Learning-Based Beam Alignment for Uplink mmWave UAVs

Praneeth Susarla, *Graduate Student Member, IEEE*, Bikshapathi Gouda, *Graduate Student Member, IEEE*, Yansha Deng, *Member, IEEE*, Markku Juntti, *Fellow, IEEE*, Olli Silvén, *Senior Member, IEEE*, and Antti Tölli, *Senior Member, IEEE*

***Abstract*— Unmanned aerial vehicles (UAVs) are the emerging vital components of millimeter wave (mmWave) wireless systems. Accurate beam alignment is essential for efficient beam based mmWave communications of UAVs with base stations (BSs). Conventional beam sweeping approaches often have large overhead due to the high mobility and autonomous operation of UAVs. Learning-based approaches greatly reduce the overhead by leveraging UAV data, like position to identify optimal beam directions. In this paper, we propose a deep Q-Network(DQN)-based framework for uplink UAV-BS beam alignment where the UAV hovers around 5G new radio (NR) BS coverage area, with varying channel conditions. The proposed learning framework uses the location information and maximize the beamforming gain upon every communication request from UAV inside the multi-location environment. We compare the proposed framework against multi-armed bandit (MAB)-based and exhaustive approaches, respectively and then analyse its training performance over different coverage area requirements, antenna configurations and channel conditions. Our results show that the proposed framework converge faster than the MAB-based approach and comparable to traditional exhaustive approach in an online manner under real-time conditions. Moreover, this approach can be further enhanced to predict the optimal beams for unvisited UAV locations inside the coverage using correlation from neighbouring grid locations.**

***Index Terms*— 5G, mmWave, beam alignment, deep Q-network.**

I. Introduction

UNMANNED aerial vehicles (UAVs) are envisioned as the vital ingredients for future wireless systems using millimeter wave (mmWave) [2], [3], [4], [5]. Especially, the deployment of cellular-enabled UAV-user equipment (UE)s (*hereafter addressed as UAVs*) adds unique features pertaining to high mobility in three dimensional space and autonomous operations, find appealing solutions to a myriad of civil applications, such as traffic surveillance, mineral exploration, internet drone delivery systems, etc. [6], [7]. It is noted that 3rd generation partnership project (3GPP) has recently released a technical report to understand existing challenges, requirements and opportunities of Unmanned aerial vehicles (UAVs) integration to fifth generation (5G) and beyond 5G communication networks [8]. The mmWave frequencies (30 GHz to 300 GHz) together with multiple input multiple output (MIMO) beamforming capabilities can provide high capacities and line-of-sight (LoS) dominant connectivity [9] to the aerial-ground communications between base station (BS) and unmanned aerial vehicle (UAV). Besides, the real-time ultra-high speed transmissions with UAVs still requires achieving challenges such as reliability and low-latency communications during aerial-ground communications. Solving these challenges for mmWave beam alignment is also essential for efficient control of UAVs in beyond 5G communications. It is noted that more flexible three-dimensional (3D) beamforming will be deployed in the forthcoming 5G systems, to enhance the beamforming gain based on the angle resolutions in both azimuth and elevation dimensions of UAVs in the sky [5].

Fast mmWave beam alignment can enhance the data throughput for both UAV-UAV and BS-UAV communications under 5G and beyond wireless systems. Especially, the availability of UAV position information at lower frequencies (following the works [10], [11]) may also provide scope for reliable communication in addition to increasing throughput. Position information for fast beam alignment has been recently studied under vehicular context in mmWave systems [10], [11], [12], [13], [14]. On the other hand, high mobility and autonomous operation of UAVs will require frequent beam realignment as well. Therefore, a faster and reliable beam alignment using UAV position information is crucial in enabling high data rate for mmWave UAVs.

An effective beam alignment or tracking scheme is usually required to ensure the consistency of beam alignment in a high mobility environment. Existing works [15], [16], [17] proposed beam tracking schemes using Kalman filters with high processing complexity. Moreover, such schemes are vulnerable to abrupt changes in environment to the high speed UAVs and

Manuscript received 1 June 2021; revised 11 October 2021, 18 January 2022, 19 April 2022, and 6 July 2022; accepted 31 August 2022. Date of publication 21 September 2022; date of current version 10 March 2023. This work was supported in part by the Academy of Finland (6Genesis Flagship) under Grant 346208; and in part by the Engineering and Physical Research Council (EPSRC), U.K., under Grant EP/W004348/1. An earlier version of this paper was presented in part at the IEEE Global Communication Conference (GLOBECOM), Madrid, Spain, August 2021 [DOI: 10.1109/GLOBECOM46510.2021.9685080]. The associate editor coordinating the review of this article and approving it for publication was X. Cheng. *(Corresponding author: Praneeth Susarla.)*

Praneeth Susarla, Bikshapathi Gouda, Markku Juntti, Olli Silvén, and Antti Tölli are with the Faculty of Information Technology and Electrical Engineering, University of Oulu, 90570 Oulu, Finland (e-mail: praneeth.susarla@oulu.fi; bikshapathi.gouda@oulu.fi; markku.juntti@oulu.fi; olli.silven@oulu.fi; antti.tolli@oulu.fi).

Yansha Deng is with the Department of Informatics, King's College London, WC2R 2LS London, U.K. (e-mail: yansha.deng@kcl.ac.uk).

Data is available on-line at https://github.com/psusarla95/Precoding_RL

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3206714.

Digital Object Identifier 10.1109/TWC.2022.3206714

tracking error accumulating over time [16]. On the other hand, conventional beam sweeping solutions [18], [19], [20] also often have large overhead as the UAVs move at high speeds and perform autonomous operations. An alternate approach is to undergo beam training [16] and perform fast beam alignment after every significant change in UAV position along the BS coverage area. Existing works in vehicular environment proposed different training-based beam alignment approaches for terrestrial systems based on stochastic methods such as genetic and evolutionary algorithms [21], [22] and the use of contextual information [10], [11], [12], [13], [14].

Contextual information generally involves data from the sensors such as position information, antenna configurations, channel state information and receiving signal power using low frequency carrier (e.g. during initial communication in 3GPP beam access protocol [23], [24]) whenever needed, as this information is used abundantly to reduce the beam training overhead. The authors in [10] assumes global positioning system (GPS) position in vehicular scenario, considered the radio fingerprints with position information as context and perform beam training at the beam level with only one feedback. They even proposed an online learning-based mmWave beam alignment solutions using position information [11], angle of departure (AoD) distribution with compressive sensing (CS)-based channel measurements [14] as the context information, respectively. The proposed methods suits well to the vehicular communication context with rapid channel variation in the environment. The method also requires maintaining a database of channel strength information with pre-processing of beam pairs before beam selection procedure, for each user location in the vehicular environment. On the other hand, UAV trajectories in general are not bound to any specific physical structures like roads, streets, pavements etc. and involves large coverage areas and high speed mobility of UAVs resulting in frequent change of UAV position information. As a result, pre-processing of beam pairs and maintaining a database could be complex, non-generic and also cause significant overhead.

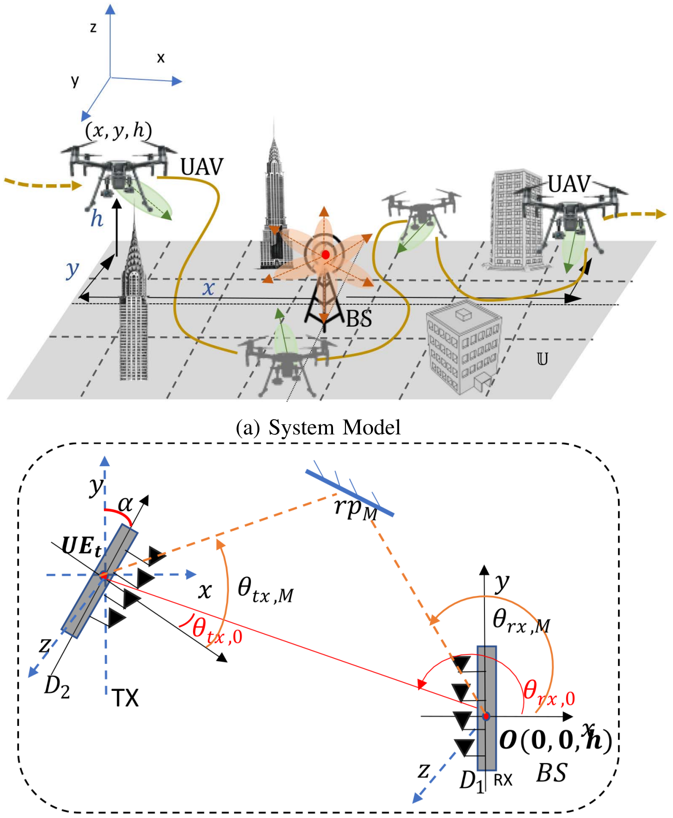
Stochastic methods such as genetic and evolutionary algorithms that require hundreds of evaluations in order to discover an optimum with sufficient precision [21]. Even though these techniques are computationally efficient, communication overhead can be very significant as each iteration of the objective function implies applying a new beam pair indices in our problem. However, the recent advancements of machine learning (ML) and neural networks (NN) techniques for stochastic methods offers a possibility to learn such objective functions and reduce overhead for communications [25], [26]. The authors in [27] and [28] investigated application of supervised ML and statistical ML techniques to reduce beam training overhead by assuming a separate data collection phase with an offline-learning environment, respectively. On the other hand, learning approaches such as multi-armed bandit (MAB) and other reinforcement learning (RL) methods such as temporal difference learning, Q-learning etc. can generate the required data through their interactions with the environment and learn the beam alignment problem in an online manner [11], [14]. Moreover, deep RL algorithms such as deep Q-network (DQN), actor-critic (A2C), policy gradient methods etc. also provide the generic ML framework with interactive learning process [29], [30].

Such generic frameworks were recently investigated for mmWave beam alignment between BS and multiple ground users [31], [32]. In [31], a generic MAB framework for mmWave beam alignment and tracking was proposed for omni-directional ground users moving randomly around the BS. The algorithm assume velocity of the users to be stochastic with the selected arm (beam subset) to be following different probability distribution models based on varying channel conditions in the environment. However, such proposed framework can result in increasing beam training overhead as the action space grows exponentially when 3D beamforming is considered between BS and directional UAVs. Moreover, the framework requires switching across multiple probability distribution models as aerial-ground channel vary differently at distinct heights [33]. In [32], a DQN-based interactive learning process was studied to design efficient mmWave beam training algorithms for both multi-user and user-centric communications. Therein, the DQN interacts with communications module to select a beam subset for efficient beam training with multiple ground users in the environment under fast varying channel conditions. Their proposed framework was interactive and generic, but their proposed beam image construction method for state space is suitable mainly to the omni-directional ground user communications. The beam image construction can be very expensive (involves more feedback information from UE with uniform planar array (UPA) antenna arrays) and less sparse when UAVs with directional beams move with high mobility speeds at distinctive heights. However, such generic and interactive algorithms can be helpful to quickly learn the beam alignment after every initial access during the high mobility and autonomous operation of UAVs. Thus, the key idea in our work is to design such ML framework using low-cost position information of UAVs and serve any random initial access requests during their mobility under BS coverage area, with efficient beam directional pairs learnt from their past beam measurements.

In our work, we formulate an online context-information-aided beam alignment problem as a partially observable Markov decision process (POMDP). We then approximate the beam-pair optimization problem for a multi-location environment using NN techniques to address the limitations of previous works [10], [11], [12], [13], [14]. The NN approximation with POMDP formulation, uses prior knowledge of the radio propagation from current and neighbouring UAV context-information in the environment to significantly reduce the beam-alignment overhead under different coverage area requirements. For the algorithm to identify significant change in UAV location, we consider a grid environment with UAV position information of the grid element as the user-context information. Thus, we combine both stochastic and user-context optimization techniques [10], [11], [12], [13], [14], [21], [22] with DQN and MAB learning frameworks, in an attempt to progress from the previous works discussed so far. Our algorithms takes into account the low-cost UAV location information and finds an optimal beam pair to maximize the beamforming gain between BS and UAV. The random mobility and autonomous operation of UAVs resulting in altitude and position information changes are captured using different grid elements around the BS coverage area. Our contributions are summarized as follows:

- We model the context-information-aided beam-pair alignment problem in uplink mmWave MIMO communication system as a POMDP. In this system, the BS serves multiple UAVs in a time domain multiple access (TDMA) manner under its coverage area, using 5G new radio (NR) based communication protocol.
- We solve the context-information-aided beam-pair alignment problem with deep reinforcement learning (DRL) e.g. using a online DQN-based algorithm. During uplink communication, the proposed method optimizes the BS-UAV beam-pair alignment generically across any UAV grid position inside the BS coverage area. Under the DQN learning-based beam alignment procedure, we study the warming and training phases to identify the optimal data rate measurements for each UAV location and then learn the corresponding optimal beam-pairs, respectively.
- We compare our DQN-based proposed approach with the adapted state-of-the-art MAB-based approach [11] and traditional exhaustive-based beam alignment under ideal channel conditions. We analyse these approaches over different coverage areas and antenna configurations under ideal channel conditions. Our results shown that DQN-based approach achieves the optimal beam alignment as the traditional exhaustive method but with reduced number of training iterations.
- We study the online training performance of our proposed DQN framework by incorporating thermal noise, shadow fading and slow channel variation at each UAV grid location under both LoS and non-line of sight (nLoS) conditions. Our results show that the online DQN-based approach is practical to the 3GPP standards and also effective to solve the beam alignment problem across multiple grid positions, for UAVs moving within the BS coverage area. Furthermore, this generic learning-based approach can be enhanced to maximize the beamforming gain for untrained UAV locations inside the coverage area as long as the predicted beam-pairs are not very far from the optimal beams for converged neighbourhood grid locations.

The rest of the paper is organized as follows. Section II presents the problem formulation and communication modelling, considered in this problem. The section also explains in details about the learning scenario considered for the problem. Section III formulates the exhaustive and learning based methods used in this work. The section also discuss in detail about the proposed DQN based RL approach for beam alignment. Section IV presents the comparison of DQN and MAB learning approaches against traditional exhaustive method and then simulates the proposed DQN-based approach through different system parameters under ideal channel conditions. The section then discuss about the online learning of DQN and its analysis under varying channel conditions. Section V presents the conclusion and future work.



(a) System Model

(b) Communication model

Fig. 1. Problem view.

II. System and Communication Model

In this section, we describe the system model, communication model for the learning framework following 3GPP 5G NR protocol [24] and also briefly describe the parameters used in the learning framework. The objective of this problem is to maximize the beamforming gain between BS and the UAV to provide efficient communication under the defined environment and channel conditions.

A. System Model

We consider a cellular mmWave MIMO uplink communication with BS serving multiple UAVs in a TDMA manner under its coverage area. The BS is fixed at $\mathcal{O}(0, 0)$ and communicates with the moving UAV (hereafter used as UE) using a multipath mmWave beamforming for urban macro-cellular (UMa) environment as shown in Figure 1a. The multi-antenna UE hovers randomly and communicates with the multi-antenna BS in the urban environment following 5G NR standard protocol [34]. We consider an analog beamforming equipped with one radio frequency (RF) chain and uniform linear array (ULA) structures of N_t and N_r antennas for both BS and UE, respectively. We note that the model considers ULA structures for simplicity in this work and can be extended to UPA as well to support 3D beamforming with better interference suppression capabilities. The UE transmits (TX) while the BS receives (RX) a radio signal in multiple beam directions following $\mathcal{B}_{\text{TX}}$ and $\mathcal{B}_{\text{RX}}$ codebook, respectively with angles defined as

$$b_i = (i-1)\frac{\pi}{N}, i \in \{1, 2, \ldots, N\}, \tag{1}$$

where b_i represents a RF radio beam direction with a fixed narrow beam width ($\frac{\pi}{N}$), N represents N_t, N_r antennas for $\mathcal{B}_{\text{TX}}$ and $\mathcal{B}_{\text{RX}}$ codebook, respectively. We note here that this choice is not important and other codebooks such as the discrete Fourier transform (DFT) codebook could be used. The codebook values are defined using the beamforming vectors $\mathbf{w}_{\text{TX}}$ and $\mathbf{w}_{\text{RX}}$ for UE and BS, respectively, given by

$$\mathbf{w}(b_i)\Big]_{n=0}^{N-1} = \frac{1}{\sqrt{N}} \exp(j\frac{2\pi nd}{\lambda}\sin(b_i)), b_i \in \mathcal{B}, \tag{2}$$

where $N = N_t$, $\mathcal{B} = \mathcal{B}_{\text{TX}}$ and $N = N_r$, $\mathcal{B} = \mathcal{B}_{\text{RX}}$ for $\mathbf{w} = \mathbf{w}_{\text{TX}}$ and $\mathbf{w} = \mathbf{w}_{\text{RX}}$, respectively. Here, d is the antenna spacing assumed to be $\frac{\lambda}{2}$ in this work. λ is the wavelength and b_i is the i^{th} codebook direction (1).

The UE moves randomly along the 3D coverage area which is divided into multiple (x, y, z) grids of equal size, the set enclosing them is denoted as $\mathbb{U}$. Here we assume UE hovers on every hop with random mobility and enters the grid position defined in $\mathbb{U}$. The communication begins with a TX request from UE, while the RX radio unit at BS starts with a random beam-pair at time $t = 0$ and learns to choose the beam direction $(b_p, b_q), b_p \in \mathcal{B}_{\text{TX}}, b_q \in \mathcal{B}_{\text{RX}}$ over time for each TX grid position. We assume TX and RX beam directions to be the same for all UE movements within each grid position. The TX and RX change their beam directions with grid positions when there is a substantial change in TX locations, due to variance in their radio measurements. Hence, the UE is expected to follow the beamforming protocol at every grid position, along the coverage area set $\mathbb{U}$. When UE moves to a different grid, it performs beam alignment initial access procedure at the grid position by following 3GPP beamforming protocol. In this work, we note here that this single BS and single UE assumption is not limited and can be extended to multi-cell, multiple UE-BS scenarios with multiple RF chains and complex antenna array structures as well.

The 3GPP beamforming protocol of physical layer in general involves mainly two procedures, initial communication (used as P_1 procedure) and beam management consisting of beam selection (used as P_2 procedure) and an optional beam refinement (used as P_3 procedure) [23]. We investigate the 3GPP based beam-pair alignment learning through P_1 and P_2 procedures between BS and UE as shown in Figure 2. P_1 procedure mainly involves requesting a connection between BS and UE using synchronization signal blocks and random-access procedure at lower frequencies [23], [24]. As a part of this procedure, the UE is assumed to send a communication request with respect to its position each time, while the learning framework at BS responds with a 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 higher frequencies for the data transmission. We consider the P_2 procedure to follow 5G NR communication with uplink transmissions at mmWave frequencies [24], [34]. Similar to the works in [11], 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.

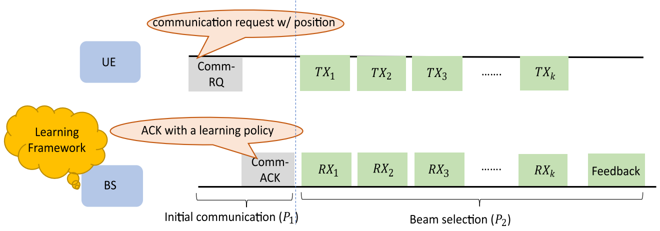


Fig. 2. 3GPP beamforming protocol design for beam-pair learning following [23]. P_1 procedure is assumed to communicate at lower frequencies (similar to the works in [11], [14]). P_2 procedure is assumed to follow 5G NR communication with uplink transmissions at mmWave frequencies [24], [34].

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 [34]. 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 due to the learning-based beam sweeping procedure in TTU units.

B. Communication Model

For the communication model, we consider a multi-path link (LoS or nLoS) radio channel between UE at time t and BS location $\mathcal{O} \in \mathbb{R}^3$ as shown in Figure 1b. $\text{UE}_l(t)$ represents the UE position on grid index $l \in \{0, 1, \ldots |\mathbb{U}|\}$ in the BS coverage area $\mathbb{U}$ at any time instant t, given by

$$\text{UE}_l(t) = (x_t, y_t, z_t), \tag{3}$$

where $\{x_t, y_t, z_t\} \in \mathbb{U}$. We assume UE_l (with respect to BS locations) is known during each P_1 procedure of 3GPP beam access protocol. α in Figure. 1b, represents the angle of rotation of UE with respect to y-axis. For simplicity, we assume $\alpha = 0$ radians, which means there is no rotation of drone. Let $\theta_{\text{tx,m}}$, $\theta_{\text{rx,m}}$ be the AoD and angle of arrival (AoA) of m^{th} communication link between BS and UE, respectively. The UE transmits radio signals in a codebook direction from $\mathcal{B}_{\text{TX}}$ while the BS receives the signal through one of its multiple beam directions following $\mathcal{B}_{\text{RX}}$ (1). Baseband equivalent of the received signal is given by

$$y[k] = \underbrace{\sum_{m=0}^{M} \sqrt{P_{tx}}\beta_m \, \mathbf{w}_{\text{RX}}^H \mathbf{a}_R(\theta_{\text{rx,m}})\mathbf{a}_T^H(\theta_{\text{tx,m}})\mathbf{w}_{\text{TX}} x[k]}_{r[k]} + \nu[k], \tag{4}$$

where P_{tx} is transmission power, M is the number of multi-paths or reflection points in the UMa environment [35], β_m, $\mathbf{a}_R(\theta_{\text{rx,m}}) \in \mathbb{C}^{N_r}$, $\mathbf{a}_T(\theta_{\text{tx,m}}) \in \mathbb{C}^{N_t}$ are the antenna channel gain and array response vectors for $\theta_{\text{rx,m}}$ and

$\theta_{\text{tx,m}}$ along the m^{th} multi-path communication link, respectively. Here, $\mathbf{a}(\theta)]_{l=0}^{N-1} = \frac{1}{\sqrt{N}} \exp(j \frac{2\pi l d}{\lambda} \sin(\theta))$, where $\theta = \theta_{\text{rx,m}}, N = N_r$ and $\theta = \theta_{\text{tx,m}}, N = N_t$ for $\mathbf{a}_R(\theta_{\text{rx,m}})$ and $\mathbf{a}_T(\theta_{\text{tx,m}})$, respectively. $\mathbf{w}_{\text{RX}} \in \mathbb{C}^{N_r}$, $\mathbf{w}_{\text{TX}} \in \mathbb{C}^{N_t}$ are the transmit and receive unit-norm beamforming vectors following (2), $\nu[k] \sim \mathcal{CN}(0, WN_0)$ 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 (from Section II-A) with $\frac{1}{K}\sum_{k=0}^{\text{K}} \|x[k]\|^2 = 1$. Here, $k = 0, 1, \ldots K$ is the number of samples spanned over TTU time. The mmWave channels in general are expected to have limited scattering and hence following a geometric channel model with m reflectors can be effective [36]. In this work, we assume each reflector m contributes to a single propagation path and follow 3GPP UMa channel conditions [33], [35]. Thus, the channel measurements H_m of the m^{th} multi-path link is given by $\mathbf{H}_m(\theta_{\text{tx},m}, \theta_{\text{rx},m}) \triangleq \beta_m \mathbf{a}_R(\theta_{\text{rx,m}}) \mathbf{a}_T^H(\theta_{\text{tx,m}})$. We define

$$r[k] = \sum_{m=0}^{M} \sqrt{P_{tx}} \mathbf{w}_{\text{RX}}^H \mathbf{H}_m(\theta_{\text{tx},m}, \theta_{\text{rx},m}) \mathbf{w}_{\text{TX}} x[k] \tag{5}$$

Similar to the formulation mentioned in [36], the signal-to-noise ratio (SNR) is given as $\text{SNR} = \frac{\frac{1}{K}\sum_{k=0}^{K} \|r[k]\|^2}{N_0 W}$ and overall rate measurement R in bits per channel use is given by

$$R = \log_2(1 + \text{SNR}). \tag{6}$$

The learning formulation (described in detail under Section III) requires characterizing the known and unknown parameters of the environment into exhaustive set of states (also known as state space) and actions (action space), respectively. Based on the beam management protocol considered in Figure 2, we define the state, action spaces $(\mathcal{S}, \mathcal{A})$ for MAB and RL learning-based beam-pair alignment, respectively as follows:

$$\begin{aligned} (\mathcal{E}_1): \ & \mathcal{S} = \{\text{UE}_l, \text{UE}_l \in \mathbb{U}\} \\ & \mathcal{A} = \{(b_p, b_q), b_p \in \mathcal{B}_{\text{TX}}, b_q \in \mathcal{B}_{\text{RX}}\}, \end{aligned} \tag{7}$$

$$\begin{aligned} (\mathcal{E}_2): \ & \mathcal{S} = \{(\text{UE}_l, b_r, b_s), \text{UE}_l \in \mathbb{U}, b_r \in \mathcal{B}_{\text{TX}}, b_s \in \mathcal{B}_{\text{RX}}\} \\ & \mathcal{A} = \{(b_p, b_q), b_p \in \mathcal{B}_{\text{TX}}, b_q \in \mathcal{B}_{\text{RX}}\}, \end{aligned} \tag{8}$$

where UE_l is the location of UE within coverage area $\mathbb{U}$ (3) while $\mathcal{B}_{\text{TX}}$, $\mathcal{B}_{\text{RX}}$ are the beam codebook sets at UE and BS side respectively (1). Here, $\mathcal{E}_1$, $\mathcal{E}_2$ are the state-action space learning environments for MAB and RL methods, respectively. The 3GPP beam alignment protocol aims to find the best (TX, RX) beam pair over time between current grid location of UE and BS, for both $\mathcal{E}_1$ and $\mathcal{E}_2$ learning environments (similar to problem formulation considered in [11]). $\mathcal{E}_2$ considers the recently applied beam-pair as part of the state information for the transition towards optimal beam-pair following POMDP. Thus, the beam-pairs influence both state and action spaces in this environment. The altitude and position changes of UE due to their mobility can be proportional to the number of grid elements considered as state-space in both learning environments. For example, under high UAV mobility, number of grid elements increases in a high channel variation scenario while it is sufficient to have less state space with large grid volume size under slow channel variation environments. Thus, we emphasize that the fixed size of a grid element depends on both state and action spaces considered under the BS coverage area in the learning environments[1]. As shown in Figure 2, the UE provides location information during P_1 procedure, while the BS responds with the sequence beam-pairs at lower frequencies that can be configured to sweep the RF beams at mmWave frequencies during P_2 procedure. Such sequence of beam-pairs are optimized over time by observing the history of multiple initial access procedures corresponding to UE locations in the grid environment. This can eventually reduce the communication overhead during initial access using the learning methods described in Section III. We assume that the UE position information during P_1 procedure can be accurately obtained either from the sensors mounted on UAV such as GPS, camera, lidar, radar etc. or using 5G localization techniques[1]. Finally, the BS determines the best-beam pair using their data-rate measurements after sweeping and communicates it to the UE for data transmission. The optimal beam-pair for UE-BS is selected based on their data rates under the scenario mentioned in Section II-A.

III. Learning Methods Formulation

In this work, we tackle the beam alignment problem at every grid location as a learning problem. Moreover, the performance of two learning methods, including MAB and RL approach are also comparable to that of traditional exhaustive search method. MAB and RL learning-based methods once converged, can significantly reduce the communication overhead during initial access procedure and maximize beamforming in $\mathcal{O}(1)$ time. On the other hand, the traditional approach always results in exhaustive search over entire action space $\mathcal{A}$ each time. The focus of this work is to design an online learning framework that is generic across both location and time, suitable to the considered environment.

A. The Exhaustive Method

The method mainly involves exhaustive search among the set of actions $\mathcal{A}$, to find the best beam pair with maximum possible beamforming between UE and BS. Since we consider a multi-location environment, exhaustive beam scanning is required for every change in grid element unit of UE inside $\mathbb{U}$. This frequent scanning results in significant communication overhead, especially with higher antenna elements. However, this method can determine the best possible beam alignment between BS and UE. If $s_t \in \mathcal{S}$ is the UE state information available at time instant t, then this method can be formulated as

$$\begin{aligned} (P1): \ & \max_{(a_t|s_t)} R(s_t, a_t), \\ & \text{s.t.} \quad a_t \in \mathcal{A} \end{aligned} \tag{9}$$

where $R(s_t, a_t)$ is the measured data rate on applying a_t to state s_t between BS and UE.

[1] Impact of contextual-errors and grid element size on beam alignment will be studied as a separate work in future.

Algorithm 1 MAB Approach Using Greedy Upper Confidence Bound (gUCB)

```
// Initialization of MAB parameters for
   beam management procedure
hyper-parameters: $M$ episodes per UE location, UCB
 multiplying factor $\alpha \in (0,1)$;
episode length $L = |\mathcal{A}|$
$Q[i] \leftarrow 0, \forall i \in \mathcal{A}$
$\eta_i^\pi \leftarrow 0, \forall i \in \mathcal{A}$ // number of selections
   of action $i$
for $n \leftarrow 1$ to $M$ // for each episode
 do
    // Apply a sequence of beam-pair
       actions over channel
    for $i \leftarrow 0$ to $L-1$ do
        // Compute UCB values
        $\tau(\eta_i^\pi) \triangleq \lceil (1+\alpha)^{\eta_i^\pi} \rceil$
        $a(n,\eta_i^\pi) \triangleq \sqrt{\frac{(1+\alpha)(1+\ln \frac{n}{\tau(\eta_i^\pi)})}{2\ \tau(\eta_i^\pi)}}$
        $\mathrm{UCB}_i \leftarrow \frac{Q[i]}{\eta_i^\pi} + a(n,\eta_i^\pi)$
        $a_n = \operatorname*{argmax}_{l \in \mathcal{A}} \mathrm{UCB}_l$ // action selected
           using greedy policy $\pi$
        $\eta_{a_n}^\pi \leftarrow \eta_{a_n}^\pi + 1$
        BS selects $a_n$ beam-pair for data transmission,
         receive signals for sequence of action beam-pairs
         from UE during uplink communication
        BS calculate the reward $r_{a_n}$ using received signal
         measurements, computing data rate and following
         (9)
        $Q[a_n] \leftarrow Q[a_n] + r_{a_n}$
```

B. k-Armed Bandit Method

Multi-armed bandit (or MAB) is a learning problem, where the best action is selected among k different actions available at the same time instant. Each action selection is associated with a reward and the best action is learnt by repeatedly selecting those actions resulting in maximum possible reward accumulation over time. Here, there are no state transitions involved, unlike in a RL framework or a markov decision process (MDP) process [30]. However, learning the best among available actions sometimes depends on a particular information or a context. Learning in such scenarios are termed as contextual multi-armed bandit (CAB) problem. Moreover, there are different algorithms such as ϵ-greedy, greedy upper confidence bound (gUCB) [11], [37], bayesian-thompson sampling (BTS) [38], fast machine learning (FML) [39] etc. used to tackle the bandit problem.

We focus on formulating the armed bandit based beam alignment problem as the CAB using gUCB algorithm (similar to the works in [11] and [37]). At any time instant t, we consider the learning environment $\mathcal{E}_1$ (7) with UE state information $s_t \in \mathcal{S}$ as the context and the BS action $a_t \in \mathcal{A}$ for the beam-alignment problem. The best action for s_t is learnt over time using the history of its previous as well as current data rate measurements among available action set $\mathcal{A}$ as the reward. The objective of this problem can be formulated as

$$(P2): \max_{\pi(a|s)} \mathbb{E}[\eta_a^\pi(t) r_a(t)],$$
$$\text{s.t.} r_a(t) = \begin{cases} 1 & \text{if } \operatorname*{argmax}_{\pi(a|s)} R(a) \\ 0 & \text{otherwise} \end{cases},$$
$$\eta_a^\pi(t) \in \mathbb{N}, \tag{10}$$

where $\eta_a^\pi(t)$ denotes the number of a $(\in \mathcal{A})$ selections until the time instant t by following a policy π (for example, greedy policy in gUCB), $r_a(t)$ and $R(a)$ are the reward and observed data rate measurement of action a at time instant t, respectively, a is the selected action beam-pair on following policy π for UE grid location s_t. We note that the initial UCB bounds for $a (\in \mathcal{A})$ are equal and set to a minimum as we apply the sequence of beam-pair actions each time over the channel following 3GPP beam alignment procedure and select a gUCB action based on the updated UCB bounds. Complete steps using the gUCB algorithm for CAB-based beam alignment objective are shown in Algorithm 1.

We also perform the convergence bound analysis of gUCB algorithm based on the reward function following (10) and the pseudo regret, one of the widely used metric for MAB problems as defined in [40]. Translating this metric to the beam alignment problem, the regret incurred at time instant is non-zero when the algorithm selects a sub-optimal beam pair with index $i, i \in [1, |\mathcal{A}|]$. Let r_i denote the mean reward on choosing the beam pair index i after n trials $(\mathbb{E}[X_{i,n}] = r_i, \forall n \geq 1)$. Here, $X_{i,n}$ is the reward experienced by beam pair with index i at time instant n $(\bar{X}_{i,n} = \frac{1}{n}\sum_{t=1}^{n} X_{i,t}, 1 \leq i \leq |\mathcal{A}|, n \geq 1)$. We define the mean reward for the optimal beam pair and its index as follows:

$$r^* = \max_{i=1,\ldots|\mathcal{A}|} r_i,$$
$$i^* \in \operatorname*{argmax}_{i=1,\ldots|\mathcal{A}|} r_i. \tag{11}$$

Let $T_i(n)$ denote the number of times the agent selects the pair with index i on the first n trails $(\sum_{i=1}^{|\mathcal{A}|} T_i(n) = n)$. We define, $T^*(n) = T_i(n), i = i^*$ and $\bar{X}_n^* = \bar{X}_{i,n}, i = i^*$. Let $\Delta_i = r^* - r_i$ be the optimality gap of sub-optimal beam pair with index i. Using these, we can now define the pseudo regret bound following [40, eq. 2.1] as

$$\overline{R_n} = \Bigg(\sum_{i=1}^{|\mathcal{A}|} \mathbb{E}\big[T_i(n)\big]\Bigg) r^* - \mathbb{E}\left[\sum_{i=1}^{|\mathcal{A}|} T_i(n) r_i\right]$$
$$= \sum_{i=1}^{|\mathcal{A}|} \Delta_i \mathbb{E}\big[T_i(n)\big]. \tag{12}$$

The pseudo-regret only increases with more selection of sub-optimal pairs over time, providing a good framework to compete against the optimal beam pair. For Algorithm 1, we consider the MAB approach using episodes and apply the rewards defined under (10) following independent and identical distribution (IID). Under each episode, a beam pair (say index i) is selected and then played until $\tau(\tilde{p}_i + 1) - \tau(\tilde{p}_i)$

times, where τ is an exponential function given as $\tau(p) = \lceil(1+\alpha)^p\rceil$. Here, $\tilde{p_i}$ is the number of episodes the i^{th} index beam pair got selected until time instant n. On assuming $p \geq 1$ and $n \geq \frac{1}{2\Delta_i^2}$ we can rewrite the exponential function as

$$\tau(p) \leq (1+\alpha)^p + 1, \tag{13}$$

$$\tau(p) \leq \tau(p-1)(1+\alpha) + 1, \tag{14}$$

$$\implies \tau(\tilde{p_i}) \leq \frac{(1+4\alpha)\ln(2en\Delta_i^2)}{2\Delta_i^2}. \tag{15}$$

Besides, the confidence bound margin of such i^{th} index is given by

$$a_{n,p} = \sqrt{\frac{(1+\alpha)\ln(\frac{en}{\tau(p)})}{2\tau(p)}}. \tag{16}$$

Using these equations and Chernoff-Hoeffding inequality, regret bounds can be derived.

Theorem 1: The pseudo regret at time n of the greedy UCB algorithm with $0 < \alpha < 1$ is upper bounded by

$$\begin{aligned}\overline{R_n} &\leq \sum_{i=1}^{|\mathcal{A}|} \Delta_i \mathbb{E}\big[T_i(n)\big] \\ &= \sum_{i:r_i<r^*} \left(\frac{(1+\alpha)(1+4\alpha)\ln(2e\Delta_i^2 n)}{2\Delta_i} + \frac{c_\alpha}{\Delta_i}\right),\end{aligned} \tag{17}$$

where

$$c_\alpha = 1 + \frac{(1+\alpha)e}{\alpha^2} + \left(\frac{1+\alpha}{\alpha}\right)^{1+\alpha}\left[1 + \frac{11(1+\alpha)}{5\alpha^2\ln(1+\alpha)}\right]. \tag{18}$$

Theorem 1 shows the regret bound for Algorithm 1. The first term increases with increase in n and dominates the bound as Δ_i decreases. Thus, the regret bound for gUCB is $\mathcal{O}(\log n)$ in finding the optimal beam-pair for beam alignment on every location. A proof of this analysis is included in Appendix. A.

C. Reinforcement Learning

RL is an interactive learning problem consisting of set of states $\mathcal{S}$, actions $\mathcal{A}$ and rewards, following a MDP or POMDP process [30]. A state transition is involved on applying each action until a terminal state is reached. The objective of the problem is to learn an optimal policy of state transitions with actions over time and reach the terminal state through reward accumulation [30].

$$\begin{aligned}(P3): \quad &\max_{\{\pi(a_t|o_t)\}} \sum_{i=t}^{\infty} \gamma^{i-t}\mathbb{E}_\pi[r_{a_i}(i)], \\ &s.t. \quad r_{a_t}(t) = \begin{cases} 1 & \text{if } R(a_t) \geq R_{max}(s_t) \\ -1 & \text{otherwise} \end{cases}, \\ &\gamma \in (0,1].\end{aligned} \tag{19}$$

In this work, the RL based beam alignment problem is modelled as a POMDP problem. At any time instant t, we define the parameters $s_t = \{(s', a')\ s' \in \mathcal{S}, a' \in \mathcal{A}\}$, $a_t \in \mathcal{A}$ and $r_t \in \mathcal{R}$ where s_t, a_t, r_t are the state, action and reward at time instant t. Here, $\mathcal{S}$ and $\mathcal{A}$ correspond to state and action spaces for scenario $\mathcal{E}_2$ (8). a' corresponds to the set of previous actions applied for state transitions until the time instant t. As shown in (8), b_r, b_s are the beam codebook directions corresponding to UE$_l$ previous time instant, following $\mathcal{B}_{\text{TX}}$ and $\mathcal{B}_{\text{RX}}$, respectively. This information is helpful to instantiate a state-transition model at UE$_l$, required for RL POMDP formulation [30]. Data rate measurements computed on applying each action are considered as the rewards for the

Algorithm 2 RL Approach Using DQN

```
$M \rightarrow$ Training Episodes;
Algorithm hyper-parameters: learning rate $\xi \in (0,1]$,
  discount rate $\gamma \in [0,1)$, $\epsilon$-greedy rate $\epsilon \in (0,1]$, update
  steps $K$;
Initialization of replay memory $M$ to capacity $C$, the
  primary Q-network with parameters $\theta_1$, the target
  Q-network with parameters $\theta_2$
$\mathcal{S}, \mathcal{A}$: State and Action space of DQN agent
for episode $\leftarrow 1$ to $M$ // for each episode
do
    Any random UAV transmits the communication
      request from the (x,y,z) location.
    $N \rightarrow$Episode limit
    BS responds with sequence of $N$ action beam-pairs
      over the channel with policy $\pi$
    Initialization of $s_1$ by executing a random action $a_0$
      and (x,y,z) location information
    n=0,
    while $True$ do
        // Episode with $\epsilon$-greedy policy $\pi$
        if $p_\epsilon < \epsilon$ then
            select a random action $a_t \in A$
        else
            select $a_t = argmax_{a \in \mathcal{A}} Q(s_t, a, \theta)$
        BS applies $a_t$ beam-pair over the channel, receive
          signal for $(t+1)^{th}$ iteration during uplink
          communication
        UE observes $s_{t+1}$, compute data rate and calculate
          the reward following (19)
        Store transition $e = (s_t, a_t, r_{t+1}, s_{t+1})$ in replay
          memory $D$
        Sample random minibatch of transitions $U(D)$
        Compute Loss and Perform gradient descent for
          $Q(s, a; \theta)$
        Update the target network parameters $\theta_2 = \theta_1$ after
          every $K$ steps
        $n = n + 1$ // Increment episode time
        if done or $(n = N)$ then
            Update the optimal data rate measurement
              $R_{\max}(s_t)$
            break // End episode
    DQN updates the sequence of action beam-pairs for
      (x,y,z) location
    // BS uses the updated sequence on
       next TX request from (x,y,z)
       location
```

problem. We denote $o_t = \{a_{t-1}, s_{t-1}, a_{t-2}, s_{t-2}, \ldots, a_1, s_1\}$ as the observed history of all such state information and past actions. After the 3GPP initial communication procedure with UE, 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 as the consecutive set of such actions until the terminal state following a policy π. The objective of this problem can be formulated as mentioned in (19), where $R_{max}(s_t)$ is the optimal data rate measurement observed among the information history o_t until its previous episode e_π, γ^{i-t} is the discount factor applied on the rewards received from future actions a_i in the episode, $r_{a_t}(t)$ is the reward and $R(a_t)$ is the data rate measurement observed on applying action beam-pair a_t, respectively. We follow DQN approach to solve this RL objective problem.

D. Deep Q-Network

DQN is a value-based approach used generally in the context of RL [29]. The approach learns an optimal approximated policy of states mapping to actions $\pi(s) = a$ by parameterizing and estimating state-action value function $Q(s, a; \theta)$ using deep neural networks (DNN). The architecture of DQN used in this learning formulation is shown in Figure 4. We denote the primary DNN network weight matrix and target DNN network weight matrix as θ and $\overline{\theta}$, respectively [29]. We consider a fully connected DNN for both the networks where $\overline{\theta}$ is updated with primary network parameters θ, after every K iterations. The input of DNN is given by the observed information in s_t. The observation for s_t contains 3D location information and UPA/ULA antenna beam steering information corresponding to the previously selected beam-pairs of that location. The features are extracted separately for each such information, mapped to a high dimensional space and then combined together for learning under the initial layers of DNN network. The intermediate layers are fully connected linear units with rectifier linear units (ReLU) by using the function $f(x) = \max(0, x)$ and the output layer is composed of linear units, which are in one-one corresponding relationship with all available actions in $\mathcal{B}$. We consider initialization of bias and weights of these layers with zeros and Kaiming normalization [41], respectively.

At a time instant t, a_t selects either a random action from $\mathcal{B}$ or perform forward propagation of $Q(s_t, a_t; \theta)$ following ϵ-greedy policy [30]. A memory buffer of experiences $D_t = \{e_1, e_2, e_3, \ldots, e_t\}$, $e_i = (s_i, a_i, r_{i+1}, s_{i+1})$ are collected, where a mini batch of them $U(D)$ are randomly sampled and sent into DQN [29]. During back propagation, a mean squared error (MSE) loss function is computed between primary, target networks and θ is updated using stocastic gradient descent (SGD) [42] and Adam Optimizer [43] as

$$\theta_{t+1} = \theta_t - \xi_{\text{Adam}} \nabla \mathcal{L}^{\text{DQN}}(\theta_t), \tag{20}$$

where ξ_{Adam} is the learning rate, $\nabla \mathcal{L}(\theta_t)$ is the gradient of the DQN loss function, given as

$$\nabla \mathcal{L}^{\text{DQN}}(\theta_t) = \mathop{\mathbb{E}}_{(s_i, a_i, r_{i+1}, s_{i+1})} \Big[(R_{i+1} + \gamma \max_{a} Q(s_{i+1}, a; \overline{\theta}_t) - Q(s_i, a_i; \theta_t)) \nabla_\theta Q(s_i, a_i; \theta_t) \Big], \tag{21}$$

where $\overline{\theta}_t$ is used to estimate future value of Q-function inside $\mathcal{L}^{\text{DQN}}$. Complete steps followed by DQN for RL based beam alignment problem are shown in Algorithm 2. Here, we define episode as the consecutive set of actions applied on the starting state until it reaches the terminal state with maximum beam alignment 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 possible under the chosen antenna configuration. For example, a configuration of 8 ULA antenna elements at both TX and RX can result in maximum episode length of 64.

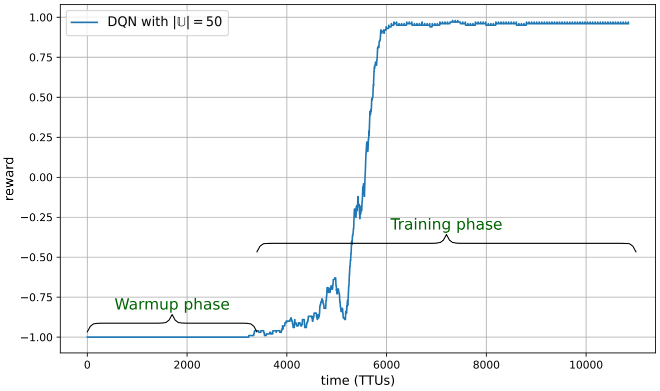


Fig. 3. DQN training procedure consisting of two phases namely, Warmup phase and Training phase.

As the RL learning objective formulation involves both current data rate R_t and best observed data rate $R_{\max}(s_t)$ measurements (shown in (19)), we consider the overall online training procedure of DQN framework under two phases namely, Warmup phase and Training phase as shown in Figure 3. During the Warmup phase, the exploration is set to maximum, in order to observe the best possible data rate for the given UE location by applying maximum episode length of actions. During the Training phase, the algorithm continues to reduce its exploration and move towards exploitation following ϵ-greedy policy. The episode starts with initial random action and applies next actions to reach the terminal state as quickly as possible. The Warmup phase results in extra training time at the start but this is later helpful in quick learning of DQN framework during the training phase. This procedure also favours quick convergence of beam alignment process for the current location based on its neighbourhood beam alignment convergence through experience replay memory buffer, thereby leading to overall faster training of DQN framework for multi-location environment.

IV. Simulation Results

As described in Section II-A and Section III, we formulate the traditional exhaustive method and MAB, RL learning methods using scenario $\mathcal{E}_1$ (7) and $\mathcal{E}_2$ (8), respectively, to solve beam alignment between BS and UE by following 3GPP-type 5G mmWave beam access protocol. In this section, we first

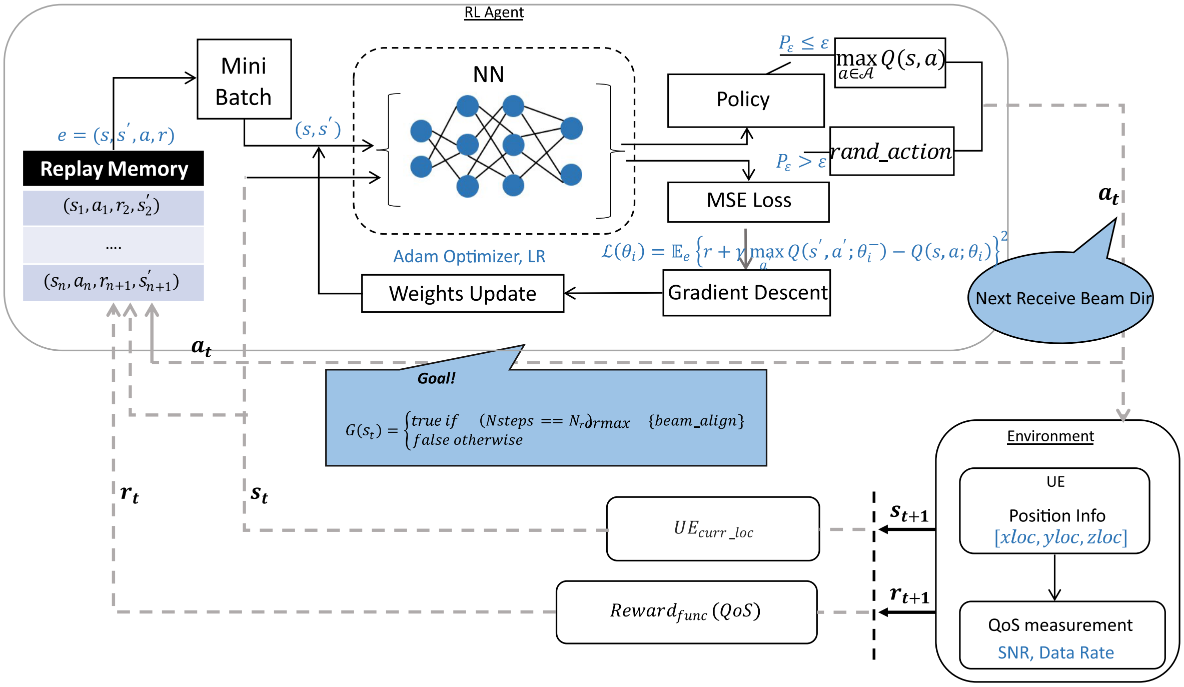

RL Agent
Mini Batch
NN
Policy
$P_\varepsilon \le \varepsilon$
$\max_{a \in \mathcal{A}} Q(s,a)$
$P_\varepsilon > \varepsilon$
$rand\_action$
MSE Loss
$e = (s, s', a, r)$
(s, s')
Replay Memory
(s_1, a_1, r_2, s_2')
....
$(s_n, a_n, r_{n+1}, s_{n+1}')$
Adam Optimizer, LR
$\mathcal{L}(\theta_i) = \mathbb{E}_e \{r + \gamma \max_a Q(s', a'; \theta_i^-) - Q(s, a; \theta_i)\}^2$
Weights Update
Gradient Descent
a_t
Next Receive Beam Dir
a_t
Goal!
$G(s_t) = \begin{cases} true\ if \quad (Nsteps == N_r) ∂ rmax \quad \{beam\_align\} \\ false\ otherwise \end{cases}$
Environment
UE
Position Info
$[xloc, yloc, zloc]$
QoS measurement
SNR, Data Rate
r_t
s_t
UE_{curr_loc}
$Reward_{func}(QoS)$
s_{t+1}
r_{t+1}

Fig. 4. DQN architecture.

investigate the training performance of our proposed RL-based approach against MAB-based learning, to maximize beam alignment under different ideal channel conditions. We note that ideal channel condition corresponds to channel without shadow fading and Rician channel variation. We implement the RL-based beam alignment using DQN, following $(P3)$ objective (19) and steps mentioned in Algorithm 2. Similarly, we implement the MAB-based approach as a CAB problem following $(P2)$ objective (10) and gUCB steps in Algorithm 1. Later, we also focus on analysing DQN approach versus the increase in coverage area, different antenna configurations on both TX and RX and channel conditions. We then combine these observations and perform online beam alignment in the presence of channel variation conditions.

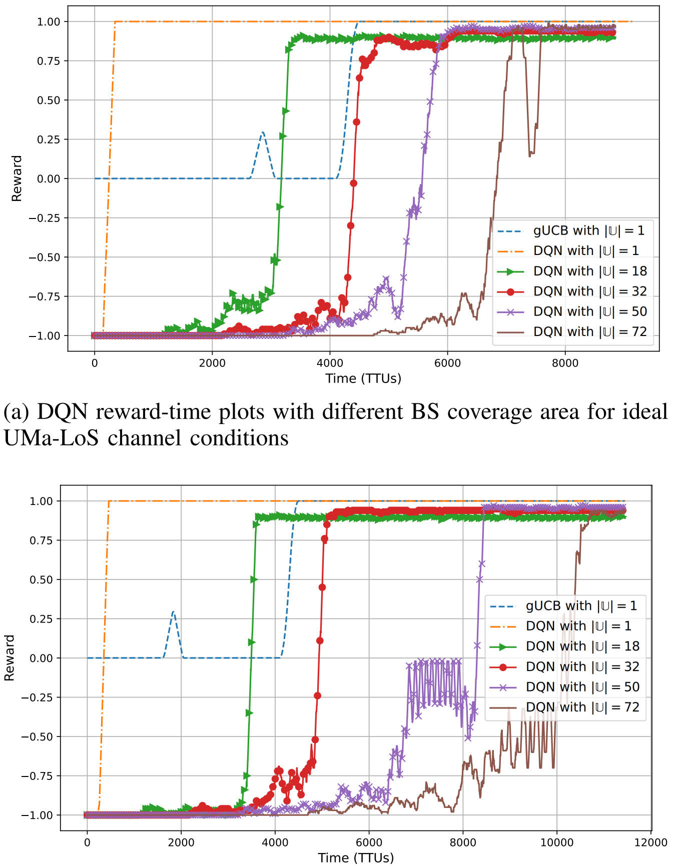
The simulation conditions for numerical results are listed in Table I. We consider a 3D coordinate system with BS located (in m) at $O = [0, 0, 25]$. The coverage area $\mathbb{U}$ around the BS is divided in the form of unit grid cells of fixed size volume $20 \times 20 \times 20$ m^3. We assume that the chosen grid volume size is enough to acquire considerable number of optimal beam-pairs across the simulated state and action space configurations. The coordinate points along x-axis, y-axis and z-axis are defined in steps of unit grid cell size along their dimensions, denoted by $\mathbb{U}_{\text{xloc}}$, $\mathbb{U}_{\text{yloc}}$ and $\mathbb{U}_{\text{zloc}}$, respectively. $\mathbb{U}_{\text{xloc}}$ and $\mathbb{U}_{\text{yloc}}$ are linearly spaced from -60 m to 60 m with step size 20 m while $\mathbb{U}_{\text{zloc}}$ points are considered above the BS height at 41.5 m and 81.5 m. For this setup, the cardinality of the state space is $|\mathbb{U}| = 72$. The transmit power of UE is considered to be 0 dBm. We asumme UE to be hovering around the BS following uniform random mobility in the defined grid environment $\mathbb{U}$. A maximum of 8 antenna elements are considered at both BS (denoted as N_{RX}) and UE (denoted as N_{TX}) following ULA antenna configuration. Thus, the cardinality of the action space is $|\mathcal{A}| = 64$. We consider a mmWave radio signal with 30 GHz carrier frequency and perform the simulations on both aerial LoS and nLoS 3GPP UMa channel conditions. The path loss models for LoS and nLoS are denoted by UMa-avLoS and UMa-avnLoS, respectively, following five parameter alpha-beta-gamma (ABG) model (from [33]) as shown in Table I. A slow variation in the communication channel is also considered using Rician fading with doppler spread of 3 kHz. Besides, a fixed number of random reflection points (4 to 6) at positions close to BS are considered for UMa-nLoS simulations.

To analyse the training performance of the proposed RL-approach, we consider normalized reward plots, beam training overhead plots and average received signal strength (RSS)

TABLE I

SIMULATION PARAMETERS

Parameters	Value
mmWave freq	30 GHz
antenna element spacing d	0.5
Transmit power P_{tx}	0 dBm
Transmit antenna elements N_{tx}	8
Receiving antenna elements N_{rx}	8
Noise Level N_0	-174 dBm
BS location	$[0, 0, 25]$
coverage xloc $\mathbb{U}_{\text{xloc}}$	$[-60, 60, 20]$ m
coverage yloc $\mathbb{U}_{\text{yloc}}$	$[-60, 60, 20]$ m
coverage zloc $\mathbb{U}_{\text{zloc}}$	$[41.5, 81.5]$ m
Cardinality of coverage area $\lvert\mathbb{U}\rvert$	72
Cardinality of Action Space $\lvert\mathcal{A}\rvert$	64
mmWave Channel	UMa
antenna configuration	ULA, UPA
UMa-avLoS Pathloss coefficients	$\alpha : 2.2, \beta : 28.0, \gamma : 2.0,$ $\sigma : 4.64 \exp(-0.0066\ \mathbb{U}_{\text{zloc}}),$ $\kappa : 0.0$ [33]
UMa-avnLoS Pathloss coefficients	$\alpha : (4.6 - 0.7 \log 10(\mathbb{U}_{\text{zloc}})),$ $\beta : -17.5, \gamma : 2.0, \sigma : 6.0,$ $\kappa : 20 \log 10(\frac{40\pi}{3})$ [33]



(a) DQN reward-time plots with different BS coverage area for ideal UMa-LoS channel conditions

(b) DQN reward-time plots with different BS coverage area for ideal UMa-nLoS channel conditions

Fig. 5. DQN reward-time plots for various BS coverage area requirements under different ideal channel conditions.

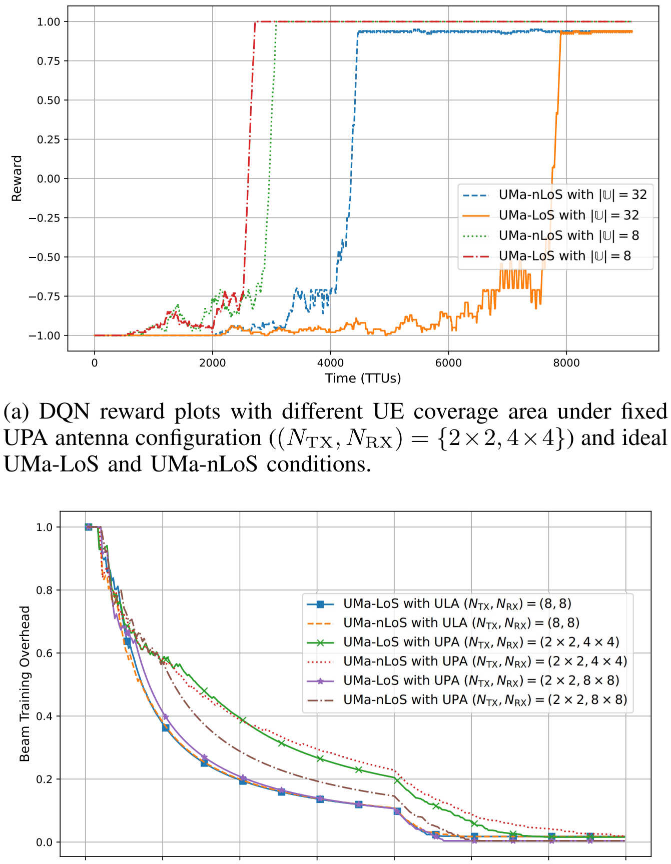
error plots against training time of the learning algorithms. An episodic reward is defined as the total reward accumulated during an episode of the training process. We note that normalized rewards are the episodic rewards normalized over the episodic length, thus are bounded between -1 and 1. The beam training overhead is defined as the average set of beam-pairs taken to maximize beamforming gain for each UE location over the training number of episodes of the RL algorithm. Average RSS error is defined as the mean difference in RSS values of proposed RL approach with respect to exhaustive approach (measured in dB scale) over UE locations. This metric helps us estimate the accuracy of learning a beam-pair in the proposed RL approach with respect to traditional exhaustive approach, at every time instant during the training procedure.

A. DQN Vs. gUCB

As shown in Figure 5, brown and blue plots represent the accumulated rewards over time for DQN and gUCB, respectively, with same action space under single grid element ($|\mathbb{U}| = 1$) coverage area and different ideal channel conditions. Thus, the coverage area considered in these simulations has exactly one UE location for performing beam alignment. As described under Section III-B and III-C, DQN and gUCB learning formulations begin the training procedure with a sequence of exhaustive beam-pairs and eventually learn the efficient beam-pairs based on the observed history for UE location over multiple episodes and time. Also, it is noted that a significant number of reflection points (6 points) are considered under UMa-nLoS based simulation. Our results show that the DQN-based approach quickly accumulate the rewards over TTU time for learning the optimal beam pair compared to the gUCB approach, under both LoS and nLoS simulations. The exploration of beam-pairs following ϵ-greedy policy, simultaneously influencing the state and action spaces for DQN-approach resulting in quicker convergence during training phase and reduced episode length. The gUCB approach involves iterating over exhaustive action space $\mathcal{A}$, thus consuming more time at every episode of the training procedure. On the other hand, DQN uses its warmup phase to quickly determine the best possible data rate measurement for the current state and then learn the determined best rate during training phase, by applying episodes with reduced action sets in a iterative manner. Thus, the reward formulations designed (10), (19) for online learning-based beam alignment schemes, favours DQN over gUCB. Now, with increase in BS coverage area, the training time of gUCB approach increases significantly as every grid element shift under the environment involves substantial change in TX location. We observe that the gUCB-based beam alignment cannot be reliable for multiple locations on convergence without frequent re-training. This is due to substantial change in RSS measurements with respect to (TX,RX) beam pairs for every change in UE location.

B. DQN With Increasing Coverage Area

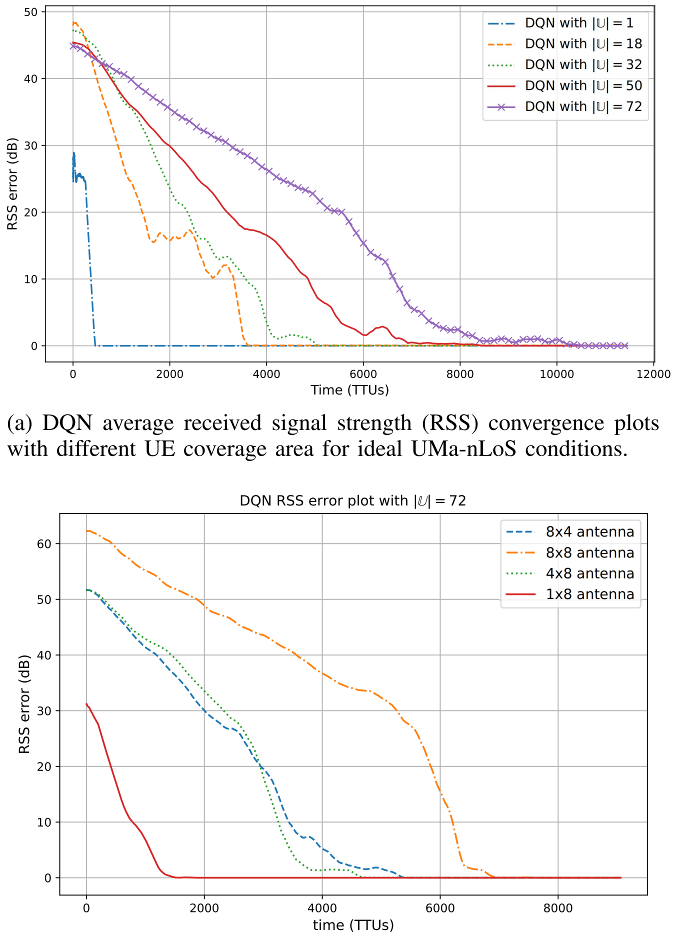
In this subsection, we study the training performance of DQN with increase in coverage area requirements under different ideal UMa channel conditions. Figure 5a, Figure 6a and Figure 5b plots the DQN accumulated normalized rewards over time across different BS coverage area under both UMa-LoS and UMa-nLoS ideal channel conditions for ULA and UPA antenna configurations. It is noted here that we consider around 4 random and fixed reflection points throughout the simulation for all the UMa-nLoS coverage area plots. With increase in coverage area of the BS, UE is provided with more grid elements to hover around and support its radio link. This is also evident under UPA antenna configuration for both UMa-LoS and UMa-nLoS ideal channel conditions. It is observed that the DQN-based approach converges well with different coverage area requirements. Thus, DQN agent with same architecture (as described in Figure 4) can still be used to learn beam alignment between BS and UE across different coverage area and different channel conditions. We note that the accumulated reward plots are shown against number of TTU's for better analysis on convergence. However, during online real-time implementation, the overall convergence time of the DQN agent is spread across multiple initial access procedures at grid locations following 3GPP protocol standards. After the convergence is obtained, BS can quickly align using learnt optimal beams for any UE position within the coverage area without frequent re-training. Also, the



(a) DQN reward plots with different UE coverage area under fixed UPA antenna configuration ($(N_{\mathrm{TX}}, N_{\mathrm{RX}}) = \{2\times 2, 4\times 4\}$) and ideal UMa-LoS and UMa-nLoS conditions.

(b) Overall beam training reduction plots for DQN-based approach with different antenna configuration for $\mathbb{U} = 8$ environment under ideal UMa-LoS conditions.

Fig. 6. DQN performance over multiple coverage area, antenna configuration under UMa-LoS and UMa-nLoS ideal channel conditions.



(a) DQN average received signal strength (RSS) convergence plots with different UE coverage area for ideal UMa-nLoS conditions.

(b) DQN average RSS convergence plots with different ULA antenna configuration for $72u$ environment under ideal UMa-LoS conditions.

Fig. 7. DQN performance over multiple coverage area under UMa-nLoS ideal channel conditions with 6 reflection points.

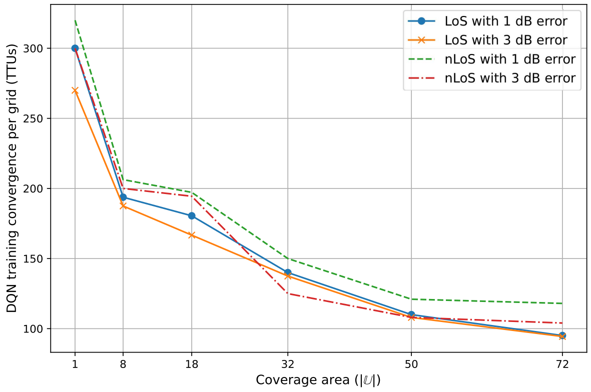


Fig. 8. DQN training convergence per grid element G measured in TTU time (s) vs coverage area of BS (m^2) under different ideal channel conditions and average RSS error (in dB) with respect to traditional exhaustive method.

learning is observed to be relatively quicker in convergence with increase in coverage area of BS under ULA antenna configuration as shown in Figure 8. With increase in coverage area, neighbouring grid elements with similar optimal beam pairs converge DQN faster as part of its MDP process, resulting in average less number of training iterations per grid element alongside convergence.

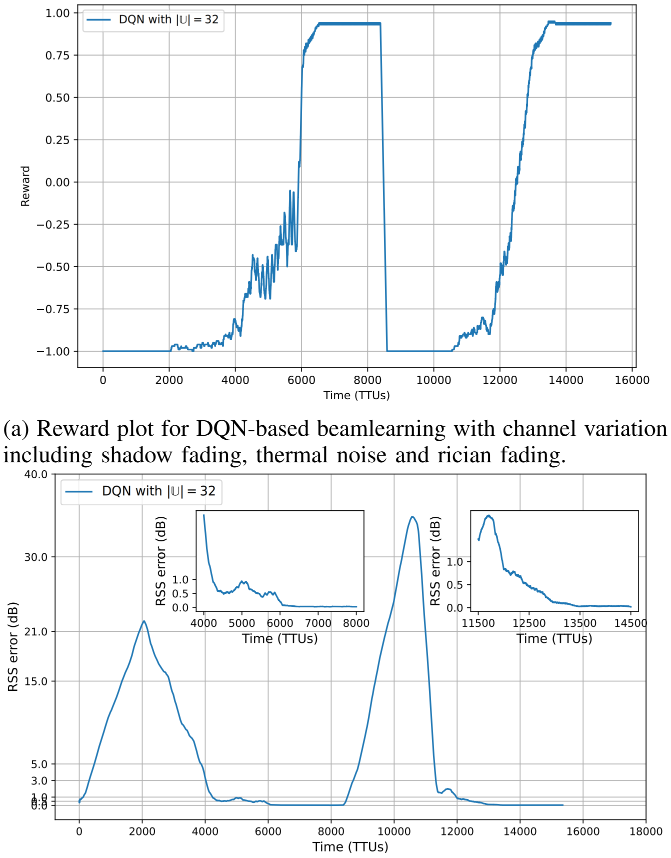
Figure 7a on the other hand, depicts the convergence accuracy of DQN-based beam alignment over the exhaustive method across multiple coverage area requirements. This is recognized by measuring the moving average RSS error of learning agent with that of the exhaustive approach (in dB scale) over TTU time. It is observed that RSS error of DQN agent with respect to exhaustive approach can always converge to 0 dB eventually for multiple coverage area requirements. The DQN convergence under UMa-nLoS for same coverage area is observed to be quicker until $1\mathrm{dB}$ RSS error on average and then convergence speed is reduced drastically until it achieves exact average global convergence. This phenomenon is observed more at higher coverage area simulations as also witnessed in Figure 8. The reason for this could be due to the increase in presence of local optimal RSS values corresponding to each UE grid element location under UMa-nLoS conditions. However, these training convergence results are observed to be significantly better compared to that gUCB-based approach as frequent re-training is avoided. Thus, the DQN agent can achieve faster and reliable beam alignment due to its training procedure (as observed in Section IV-A) and the neighbourhood grid element convergence.

C. DQN Performance Across Multiple Antenna Configurations

In this subsection, we analyze the DQN training performance over different antenna configurations between BS and UE, under the same ($|\mathbb{U}| = 72$) coverage area conditions. In terms of RL formulation (as described in section III), this corresponds to simulating DQN agent across different action spaces $\mathcal{A}$ under the same state space $\mathcal{S}$. The maximum number of ULA antenna elements considered either at TX or RX side is 8. Similarly under UPA, the maximum number of antenna elements at TX side and RX side are set as $2 \times 2 = 4$ and $8 \times 8 = 64$, respectively.

Figure 6b plots the normalized average beam training overhead for a coverage area of $\mathbb{U} = 8$ grid elements under different ULA and UPA antenna configurations. Three different configurations of $\{N_{\mathrm{TX}}, N_{\mathrm{RX}}\}$, i.e., $\{8, 8\}$, $\{(2\times2), (4\times4)\}$, $\{(2 \times 2), (8 \times 8)\}$ are considered under both UMa-LoS and UMa-nLoS ideal channel conditions. For the same coverage area the proposed approach is shown to converge well with significant reduction in overall beam training over time, using the same DQN architecture. Compared to ULA, the beam training overhead for UPA is shown to be relatively higher especially during nLoS conditions. This is due to more unique beam pairs required while using both azimuthal and elevation angles for UAV locations under the coverage area. However, increasing action spaces, the normalized beam training overhead decreases relatively quicker as shown for $\{N_{\mathrm{TX}}, N_{\mathrm{RX}}\} = \{(2\times2), (8\times8)\}$ configuration. This shows that the proposed algorithm converge faster by eliminating unwanted beam pairs under AoA-AoD distribution for UAV positions in mmWave scenarios. However, there is less reduction in overhead when there are more sub-optimal beam pairs as observed with wider beams under $\{N_{\mathrm{TX}}, N_{\mathrm{RX}}\} = \{(2\times2), (4\times4)\}$ configuration. This shows that the proposed approach converge faster (in terms of number of episodes) for higher antenna configurations despite the increase in action space. We note that overall beam training overhead still depends on both number of episodes and each episode length. Once converged, the maximum beam forming gain for any random UAV position within coverage area can be obtained in very few steps using DQN-based beam alignment procedure, significantly minimizing the beam training overhead.

Figure 7b depicts DQN average RSS error (in dB) plots under ideal UMa-LoS for 4 different ULA configurations of $N_{\mathrm{TX}} \times N_{\mathrm{RX}}$, i.e., $4\times4, 4\times8, 8\times4$,and 8×8. It is observed that the DQN learning time increases with increase in exhaustive set of actions under the $(\mathcal{S}, \mathcal{A})$ formulation. From the blue and green plots in Figure 7b, it is evident that DQN performs similar during training under same state $\mathcal{S}$ and action space $\mathcal{A}$ even with different antenna configurations. The reason for this is intuitive as the increase in action space also increases the episode length of DQN creating significant impact in training iterations especially during warmup phase of DQN. It is also observed that RSS error of DQN agent with respect to exhaustive approach can always converge to 0 dB eventually for multiple antenna configurations. Thus, the DQN agent with the same learning framework is still generic to learn under different antenna configurations.



(a) Reward plot for DQN-based beamlearning with channel variation including shadow fading, thermal noise and rician fading.

(b) RSS Error plot for DQN based beam learning with channel variation including shadow fading, thermal noise and rician fading.

Fig. 9. Performance of DQN based beam learning in the presence of channel variation with UMa-nLoS to UMa-LoS change in environment.

D. DQN Training Performance With Channel Variation and Shadow Fading

Having realized the DQN training performance for different parameters such as coverage area, channel conditions, antenna configuration etc. It is now more evident that the proposed DQN can effectively learn beam alignment for the defined environment. In this subsection, we plot DQN training performance in real-time conditions in an online manner by considering change in channel conditions, thermal noise, slow fading and slow channel variation as shown in Figure 9.

Figure 9a and Figure 9b plot the rewards of the DQN learning agent and the RSS errors of agent with respect to exhaustive approach, respectively. For these simulations, we consider a ($|\mathbb{U}| = 32$) coverage area environment between BS and UE. The environment is equipped with thermal noise, shadow fading, UMa-nLoS to UMa-LoS channel conditions along with a slow rician channel variation. The parameters used for this simulation are disclosed under Table. I. Depending on the defined coverage area, different path loss models corresponding to aerial view of UE for both UMa-nLoS and UMa-LoS conditions are included as well following 3GPP standard mentioned in [33]. It is noted here that same set of hyper-parameters are used for DQN during both LoS and nLoS channel conditions during this simulation. We observe that

DQN agent performs similar to previous results, converging well under both varying channel conditions. The disturbance in the smoothness of the reward plots observed in Figure 9a could be due to the impact of channel variation under nLoS conditions. This concludes that the proposed DQN framework is generic, converge faster than the gUCB approach and can also learn beam alignment in an online manner under varying channel conditions.

V. Conclusion and Future Work

In this paper, we developed a learning-based beam alignment framework for mmWave MIMO uplink BS-UAV communication. We proposed a RL-based framework using DQN to maximize the beam alignment for any UAV position within the BS coverage area following 3GPP standard. We also analyze the same DQN architecture over different coverage requirements, antenna configurations and channel conditions to demonstrate the generalization of the proposed RL-based framework for beam learning. Our results show that, the proposed approach significantly outperforms the MAB-based method in terms of the convergence and also learns optimal beam pairs comparable to that of the traditional exhaustive search method. We further demonstrate that the DQN-based beam alignment can be performed in an online manner under varying channel conditions. Thus, we conclude that our proposed DQN framework is generic and converge faster than the MAB-based (gUCB) approach. The framework also learn optimal beam alignment in an online manner under real-time conditions.

Even though we demonstrate promising results following 3GPP standard in our work, we are yet to explore the full capabilities of these generic learning architectures for mmWave UAVs. From that perspective, a promising future direction is to predict beams over unexplored grid locations, support higher MIMO antenna configurations, multi BS-UAV distributed environments by optimizing parameters such as better connectivity, large number of beam-directional pairs, interference mitigation etc. Such multi-objective optimization is quite challenging and should be addressed in future.

Appendix A
Proof of Theorem 1

Proof: From (12), it is understood that bounding the regret can be done by simply bounding each $\mathbb{E}[T_i(n)]$ if $\Delta_i > 0$. We will now compute a bound for $\mathbb{E}[T_i(n)]$. A necessary condition for the beam pair with index i to be selected over optimal beam pair index i^* is $\mathrm{UCB}_i > \mathrm{UCB}_{i^*}$. Once a beam pair with index i is selected $\tau(\tilde{p_i})$ times (as shown in (15)), the number of times i is selected instead of i^* until time instant n after its p episodes is bounded by

$$\begin{aligned} T_i(n) &\le \tau(\tilde{p_i}) \\ &\quad + \sum_{p>\tilde{p_i}} (\tau(p)-\tau(p-1))\mathbf{1}\big\{\mathrm{UCB}_i > \mathrm{UCB}_{i^*}\big\}, \end{aligned} \tag{22}$$

$$\begin{aligned} T_i(n) \le \tau(\tilde{p_i}) + \sum_{p>\tilde{p_i}} (\tau(p)-\tau(p-1))\mathbf{1}\big\{\bar{X}_{i,\tau(p-1)} + a_{t,p-1} \\ \ge \bar{X}^*_{\tau(p)} + a_{t,p}, t \ge \tau(\tilde{p_i}) + \tau(p)\big\}. \end{aligned} \tag{23}$$

Now, beam index pair i on completing p episodes has the following chain of implications.

$$\begin{aligned} \implies &\exists\ k \ge 0, \exists t \ge \tau(p-1) + \tau(k) \\ &\quad \text{s.t. } (\bar{X}_{i,\tau(p-1)} + a_{t,p-1}) \ge (\bar{X}^*_{\tau(k)} + a_{t,k}), \end{aligned} \tag{24}$$

$$\begin{aligned} \implies &\exists t \ge \tau(p-1) \\ &\text{s.t. } \ (\bar{X}_{i,\tau(p-1)} + a_{t,p-1}) \ge r^* - \alpha\Delta_i/2, \\ &\qquad \text{(or) } \exists k \ge 0, \exists t' \ge \tau(p-1) + \tau(k) \\ &\qquad \text{s.t. } (\bar{X}^*_{\tau(k)} + a_{t',k}) \le r^* - \alpha\Delta_i/2, \end{aligned} \tag{25}$$

$$\implies (\bar{X}_{i,\tau(p-1)} + a_{n,p-1}) \ge r^* - \alpha\Delta_i/2, \tag{26}$$

$$\text{(or) } \exists k \ge 0,\ \text{s.t.}(\bar{X}^*_{\tau(k)} + a_{\tau(p-1)+\tau(k),k}) \le r^* - \alpha\Delta_i/2. \tag{27}$$

Using (22)-(27), $\mathbb{E}[T_i(n)]$ can be bounded by

$$\begin{aligned} \mathbb{E}[T_i(n)] &\le \tau(\tilde{p_i}) \\ &\quad + \sum_{p>\tilde{p_i}} (\tau(p)-\tau(p-1))\ \mathbb{P}\{(26) \text{ is True}\} \\ &\quad + \sum_{p>\tilde{p_i}} \sum_{k\ge 0} (\tau(p)-\tau(p-1))\ \mathbb{P}\{(27) \text{ is True}\}. \end{aligned} \tag{28}$$

From (15) and (16), $a_{n,p-1}$ can also be bounded as $a_{n,p-1} \le \Delta_i\sqrt{\frac{1+\alpha}{1+4\alpha}}$. Now, by using Chernoff-Hoeffding inequality, first and second summation terms in (28) can also be bounded as the following ([44], p.251-252)

$$\begin{aligned} &\sum_{p>\tilde{p_i}} (\tau(p)-\tau(p-1))\ \mathbb{P}\{(26) \text{ is True}\} \\ &\qquad \le \frac{(1+\alpha)e}{(\Delta_i\alpha)^2}, \end{aligned} \tag{29}$$

$$\begin{aligned} &\sum_{p>\tilde{p_i}} \sum_{k\ge 0} (\tau(p)-\tau(p-1))\ \mathbb{P}\{(27) \text{ is True}\} \\ &\qquad \le (\frac{1+\alpha}{\alpha})^{(1+\alpha)}\left[1 + \frac{11(1+\alpha)}{5(\alpha\Delta_i)^2\ln(1+\alpha)}\right]. \end{aligned} \tag{30}$$

Combining (29) and (30) in (28) and substituting in (12), pseudo regret can be upper bounded ([44], p.253) as obtained in (17). □

References

[1] P. Susarla, B. Gouda, Y. Deng, M. Juntti, O. Silven, and A. Tolli, "DQN-based beamforming for uplink mmWave cellular-connected UAVs," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Aug. 2021, pp. 1–6.

[2] H. Shakhatreh *et al.*, "Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48572–48634, 2019.

[3] S. D. Muruganathan *et al.*, "An overview of 3GPP release-15 study on enhanced LTE support for connected drones," 2018, *arXiv:1805.00826*.

[4] L. Zhang *et al.*, "A survey on 5G millimeter wave communications for UAV-assisted wireless networks," *IEEE Access*, vol. 7, pp. 117460–117504, 2019.

[5] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, Feb. 2018.

[6] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[7] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, Nov. 2015.

[8] "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN)," Version 15.10.0, Tech. Rep. 36.300, Jul. 2020.

[9] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[10] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, Jr., "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4042–4058, Dec. 2017.

[11] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, Jr., "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30507–30526, 2019.

[12] F. Devoti, I. Filippini, and A. Capone, "Facing the millimeter-wave cell discovery challenge in 5G networks with context-awareness," *IEEE Access*, vol. 4, pp. 8019–8034, 2016.

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

[14] Y. Wang, N. J. Myers, N. González-Prelcic, and R. W. Heath, Jr., "Site-specific online compressive beam codebook learning in mmWave vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3122–3136, May 2021.

[15] J. Zhao, F. Gao, W. Jia, S. Zhang, S. Jin, and H. Lin, "Angle domain hybrid precoding and channel tracking for millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6868–6880, Oct. 2017.

[16] C. Zhang, W. Zhang, W. Wang, L. Yang, and W. Zhang, "Research challenges and opportunities of UAV millimeter-wave communications," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 58–62, Feb. 2019.

[17] J. Zhao, F. Gao, L. Kuang, Q. Wu, and W. Jia, "Channel tracking with flight control system for UAV mmWave MIMO communications," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1224–1227, Jun. 2018.

[18] J. Saloranta, G. Destino, and H. Wymeersch, "Comparison of different beamtraining strategies from a rate-positioning trade-off perspective," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

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

[20] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5689–5701, Sep. 2017.

[21] T. Kadur, H.-L. Chiang, and G. Fettweis, "Effective beam alignment algorithm for low cost millimeter wave communication," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2016, pp. 1–5.

[22] P. Susarla *et al.*, "Learning-based trajectory optimization for 5G mmWave uplink UAVs," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–7.

[23] "Study on 5G new radio(NR) access technology," Version 15.0.0, Tech. Rep. 38.912, May 2017.

[24] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2020.

[25] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Dec. 2016.

[26] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu, "Machine learning for vehicular networks: Recent advances and application examples," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 94–101, Jun. 2018.

[27] 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.

[28] Y. Wang, M. Narasimha, and R. W. Heath, Jr., "mmWave beam prediction with situational awareness: A machine learning approach," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, Jun. 2018, pp. 1–5.

[29] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, Feb. 2015.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, Nov. 2018.

[31] J. Zhang, Y. Huang, Y. Zhou, and X. You, "Beam alignment and tracking for millimeter wave communications via bandit learning," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5519–5533, Sep. 2020.

[32] J. Zhang, Y. Huang, J. Wang, X. You, and C. Masouros, "Intelligent interactive beam training for millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2034–2048, Mar. 2020.

[33] "Study on enhanced LTE support for aerial vehicles," Version 15.0.0, Tech. Rep. 36.777, Jan. 2018.

[34] J. Campos, "Understanding the 5G NR physical layer," Keysight Technol., Santa Rosa, CA, USA, Nov. 2017. [Online]. Available: https://www.keysight.com/fi/en/assets/9921-03326/training-materials/Understanding-the-5G-NR-Physical-Layer.pdf

[35] "Study on channel model for frequencies from 0.5 to 100 GHz," Version 14.3.0, Tech. Rep. 38.901, Jan. 2018.

[36] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[37] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 2393–2401.

[38] I. Aykin, B. Akgun, M. Feng, and M. Krunz, "MAMBA: A multi-armed bandit framework for beam tracking in millimeter-wave systems," in *Proc. IEEE Conf. Comput. Commun.*, Toronto, ON, Canada, Jul. 2020, pp. 1469–1478.

[39] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2487–2500, Dec. 2018.

[40] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," 2012, *arXiv:1204.5721*.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 1026–1034.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[44] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, May 2002.

Praneeth Susarla (Graduate Student Member, IEEE) received the master's degree in information technology specialized in computer vision and embedded systems from the International Institute of Information Technology Bangalore (IIIT Bangalore), India, in 2017. He is currently pursuing the Ph.D. degree in applying artificial intelligence to 5G and beyond mmWave beamforming problems under Prof. Olli Silven and Prof. Markku Juntti. His research interests are in artificial intelligence, radio communications, computer vision, and embedded systems. He is actively involved in Nokia-5G Champion Project with the Center for Wireless Communications (CWC), University of Oulu, from 2017 to 2018. He received the Best Student Paper Award at IAPR-MedPRAI, 2018. He was a MITACS Global Research Internee at the University of British Columbia, Vancouver, in Summer 2016. He was a Visiting Researcher at Kings College London, in Summer 2019, as a part of collaboration with Prof. Toktam Mahmoodi and Prof. Yansha Deng.

Bikshapathi Gouda (Graduate Student Member, IEEE) received the master's degree in communication systems from the Indian Institute of Technology Madras, India, in 2013. He is currently pursuing the Ph.D. degree with the Centre for Wireless Communications, University of Oulu, Finland. In 2013, he joined a startup company to work on the physical-layer design of 802.11ad WiFi Technologies. In 2017, he moved to Cypress Semiconductor, India, to work on 802.11ac/ax WiFi Technologies. His research interests include wireless communications and signal processing for distributed multi-antenna systems.

Yansha Deng (Member, IEEE) is currently a Senior Lecturer (Associate Professor) with the Department of Engineering, King's College London, London, U.K. She has secured more than 2.3 million of research funding as the Principal Investigator and has received EPSRC NIA Award. She has published more than 90 journal articles and more than 40 IEEE/ACM conference papers. She is currently a Senior Editor of IEEE COMMUNICATIONS LETTERS since 2020, an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS since 2017, an Associate Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS since 2022, an Associate Editor of IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING since 2022, an Associate Editor of IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATIONS since 2019, an Associate Editor of IEEE Open Journal of Communications Society since 2019, and the Vertical Area Editor of *IEEE Internet of Things Magazine* since 2021.

Markku Juntti (Fellow, IEEE) received the M.Sc. (EE) and D.Sc. (EE) degrees from the University of Oulu, Oulu, Finland, in 1993 and 1997, respectively.

He was with the University of Oulu from 1992 to 1998. In academic year 1994–1995, he was a Visiting Scholar at Rice University, Houston, Texas, USA. From 1999 to 2000, he was a Senior Specialist with Nokia Networks, Oulu. Since 2000, he has been a Professor of communications engineering with the University of Oulu, Centre for Wireless Communications (CWC), where he leads the Communications Signal Processing (CSP) Research Group. He also serves as Head of CWC—Radio Technologies (RT) Research Unit. His research interests include signal processing for wireless networks as well as communication and information theory. He is the author or coauthor in almost 500 papers published in international journals and conference records as well as in books *Wideband CDMA for UMTS* from 2000 to 2010, *Handbook of Signal Processing Systems* in 2013 and 2018, and *5G Wireless Technologies* in 2017. He is also an Adjunct Professor with the Department of Electrical and Computer Engineering, Rice University.

Dr. Juntti is an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and served previously in similar role in IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was a Secretary of IEEE Communication Society Finland Chapter from 1996 to 1997 and the Chairperson for years 2000 to 2001. He has been a Secretary of the Technical Program Committee (TPC) of the 2001 IEEE International Conference on Communications (ICC), and the Chair or Co-Chair of the Technical Program Committee of several conferences including 2006 and 2021 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), the Signal Processing for Communications Symposium of IEEE GLOBECOM 2014, Symposium on Transceivers and Signal Processing for 5G Wireless and mm-Wave Systems of IEEE GlobalSIP 2016, ACM NanoCom 2018, and 2019 International Symposium on Wireless Communication Systems (ISWCS). He has also served as the General Chair of 2011 IEEE Communication Theory Workshop (CTW 2011) and 2022 IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC).

Olli Silvén (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Oulu, Oulu, Finland, in 1982 and 1988, respectively. Since 1996, he has been a Professor of signal processing engineering with the University of Oulu. He has contributed to the development of solutions from real-time 3-D imaging in reverse vending machines to IP blocks for mobile video coding. His research interests include ultra-energy-efficient embedded signal processing and machine vision system design.

Antti Tölli (Senior Member, IEEE) received the D.Sc. (Tech.) degree in electrical engineering from the University of Oulu, Oulu, Finland, in 2008. He is a currently a Professor with the Centre for Wireless Communications (CWC), University of Oulu. From 1998 to 2003, he worked at Nokia Networks, Finland and Spain, as a Research Engineer and the Project Manager. In May 2014, he was granted a five year (2014–2019) Academy Research Fellow post by the Academy of Finland. During the academic year 2015–2016, he visited at EURECOM, Sophia Antipolis, France. From August 2018 to June 2019, he was visiting at the University of California Santa Barbara, USA. He has authored numerous papers in peer-reviewed international journals and conferences and several patents all in the area of signal processing and wireless communications. His research interests include radio resource management and transceiver design for broadband wireless communications with a special emphasis on distributed interference management in heterogeneous wireless networks. From 2017 to 2021, he served as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING.