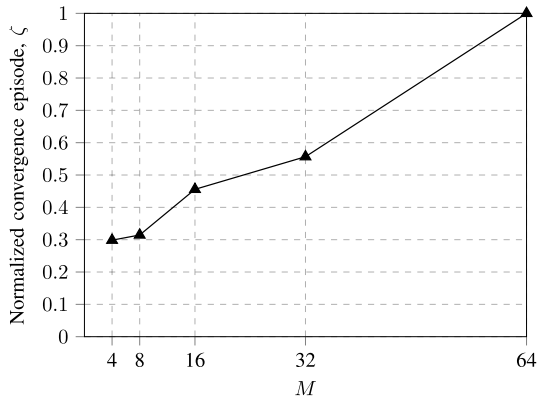

Fig. 9. The normalized convergence time for the proposed deep Q-learning algorithm as a function of the number of antennas M.

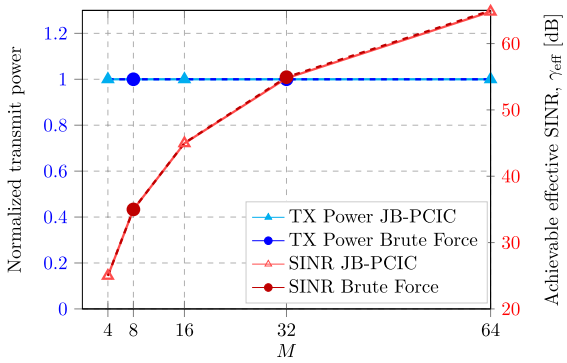

Fig. 10. Achievable SINR and normalized transmit power for both the brute force and proposed JB-PCIC algorithms as a function of the number of antennas M.

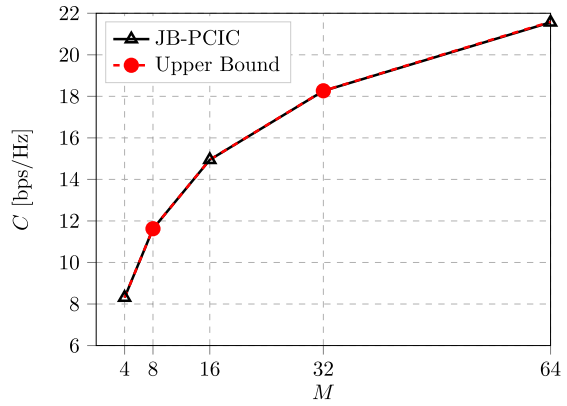

Fig. 11. Sum-rate capacity of the convergence episode as a function of the number of antennas M.

data bearers, then the impact of the convergence time would cause these b bits to be transmitted over a duration of $T^{\text{bf}}\zeta$. The throughput due to convergence then becomes $b/T^{\text{bf}}\zeta$. For voice, the number of lost voice frames due to this convergence time is $\lceil \nu\zeta \rceil$.

The achieved SINR is proportional to the ULA antenna size M as shown in Fig. 10. This is expected as the beamforming array gain is $\|\mathbf{f}_b\|^2 \leq M$. The transmit power is almost equal to the maximum. Fig. 10 also shows the relative performance of JB-PCIC compared with the brute force performance outlined in Section VI-B. We observe that the performance gap of both the transmit power of the base stations and the SINR is almost diminished all across M. This is because of the DQN ability to estimate the function that leads to the upper limit of the performance. Further, we observe that the solution for the race condition is for both BSs to transmit at maximum power.

Finally, Fig. 11 shows the sum-rate capacity of both the JB-PCIC algorithm and the upper limit of performance. Similarly, the performance gap diminishes across all M for the same reason discussed earlier.

IX. Conclusion

In this paper, we sought to maximize the downlink SINR in a multi-access OFDM cellular network from a multi-antenna base station to single-antenna user equipment. The user equipment experienced interference from other multi-antenna base stations. Our system used sub-6 GHz frequencies for voice and mmWave frequencies for data. We assumed that each base station could select a beamforming vector from a finite set. The power control commands were also from a finite set. We showed that a closed-form solution did not exist, and that finding the optimum answer required an exhaustive search. An exhaustive search had a run time exponential in the number of base stations.

To avoid an exhaustive search, we developed a joint beamforming, power control, and interference coordination (JB-PCIC) algorithm using deep reinforcement learning. This algorithm resides at a central location and receives UE measurements over the backhaul. For voice bearers, our proposed algorithm outperformed both the tabular Q-learning algorithm and the industry standard fixed power allocation algorithm.

Our proposed algorithm for joint beamforming, power control and interference coordinations requires that the UE sends its coordinates and its received SINR every millisecond to the base station. The proposed algorithm, however, does not require the knowledge of the channel state information, which removes the need for channel estimation and the associated training sequences. Moreover, the overall amount of feedback from the UE is reduced because the UE sends its coordinates and would not need to send explicit commands for beamforming vector changes, power control, or interference coordination.

References

[1] F. B. Mismar and B. L. Evans, "Q-learning algorithm for VoLTE closed loop power control in indoor small cells," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018.

[2] S. Yun and C. Caramanis, "Reinforcement learning for link adaptation in MIMO-OFDM wireless systems," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2010, pp. 1–5.

[3] M. Bennis and D. Niyato, "A Q-learning based approach to interference avoidance in self-organized femtocell networks," in *Proc. IEEE Globecom Workshops*, Dec. 2010, pp. 706–710.

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

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

[6] C. Luo, J. Ji, Q. Wang, L. Yu, and P. Li, "Online power control for 5G wireless communications: A deep Q-network approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[7] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Commun.*, vol. 46, no. 10, pp. 1313–1324, Oct. 1998.

[8] *Evolved Universal Terrestrial Radio Access (E-UTRA); Overall Description*, document TS 36.300, 3GPP, Jan. 2019. [Online]. Available: http://www.3gpp.org/dynareport/36300.htm

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.

[9] R. Kim, Y. Kim, N. Y. Yu, S.-J. Kim, and H. Lim, "Online learning-based downlink transmission coordination in ultra-dense millimeter wave heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2200–2214, Apr. 2019.
[10] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.
[11] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.
[12] H. S. Jang, H. Lee, and T. Q. S. Quek, "Deep learning-based power control for non-orthogonal random access," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 2004–2007, Aug. 2019.
[13] M. K. Sharma, A. Zappone, M. Debbah, and M. Assaad, "Deep learning based online power control for large energy harvesting networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 8429–8433.
[14] W. Lee, M. Kim, and D.-H. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.
[15] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.
[16] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, May 2019. [Online]. Available: https://arxiv.org/abs/1905.03761
[17] T. Maksymyuk, J. Gazda, O. Yaremko, and D. Nevinskiy, "Deep learning based massive MIMO beamforming for 5g mobile network," in *Proc. IEEE Int. Symp. Wireless Syst.*, Sep. 2018, pp. 241–244.
[18] F. B. Mismar, J. Choi, and B. L. Evans, "A framework for automated cellular network tuning with reinforcement learning," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7152–7167, Oct. 2019.
[19] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, "Deep learning-based beam management and interference coordination in dense mmWave networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 592–603, Jan. 2019.
[20] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, early access, Dec. 2019, doi: 10.1109/TCOMM.2019.2960361.
[21] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TS 36.213, 3GPP, Dec. 2015. [Online]. Available: http://www.3gpp.org/dynareport/36213.htm
[22] F. B. Mismar and B. L. Evans, "Deep learning in downlink coordinated multipoint in new radio heterogeneous networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1040–1043, Aug. 2019.
[23] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[24] R. W. Heath, Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
[25] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter wave communications: The sparse way," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014.
[26] T. S. Rappaport, F. Gutierrez, Jr., E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.
[27] T. S. Rappaport, R. W. Heath, Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. London, U.K.: Pearson, 2014.
[28] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," in *Proc. NIPS Deep Learn. Workshop*, 2013. [Online]. Available: http://arxiv.org/abs/1312.5602
[29] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1998.
[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2016.
[31] 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, Mar. 2003.
[32] L.-J. Lin, "Reinforcement learning for robots using neural networks," Ph.D. dissertation, School Comput. Sci., Carnegie-Mellon Univ., Pittsburg, PA, USA, 1993.
[33] M. Simsek, A. Czylwik, A. Galindo-Serrano, and L. Giupponi, "Improved decentralized Q-learning algorithm for interference reduction in LTE-femtocells," in *Proc. Wireless Adv.*, Jun. 2011.
[34] D. Silver. (2015). *Advanced Topics–Reinforcement Learning*. [Online]. Available: http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html
[35] F. B. Mismar and B. L. Evans, "Partially blind handovers for mmWave new radio aided by sub-6 GHz LTE signaling," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2018, pp. 1–5.
[36] *NR; Physical Channels and Modulation*, document TS 38.211, 3GPP, Jun. 2018.
[37] A. I. Sulyman, A. Alwarafy, G. R. MacCartney, T. S. Rappaport, and A. Alsanie, "Directional radio propagation path loss models for millimeter-wave wireless networks in the 28-, 60-, and 73-GHz bands," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6939–6947, Oct. 2016.
[38] T. Bai and R. W. Heath, Jr., "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.
[39] F. B. Mismar. (2019). *Source Code*. [Online]. Available: https://github.com/farismismar/Deep-Reinforcement-Learning-for-5G-Networks

Faris B. Mismar (Senior Member, IEEE) received the B.S. degree in electrical and computer engineering from The University of Jordan in 2004, the M.S. degree in electrical engineering from The University of Texas at Dallas in 2011, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin in 2019. He was a consultant with many leading wireless operators across the globe. He held various senior positions at Motorola, Ericsson, and Samsung. He has been working in the wireless communication industry since 2004. His research interests include machine learning, artificial intelligence, and wireless communications.

Brian L. Evans (Fellow, IEEE) received the B.S. degree in electrical engineering and computer science from the Rose-Hulman Institute of Technology in 1987, and the M.S. and the Ph.D. degrees in electrical engineering from the Georgia Institute of Technology in 1993 and 1988, respectively. From 1993 to 1996, he was a Post-Doctoral Researcher with the University of California, Berkeley. In 1996, he joined the Faculty of The University of Texas at Austin (UT Austin). He is an Engineering Foundation Professor of electrical and computer engineering with UT Austin. He has published 270 refereed conference and journal articles, and graduated 29 Ph.D. and 13 M.S. students. His recent projects have included cloud radio access networks, image quality assessment, smart phone video acquisition, and wireless interference mitigation. His research and teaching interests include signal processing theory and algorithms to increase connection speeds and reliability in communication systems. His research group develops algorithms with implementation constraints in mind and translates algorithms into design methods and embedded prototypes. His most recent research efforts are focused on multiantenna communication systems. He has been awarded three best/top paper awards at IEEE conferences and five teaching awards at UT Austin. He has received the 1997 U.S. National Science Foundation CAREER Award.

Ahmed Alkhateeb received the B.S. degree (Hons.) and the M.S. degree in electrical engineering from Cairo University, Egypt, in 2008 and 2012, respectively, and the Ph.D. degree in electrical engineering from The University of Texas at Austin, USA, in 2016. From 2016 to 2017, he was a Wireless Communications Researcher with the Connectivity Lab, Facebook Inc., Menlo Park, CA, USA. He has held research and development internship positions at FutureWei Technologies, Huawei, Chicago, IL, USA, and Samsung Research America, Dallas, TX, USA. He joined Arizona State University in 2018, where he is currently an Assistant Professor with the School of Electrical, Computer and Energy Engineering. His research interests include wireless communications, communication theory, signal processing, machine learning, and applied maths. He was a recipient of the 2012 MCD Fellowship from The University of Texas at Austin and the 2016 IEEE Signal Processing Society Young Author Best Paper Award for his work on hybrid precoding and channel estimation in millimeter wave communication systems.

Authorized licensed use limited to: UCLA Library. Downloaded on March 04,2024 at 18:32:54 UTC from IEEE Xplore. Restrictions apply.