Hierarchial-DQN Position-Aided Beamforming for Uplink mmWave Cellular-Connected UAVs

Praneeth Susarla*, Yansha Deng†, Markku Juntti*, Olli Sílven*
*University of Oulu, Finland
†King's College London, United Kingdom

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001044

***Abstract*—Unmanned aerial vehicles (UAVs) are the vital components of sixth generation (6G) millimeter wave (mmWave) wireless networks. Fast and reliable beam alignment is essential for efficient beam-based mmWave communications between UAVs and the base stations (BSs). Learning-based approaches may greatly reduce the overhead by leveraging UAV data, such as position, to identify the optimal beam directions. In this paper, we propose a deep reinforcement learning (DRL)-based framework for UAV-BS beam alignment using the hierarchical deep Q-Network (hDQN) in a mmWave radio setting. We consider uplink communications where the UAV hovers around 5G new radio (NR) BS coverage area, with three dimensional (3D) beams under diverse channel conditions. A BS serves with learnt beam-pairs in an uplink manner upon every communication request from UAV inside the multi-location environment. Compared to our prior DQN-based method, the proposed hDQN framework uses the location information and the fixed spatial arrangement of the antenna elements to reduce the beam search complexity and maximize the data rates efficiently. The results show that our proposed hDQN-based framework converges faster than the DQN-based approach with an average overall training reduction of 43% and, is generic to multi-location environments across different uniform planar array (UPA) configurations and diverse channel conditions.**

***Index Terms*—6G, 5G and beyond, mmWave, hierarchical Beam alignment, Deep Q-Network**

I. Introduction

Position information can be leveraged for fast beam alignment in the upcoming sixth generation (6G) millimeter wave (mmWave) communications. Such information is widely available in base station (BS)-unmanned aerial vehicle (UAV) communications, from sensors at lower frequencies mounted on UAV such as global positioning system (GPS), camera, lidar etc. Also, the 6G radio wave directionality obtained through mmWave frequencies and multiple input multiple output (MIMO) beamforming enable high speed data access and line-of-sight (LoS) dominant connectivity to unmanned aerial vehicles (UAVs), envisioned as future cellular networks. Especially, the deployment of cellular-enabled UAV-user equipments (UE) (*hereafter addressed as UAVs*) adds unique features pertaining to high mobility and autonomous operations traffic surveillance, mineral exploration, internet drone delivery systems, etc. [1].

However, the UAVs bring additional challenges and requirements to the existing terrestrial networks, including reliability, low-latency communications, interference during aerial-ground communications, etc. Aerial-ground interference management of UAVs has been extensively studied in BS-UAV communications over recent years [2], [3]. For reliability and low-latency challenges, the UAV position information with flexible three-dimensional (3D) beamforming can be effectively used to enhance their data throughput through fast mmWave beamalignment in forthcoming 6G systems.

Position information has been used for fast mmWave beam alignment in vehicular communications [4]–[7]. In [4], Va *et al.* proposed an inverse fingerprinting approach by exploiting vehicular position information during beam training in non-line-of-sight (nLoS) conditions. The authors in [6], [7] proposed a learning-based beam training schemes using multi-armed bandit (MAB) approach, by building a database of finite beam-pairs useful for beam training based on vehicular position information. Their key idea here is that the machine learning (ML)-based approaches can effectively use the position information for fast mmWave beamalignment in an online manner.

High mobility and autonomous operation of UAVs will also require frequent beam realignment and can be jointly optimized effectively using reinforcement learning (RL)-based beam training [8]. In our previous work [9], we proposed a context-information aided beam pair alignment problem for cellular-connected mmWave UAV networks using deep reinforcement learning (DRL) technique like deep Q-network (DQN) at the BS using the UAV position information. We have shown that a generic DQN framework can enhance the beamforming gains in an online manner under different 3rd generation partnership project (3GPP) conditions. However, the proposed DQN-based approach is slightly impractical for uniform planar array (UPA) antenna configurations due to their large action spaces. DQN action space increase exponentially with increase of beam pairs for large antenna elements affecting the convergence due to curse of dimensionality phenomena. Also, the work assumed independent and fixed grid element in the BS-UAV environment.

In this paper, we model the BS-UAV beam pair alignment problem using hierarchical deep Q-network (hDQN) with the aim to reduce the beam search complexity for UPA configurations.The fixed spatial arrangement of the antenna elements can be effectively exploited alongside the UAV position information by considering mmWave beams with different beam width resolutions in a hierarchical manner during beam-training. Our simulations show that the hDQN approach reduces the beam training overhead by 43% from

our prior DQN method.

The rest of the paper is organized as follows. Section II-A and II-B presents the problem formulation and communication modelling, considered in this problem, respectively. Section II-C and Section III discuss in detail the problem formulation of the proposed hierarchical deep reinforcement learning (hDRL) approach, namely hDQN and its algorithmic implementation for BS-UAV beam pair alignment problem. Section IV presents the comparison of the proposed hDQN approach against our previous DQN-based method under different UPA antenna configurations and channel conditions. Section V summarizes the conclusion and future work.

II. System and Communication Model

A. System Model

As shown in Figure 1, we consider a cellular mmWave MIMO uplink communication with BS serving multiple UAVs in a time domain multiple access (TDMA) manner under its spherical coverage area. The BS is fixed at $\mathcal{O}(0,0,h_{\mathrm{BS}}) \in \mathbb{R}^3$ and communicates with the moving UAV (hereafter used as user equipment (UE)) using a multi-path mmWave beamforming. The UE moves randomly around the BS 3D spherical coverage area composed of multiple grids, the set enclosing them is denoted as $\mathbb{U}$. Following the 3D spherical coordinate system, let ξ_h, θ_h, ϕ_h represent the radial distance, elevation and azimuthal angles of grid index $h \in \{0,1,...|\mathbb{U}|\}$ with respect to BS, then $\mathrm{UE}_h(t)$ in cartesian form at any time instant t is given by

$$\mathrm{UE}_h(t) = (\xi_h \sin\theta_h \cos\phi_h, \xi_h \sin\theta_h \sin\phi_h, \xi_h \cos\theta_h), \quad (1)$$

The UE transmits (TX) while the BS receives (RX) a radio signal in multiple beam directions following $\mathcal{W}$ and $\mathcal{F}$ predefined analog codebook directions, respectively. The BS usually has more antenna elements compared to UE and hence, we assume a hierarchical multi-resolution and the narrowest possible angular resolution codebooks at $\mathcal{F}$ and $\mathcal{W}$, respectively. The hierarchical codebook at $\mathcal{F}$ follows a joint subarray and deactivation approach proposed in [10].

The communication begins with a TX request from UE_h while the RX radio unit at BS starts with a random beam-pair at time $t = 0$ and learns to choose the beam-pair direction $(b_p, b_q), b_p \in \mathcal{W}, b_q \in \mathcal{F}$ over time for the grid position with index $h \in \{0,1,...|\mathbb{U}|\}$. The BS receives the initial radio beam b_q at broader angular-resolution level from $\mathcal{F}$ and then switch to narrow radio-beams over time, to reduce the beam search space and still achieve efficient beamforming gains for UPA antenna configurations. Here, we assume the moving UE transmit radio signals in the same narrow beam directions within each grid position. Thus, the BS selects a sequence of beam-pair directions for TX and RX, with every change in grid position as the substantial change in TX location induces a variance in their radio measurements, following 3GPP fifth generation (5G) new radio (NR) beam alignment protocol [11].

The 3GPP 5G NR beam alignment protocol for physical layer consists of initial communication (used as P_1 procedure), beam selection (used as P_2 procedure) and an optional beam refinement (used as P_3 procedure) [11]. Herein, we consider BS and UE following P_1 and P_2 procedures at every grid position, along the coverage area set $\mathbb{U}$. During P_1 procedure, the UE is assumed to send a communication request with respect to its position, while the learning framework at BS responds with a hierarchical sequence of radio beam-pairs to be considered for next phase of uplink based beam access protocol. P_2 generally implies the radio beam selection procedure at mmWave frequencies later used for the data transmission [11]. Similar to the works in [7]–[9], the BS and UE in P_2 are assumed to undergo the beam-training procedure following the sequence of beam-pairs configured by the BS-side learning framework from initial communication procedure.

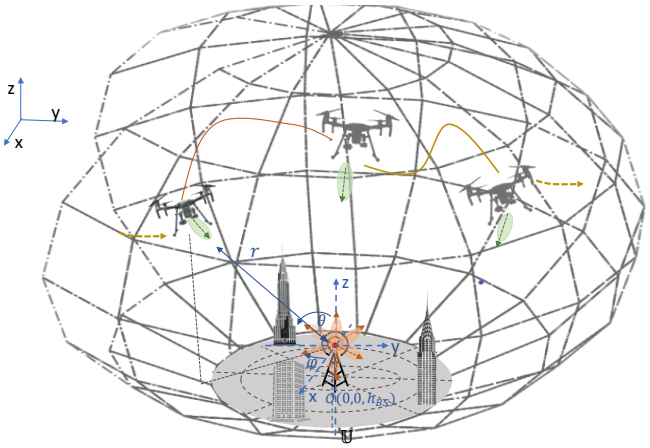

Figure 1: System model.

The received signal measurement can be observed at the BS for different TX-RX beam pairs during these procedures and their timing information can be estimated using 5G protocol frame structure [11]. We define travel time unit (TTU) as the orthogonal frequency division multiple access (OFDM) symbol time during every beam transmission or reception from the 5G frame structure. In this work, we use this definition to measure the communication overhead for the learning-based beam sweeping procedure in TTU units.

B. Communication Model

We consider a multi-path link (LoS or nLoS) radio channel between UE at time t and BS location $\mathcal{O} \in \mathbb{R}^3$. The BS and UE are equipped with single radio frequency (RF) chains of $(N_x^{\mathrm{rx}}, N_y^{\mathrm{rx}})$ receive and $(N_x^{\mathrm{tx}}, N_y^{\mathrm{tx}})$ transmit antennas respectively. As the BS serves multiple UEs in a TDMA manner, we model the communication between a single UE and single BS with UPA for the urban macro-cellular (UMa) environments [11]. We assume each UPA beam at both BS and UE projected with azimuthal ϕ and elevation θ main lobe broadside direction. Let M denote the number of multi-paths or reflection points in the environment, the channel matrix corresponding to the m^{th} path is given by

$$\mathbf{H}_m \triangleq \beta_m \mathbf{a}_R(\theta_{\mathrm{m}}^{\mathrm{rx}}, \phi_{\mathrm{m}}^{\mathrm{rx}}) \mathbf{a}_T^H(\theta_{\mathrm{m}}^{\mathrm{tx}}, \phi_{\mathrm{m}}^{\mathrm{tx}}) \quad (2)$$

where, β_{m} is the antenna channel gain, $\theta_{\mathrm{m}}^{\mathrm{tx}}$, $\theta_{\mathrm{m}}^{\mathrm{rx}}$ are the azimuthal angle of departure (AoD) and angle of arrival (AoA), $\phi_{\mathrm{m}}^{\mathrm{tx}}$, $\phi_{\mathrm{m}}^{\mathrm{rx}}$ are the elevation AoD and AoA of m^{th} communication link between BS and UE. $\mathbf{a}_R(\theta_{\mathrm{m}}^{\mathrm{rx}}, \phi_{\mathrm{m}}^{\mathrm{rx}}) \in \mathbb{C}^{N_x^{\mathrm{rx}} N_y^{\mathrm{rx}}}$, $\mathbf{a}_T(\theta_{\mathrm{m}}^{\mathrm{tx}}, \phi_{\mathrm{m}}^{\mathrm{tx}}) \in \mathbb{C}^{N_x^{\mathrm{tx}} N_y^{\mathrm{tx}}}$ are the antenna array steering vectors for $(\theta_{\mathrm{m}}^{\mathrm{rx}}, \phi_{\mathrm{m}}^{\mathrm{rx}})$ and $(\theta_{\mathrm{m}}^{\mathrm{tx}}, \phi_{\mathrm{m}}^{\mathrm{tx}})$, respectively. Let $\omega_x = \frac{2\pi}{\lambda} d_x \sin\theta\cos\phi$, $\omega_y = \frac{2\pi}{\lambda} d_y \sin\theta\sin\phi$, λ is the wavelength, $\otimes$ denote the Kronecker product, N_x and N_y are the antenna elements along x and y-axis, d_x and d_y are the antenna element spacing in x and y-direction, respectively. Then, the array steering vector is given by

$$a(\theta,\phi) = \frac{1}{\sqrt{N_x N_y}} \begin{bmatrix} 1 \\ e^{j\omega_y} \\ \vdots \\ e^{j(N_y-1)\omega_y} \end{bmatrix} \otimes \begin{bmatrix} 1 \\ e^{j\omega_x} \\ \vdots \\ e^{j(N_x-1)\omega_x} \end{bmatrix} \tag{3}$$

where $(\theta,\phi) = (\theta_{\mathrm{m}}^{\mathrm{rx}}, \phi_{\mathrm{m}}^{\mathrm{rx}}), (N_x, N_y) = (N_x^{\mathrm{rx}}, N_y^{\mathrm{rx}})$ and $(\theta,\phi) = (\theta_{\mathrm{m}}^{\mathrm{tx}}, \phi_{\mathrm{m}}^{\mathrm{tx}}), (N_x, N_y) = (N_x^{\mathrm{tx}}, N_y^{\mathrm{tx}})$ for $\mathbf{a}_R(\theta_{\mathrm{m}}^{\mathrm{rx}}, \phi_{\mathrm{m}}^{\mathrm{rx}})$ and $\mathbf{a}_T(\theta_{\mathrm{m}}^{\mathrm{tx}}, \phi_{\mathrm{m}}^{\mathrm{tx}})$, respectively. For a unit-norm transmit and receive beamforming vectors namely, $\mathbf{w}_k \in \mathbb{C}^{N_x^{\mathrm{tx}} N_y^{\mathrm{tx}}}$ and $\mathbf{f}_k \in \mathbb{C}^{N_x^{\mathrm{rx}} N_y^{\mathrm{rx}}}$, baseband equivalent of the received signal at discrete symbol time k is given by

$$y_k = \underbrace{\sum_{m=0}^{M} \sqrt{P_{tx}} \mathbf{f}_k^H \mathbf{H}_m \mathbf{w}_k x_k}_{r_k} + \nu_k, \tag{4}$$

where P_{tx} is transmission power, $\nu_k \sim \mathcal{CN}(0, WN_0)$ is the effective noise with zero mean and two-sided power spectral density $\frac{N_0}{2}$, x_k represents one OFDM symbol of the time-domain transmitted signal with bandwidth W and TTU time period with $\frac{1}{K}\sum_{k=0}^{\mathrm{K}} \|x_k\|^2 = 1$. Here, we assume $\mathbf{H}_m$ to follow 3GPP UMa conditions [11] and $k = 0, 1, ...K$ denotes the number of samples spanned over TTU time. $\mathbf{w}_k$ and $\mathbf{f}_k$ for UPA beams are measured using (3) for selected codebook directional pairs (θ_k, ϕ_k) from $\mathcal{W}$ and $\mathcal{F}$, respectively. Similar to our previous work on linear codebook direction sets [9] and following (3), the UPA radio beam directions are determined by linear array angular resolutions namely, $\frac{2}{N_x}$ and $\frac{2}{N_y}$ with their physical angles $(-\frac{\pi}{2}, \frac{\pi}{2})$ along x and y direction, respectively. Thus, we assume that the UE transmits radio signals with a narrowest angular resolutions in $\mathcal{W}$ codebook directions while the BS receives the signal through one of its hierarchical multi-angular resolution codebook directions from $\mathcal{F}$.

The hierarchical directional set $\mathcal{F}$ consists of L (for example, $0 \le l \le L, L = \log_2(N_x^{\mathrm{rx}})$ along x-axis) multiple angular resolution levels along x and y directions separately, with the l^{th} level codebook directional subset $\mathcal{F}_l = \{f_1^{(l)}, f_2^{(l)}, ..., f_{\mathbf{card}(\mathcal{F}_l)}^{(l)}\}$ designed to uniformly cover all the spatial frequency range $(-1, 1)$ (physical angles $(-\frac{\pi}{2}, \frac{\pi}{2})$) along x and y-directions separately and satisfy the relation $\mathbf{card}(\mathcal{F}_1) < ... < \mathbf{card}(\mathcal{F}_L)$ as shown in Figure 2. Here, $\mathbf{card}(\mathcal{F})$ denotes the cardinality of $\mathcal{F}$. For this work, we consider a two level angular-resolution subsets at RX, namely, $\mathcal{F}_\mathcal{B}$ and $\mathcal{F}_\mathcal{N}$ ($\mathbf{card}(\mathcal{F}_\mathcal{B}) < \mathbf{card}(\mathcal{F}_\mathcal{N})$) with their beam widths $\psi = \frac{2}{\mathbf{card}(\mathcal{F})}$, where $\psi = \psi^\mathcal{B}$ and $\psi = \psi^\mathcal{N}$ ($\psi^\mathcal{B} > \psi^\mathcal{N}$) for $\mathcal{F}_\mathcal{B}$ and $\mathcal{F}_\mathcal{N}$, respectively. We select $(\psi_x^\mathcal{N}, \psi_y^\mathcal{N}) = (\frac{2}{N_x^{\mathrm{rx}}}, \frac{2}{N_y^{\mathrm{rx}}})$ and $(\psi_x^\mathcal{B}, \psi_y^\mathcal{B}) = (\frac{2^{(L-l+1)}}{N_x^{\mathrm{rx}}}, \frac{2^{(L-l+1)}}{N_y^{\mathrm{rx}}}), l \in [0, L)$. We assume an elliptical surface for every rectangular grid element in $\mathbb{U}$ and is proportional to $\psi^\mathcal{B}$ given by $(\psi_x^\mathcal{B}, \psi_y^\mathcal{B}) = (\eta_\theta cos\theta_0, \eta_\phi)$ where η_θ and η_ϕ are the elevation and azimuthal angular resolution in $\mathbb{U}$, respectively [12]. θ_0 is the elevation angle of a UE grid element $g, g \in \mathbb{U}$ from the BS. Thus, with $\psi^\mathcal{B}$ selection, η_θ and η_ϕ can be chosen to favour a single broad beam projection from BS, for every grid element in $\mathbb{U}$. We define $r_k = \sum_{m=0}^{M} \sqrt{P_{tx}} \mathbf{f}_k^H \mathbf{H}_m \mathbf{w}_k x_k$. Then, the signal-to-noise ratio (SNR) is given as $\mathrm{SNR} = \frac{\frac{1}{K}\sum_{k=0}^{K} \|r_k\|^2}{N_0 W}$ and overall rate measurement R in bits per channel use is given by

$$R = \log(1 + \mathrm{SNR}). \tag{5}$$

Thus, the optimal beam-pair for UE-BS during P_1 procedure is selected based on the data rate measurements.

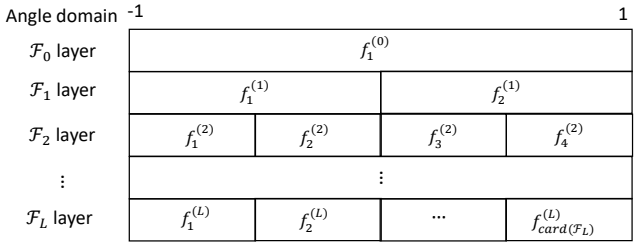

Figure 2: Beam coverage of a hierarchical beam structure codebook.

C. Problem Formulation

We consider an uplink communication between BS and UE following 3GPP beam access protocol [11]. We formulate the learning-based beam-pair alignment as a partially observable Markov decision process (POMDP) during P_1 and P_2 procedures, and maximize the beamforming gain for any UE position around the BS coverage area $\mathbb{U}$. We consider received signal strength (RSS) of radio signal and radio beam pair directions (both TX and RX) as the known and unknown parameters of this multi-location environment, respectively.

In this work, we model an interactive RL-based beam-pair alignment problem as a POMDP. At any time instant t, we define the parameters $s_t = \{(\mathrm{UE}_h, b_r, b_s, b_u), \mathrm{UE}_h \in \mathbb{U}, b_r \in \mathcal{W}, b_s \in \mathcal{F}_\mathcal{B}, b_u \in \mathcal{F}_\mathcal{N}\}$, $a_t^\mathcal{B} = \{(b_p, b_q), b_p \in \mathcal{W}, b_q \in \mathcal{F}_\mathcal{B}\}$, $a_t^\mathcal{N} = \{(b_m, b_n), b_m \in \mathcal{W}, b_n \in \mathcal{F}_\mathcal{N}\}$ where s_t, $a_t^\mathcal{B}$, $a_t^\mathcal{N}$, r_t are the state, broad beam-pair action, narrow beam-pair action and reward at time instant t. Data rate measurements computed for each applied action are considered as the rewards for the problem. As shown in Figure 2, every broad beam codebook direction $f_c^{(l)}$ in $\mathcal{F}_\mathcal{B}$ comprise of a finite subset of narrow beam codebook directions $\{f_c^{(l)} - \frac{\psi^\mathcal{B}}{2} \le f_d^{(L)} \le f_c^{(l)} + \frac{\psi^\mathcal{B}}{2}, f_d^{(L)} \in \mathcal{F}_\mathcal{N}\}$ with cardinality defined as V. Let π_1 and π_2 denote the broad beam action and narrow beam action policies for state transitions $(s_t, a_t^\mathcal{B}, r_{t+V}, s_{t+V})$ and $(s_t, a_t^\mathcal{N}, r_t, s_{t+1})$, respectively. After UE's P_1 procedure, BS

starts with a random receiving beam direction and then proceeds towards the maximum beamforming gain by applying actions and undergoing state transitions, accordingly. The current applied action becomes part of the next state, undergoing state transition. We define an episode e_{π_1,π_2} as the consecutive set of such broad beam and narrow beam actions until the terminal state following policies π_1 and π_2, respectively. The objective of this problem for broad beam and narrow beam actions can be formulated as

$$\begin{aligned}(P1): &\max_{\{\pi_1(a_t^{\mathcal{B}})\}} \sum_{t\le i<\infty} \gamma^{(i-t)V}\mathbb{E}_\pi[r(a_{iV}^{\mathcal{N}})],\\ (P2): &\max_{\{\pi_2(a_t^{\mathcal{N}})\}} \sum_{t\le i<\infty} \gamma^{i-t}\mathbb{E}_\pi[r(a_i^{\mathcal{N}})],\\ &\text{s.t.}\\ r(a_t^{\mathcal{N}}) &= \begin{cases}1 & \text{if } R(a_t^{\mathcal{N}}) \ge R_{max}(s_t)\\ -1 & \text{otherwise}\end{cases},\\ &\gamma \in (0,1],\end{aligned} \tag{6}$$

where $R_{max}(s_t)$ is the optimal data rate measurement observed among the information history o_t until its previous episode e_{π_1,π_2}, $r(a_t^{\mathcal{N}})$ and $R(a_t^{\mathcal{N}})$ are the rewards and data rate measurement observed on applying action beam-pair $a_t^{\mathcal{N}}$, respectively. We maximize the objective formulation by learning the hierarchical sequence of beam-pair actions starting with broad beam level selection from $\mathcal{F}_{\mathcal{B}}$ and switch to narrow-beam level selection from $\mathcal{F}_{\mathcal{N}}$ following the same reward function (6). We consider a hDQN approach to solve this objective problem.

III. Hierarchical DQN-based beam alignment

DQN is a value-based approach, learning an optimal approximated policy of states mapping to actions $\pi(s) = a$ by parameterizing and estimating state-action value function $Q(s, a; \theta)$ where θ denotes the weight matrix of the primary deep neural networks (DNN) [13]. The hDQN framework integrates hierarchical action-value functions operating at different temporal scales using DQN approach and learns optimal approximated policies $\pi_1(s) = a, a \in \mathcal{A}_{\mathcal{B}}$ and $\pi_2(s) = a, a \in \mathcal{A}_{\mathcal{N}}$, respectively [14], [15]. Under our hDQN framework, we consider a broad beam (BB) and narrow beam (NB) DQN agents over the same state space $\mathcal{S}$ but different action spaces $\mathcal{A}_{\mathcal{B}}$ and $\mathcal{A}_{\mathcal{N}}$, respectively as shown in Figure 3.

For the BB agent, we denote the primary DNN network weight matrix and target DNN network weight matrix as θ_1 and $\overline{\theta_1}$, respectively [13]. We consider a fully connected DNN for both the networks where $\overline{\theta_1}$ is updated with primary network parameters θ_1, after every K_1 iterations. The input of DNN is given by the variables in s_t. The intermediate layers are fully connected linear units with rectifier linear units (ReLU) and the output layer is composed of linear units in one-one correspondence with $\mathcal{A}_{\mathcal{B}}$. We consider both DNNs with zeros initialization bias and Kaiming normalization weights. A memory buffer of experiences $D_1 = \{e_1, e_2, e_3, ..., e_t\}$, $e_i = (s_i, a_i^{\mathcal{B}}, r_{i+V}, s_{i+V})$ are collected, where a mini batch of them $U(D_1)$ are randomly sampled and sent into BB-DQN [13]. For the NB agent, we follow the same procedure with network weight parameters as θ_2, $\overline{\theta_2}$, target network updated every K_2 iterations, the output layer mapped to $\mathcal{A}_{\mathcal{N}}$ with disjoint (from D_1) memory buffer of experiences as $D_2 = \{e'_1, e'_2, e'_3, ..., e'_t\}$ and collected transitions as $e'_i = (s_i, a_i^{\mathcal{N}}, r_{i+1}, s_{i+1})$ respectively.

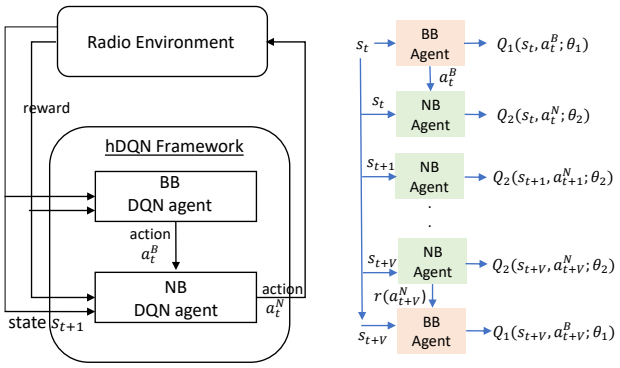

Figure 3: hDQN framework with BB and NB DQN agents

Let $Q_1(s, a; \theta_1)$ and $Q_2(s, a; \theta_2)$ denote the state-action value functions of BB and NB agents, respectively as shown in Figure 3. For both the DQN agents, a mean squared error (MSE) loss function is computed between primary, target networks during back propagation, and θ is updated using stocastic gradient descent (SGD) and Adam Optimizer as

$$\theta_{t+1} = \theta_t - \zeta_{\text{Adam}} \nabla \mathcal{L}(\theta_t), \tag{7}$$

where ζ_{Adam} is the learning rate, $\nabla\mathcal{L}(\theta_t)$ is the gradient of the DQN loss function. Here, $(\theta, \nabla\mathcal{L}) = (\theta_1, \nabla\mathcal{L}_1)$ and $(\theta, \nabla\mathcal{L}) = (\theta_2, \nabla\mathcal{L}_2)$ for BB and NB agents, respectively. Complete steps followed by hDQN-based beam alignment problem are shown in Algorithm 1. Thus, we note that hDQN is practical in applying only narrow beams over the channel by using BB agent as a meta-controller. Here, we define episode as the consecutive set of hierarchical actions applied on the starting state until it reaches the terminal state with maximum beamforming gain for that location. In order to prevent episodes with infinite set of actions during training, we confine maximum episode length to exhaustive set of beam pairs under the chosen UPA configuration. Hence in the proposed hDQN, the maximum episode lengths are $\mathbf{card}(\mathcal{A}_{\mathcal{B}})$ (say $K_{\mathcal{B}}$) and V for the BB and NB agents, respectively.

We consider the overall hDQN training procedure into Warmup and Training phases, similar to our prior work in [9]. As the reward formulation in (6) involve computing $R_{\max}(s_t)$ measurements over $\mathcal{A}_{\mathcal{N}}$, we consider the warmup phase only for NB agent of hDQN. During Training phase, the hDQN perform exploration and exploitation using ϵ-greedy policies ϵ_1 and ϵ_2 for BB and NB agents, respectively. The Warmup phase results in extra training time at the start but favours quick convergence of hDQN during training phase resulting in faster beam-alignment training for the multi-location environment.

Algorithm 1: Hierarchical DRL using DQN

$M \rightarrow$ Training Episodes; Algorithm hyper-parameters: BB learning rate $\zeta_1 \in (0, 1]$, BB ϵ-greedy rate $\epsilon_1 \in (0, 1]$, BB episode limit $K_{\mathcal{B}}$, NB learning rate $\zeta_2 \in (0, 1]$, NB ϵ-greedy rate $\epsilon_2 \in (0, 1]$, NB episode limit V;
Initialization of replay memory D_1 to capacity C_1, D_2 to capacity C_2, BB network parameters $\theta_1, \overline{\theta_1}$ and NB network parameters $\theta_2, \overline{\theta_2}$;
$\mathcal{S}$: State space of BB, NB agent;
$\mathcal{A}_{\mathcal{B}}, \mathcal{A}_{\mathcal{N}}$: Action space of BB and NB agent, respectively;
for *episode* $\leftarrow 1$ *to* M `// for each episode`
do
    Any random UAV transmits the communication request from the (x,y,z) location;
    BS responds with a sequence of $VK_{\mathcal{B}}$ action beam-pairs over the channel with π_1, π_2 policies;
    Initialization of s_0 by executing a random action $a_0^{\mathcal{B}}$, $a_0^{\mathcal{N}}$ and (x,y,z) location information;
    $k = 0$
    while $True$ **do**
        if *done or* $(k = K_{\mathcal{B}})$ **then**
            `// End Training episode`
            Reset Env and obtain new s_0
        select $a_t^{\mathcal{B}}$ from BB network following ϵ_1
        BS selects the NB action subset $\mathcal{P}$ $(|\mathcal{P}| \leq V)$ corresponds to $a_t^{\mathcal{B}}$;
        $p = 0$;
        for $p \leq V$ **do**
            if *done or* $(p = V)$ **then**
                `// End NB episode`
                Update $R^{\mathcal{N}}_{\max}(s_{t+p})$;
            if *warmup* **then**
                Randomly select $a_t^{\mathcal{N}} \in \mathcal{A}_{\mathcal{N}}$
            else
                select $a_t^{\mathcal{N}}$ from NB network following ϵ_2
            BS applies $a_t^{\mathcal{N}}$ over the channel, receive signal for $(t+1)^{th}$ episode during uplink communication;
            UE observes s_{t+1} and calculate the reward $r(a_t^{\mathcal{N}})$;
            Store the experience $(s_t, a_t^{\mathcal{N}}, r_t, s_{t+1})$ to D_2;
            Train and update NB parameters θ_2;
        Store the experience $(s_t, a_t^{\mathcal{B}}, r(a_{t+p}^{\mathcal{N}}), s_{t+p})$ to D_1;
        Train and update BB parameters θ_1;
    $k = k + 1$ `// Increment episode time`
    hDQN updates the sequence of action beam-pairs for (x,y,z) location;

IV. Simulation Results

As described in Section II-C and Section III, we implement the hDQN-based beam-pair alignment, following $(P1)$ objective (6) and Algorithm 1. Similarly, we implement the state-of-the-art DQN-based approach [9] over UPA configuration and compare our results. We note here that both the RL-based methods once converged, can significantly reduce the communication overhead during P_2 procedure [11] and maximize the beamforming gain in $\mathcal{O}(1)$ time. In this section, we first investigate the training performance of our proposed hDQN-based approach against DQN-based method under diverse channel conditions. Later, we evaluate the time and rate complexities of the proposed hDQN, DQN training procedures over different UPA configurations and UMa-nLoS channel conditions. Here, we select 4 random reflection points within BS coverage and fix them throughout the hDQN and DQN nLoS simulations. For simplicity, we consider UE hovers in $\mathbb{U}$ with fixed radial distance from BS $\xi = 20$ m. The simulation conditions for all the numerical results are listed in Table I.

Table I: Simulation Parameters

Parameters	Value
mmWave freq	30 GHz
Bandwidth W	120 MHz
antenna element spacing d	0.5
Transmit power P_{tx}	0 dBm
Transmit antenna elements N_{tx}	$\{(2 \times 2), (4 \times 4)\}$
Receiving antenna elements N_{rx}	$\{(4 \times 4), (8 \times 8)\}$
Noise Level N_0	-174 dBm
BS location (in m)	$[0, 0, 25]$
UMa-nLoS pathloss coefficients	$\alpha : 4.6 - 0.7 \log_{10}(\mathbb{U}_{\text{zloc}})$, $\beta : -17.5, \gamma : 2.0, \sigma : 6.0$, $\kappa : 20 \log 10(\frac{40\pi}{3})$ [11]

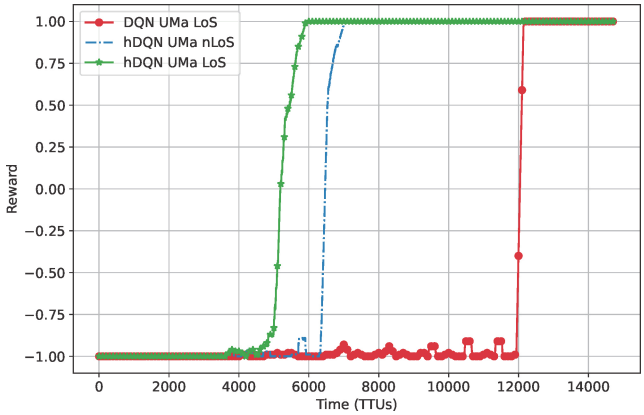

Figure 4: hDQN, DQN training convergence for $(N_{\text{TX}}, BN_{\text{RX}}, N_{\text{RX}}) = (2\times2, 4\times4, 8\times8)$ UPA configuration.

A. hDQN vs DQN Training Performance

As shown in Figure. 4, red plot shows the DQN overall reward performance under UMa-LoS channel while the green and blue plots depict the rewards (following (6)) over hDQN overall training time under UMa-LoS and UMa-nLoS conditions, respectively. We note that all the simulations are carried out with $(N_{\text{TX}}, N_{\text{RX}}) = (2 \times 2, 8 \times 8)$ and $(l_x, l_y) = (2, 2)$ i.e. $BN_{\text{RX}} = 4 \times 4$ UPA is selected under UMa channel

with thermal noise but no shadow fading and channel variation conditions. DQN simulations are performed over $A_{\mathcal{N}}$ action space while the hDQN method apply broad ($\mathcal{A}_{\mathcal{B}}$) and narrow beams ($\mathcal{A}_{\mathcal{N}}$) using BB and NB networks, respectively. We observe that under both UMa-LoS and UMa-nLoS conditions, the hDQN training procedure attain the maximum reward with significantly less training time compared to the DQN method, resulting in faster training convergence.

B. hDQN For Different UPA Configurations

In this subsection, we plot the training times (TTUs) and maximum achievable data rates (5) of hDQN and DQN-based approaches under different UPA antenna configurations with UMa-nLoS conditions. As shown in Figure. 5, blue and red bars show the training times of hDQN and DQN, respectively while the black plot depict maximum learnt data rates obtained in both the methods. We note that same DNN architecture and hyper-parameters values are used for all the hDQN simulations. We observe that hDQN converges faster than DQN and achieves the maximum data rate with average reduction in training overhead of 43% among all UPA simulations. Under the same $(N_{\mathrm{TX}}, N_{\mathrm{RX}})$ configuration, we observe that selection of higher BN_{RX} increases the reliability of providing maximum achievable rate across narrow grid element area in $\mathbb{U}$. This also impacts the training time due to the increase in state space S for DQN, both S and $\mathcal{A}_{\mathcal{B}}$ in hDQN. However, we notice that hDQN converges faster as the selection of broad beam actions depends on both ϵ_1 policy and the convergence of NB network. Now, increasing the BN_{RX}, decreases the cardinality (V) of narrow beam subset for each broad beam-pair in $\mathcal{A}_{\mathcal{B}}$, resulting in faster overall convergence. Thus, the observed results show that broad beam level selection is crucial and has more impact on both training and rate performance under the hDQN approach. This can be useful to trade-off rate and training performances over broad beam level selections for different cellular UAV applications.

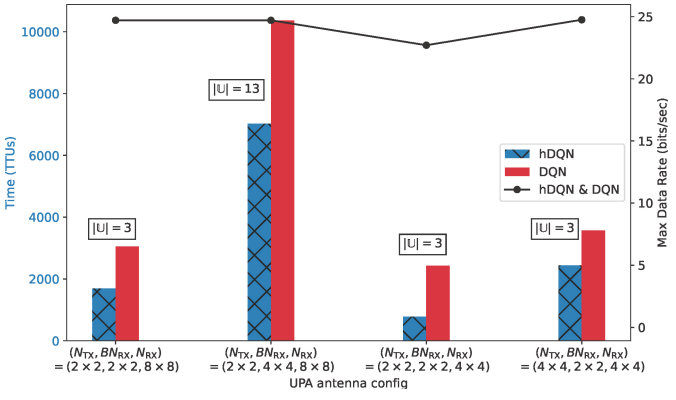

Figure 5: hDQN overall training performance under UMa-nLoS conditions.

V. Conclusion and Future Work

In this paper, we proposed a hDQN-based position-aided beam alignment framework for cellular-connected mmWave UAVs and maximize their beamforming gain within the BS coverage area in an online manner. We also analyzed the hDQN approach over state-of-the-art DQN-based method under different UPA antenna configurations and diverse channel conditions. Our results shown that, the proposed hDQN approach converges faster than the DQN method with an average overall training reduction of 43% for UPA configurations. Having shown some promising results, we will address the hDQN architecture under different UAV radial distances from BS, large number of beam-directional pairs, interference mitigation etc. as future works.

VI. Acknowledgements

The research was supported by 6G Flagship (Grant No. 346208), Finland and the Engineering and Physical Research Council (EPSRC), U.K., under Grant EP/W004348/1.

References

[1] L. Zhang, H. Zhao, S. Hou, Z. Zhao, H. Xu, X. Wu, Q. Wu, and R. Zhang, "A Survey on 5G Millimeter Wave Communications for UAV-Assisted Wireless Networks," *IEEE Access*, vol. 7, pp. 117 460–117 504, July 2019.

[2] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P. E. Wang, "The sky is not the limit: Lte for unmanned aerial vehicles," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 204–210, April 2018.

[3] M. M. Azari, F. Rosas, A. Chiumento, and S. Pollin, "Coexistence of terrestrial and aerial users in cellular networks," in *2017 IEEE Globecom Workshops (GC Wkshps)*. IEEE, Dec 2017, pp. 1–6.

[4] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave v2i beam alignment," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4042–4058, Dec 2017.

[5] J. C. Aviles and A. Kouki, "Position-aided mm-wave beam training under nlos conditions," *IEEE Access*, vol. 4, pp. 8703–8714, Nov 2016.

[6] I. Aykin, B. Akgun, M. Feng, and M. Krunz, "Mamba: A multi-armed bandit framework for beam tracking in millimeter-wave systems," in *Proc. of the IEEE INFOCOM 2020 Conference, Toronto, Canada*, July, 2020.

[7] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online Learning for Position-Aided Millimeter Wave Beam Training," *IEEE Access*, vol. 7, pp. 30 507–30 526, Mar 2019.

[8] P. Susarla, Y. Deng, G. Destino, J. Saloranta, T. Mahmoodi, M. Juntti, and O. Sílven, "Learning-based trajectory optimization for 5g mmwave uplink uavs," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, June, 2020, pp. 1–7.

[9] P. Susarla, B. Gouda, Y. Deng, M. Juntti, O. Sílven, and A. Tölli, "Dqn-based beamforming for uplink mmwave cellular-connected uavs," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[10] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3380–3392, Jan 2016.

[11] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020.

[12] R. S. Elliott, "Antenna theory and design, a john wiley & sons," *INC., Publication*, 2003.

[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, Feb, 2015.

[14] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," *Advances in neural information processing systems*, vol. 29, 2016.

[15] F. Hu, Y. Deng, and A. H. Aghvami, "Cooperative multigroup broadcast 360° video delivery network: A hierarchical federated deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, 2021.