Learning-Based Trajectory Optimization for 5G
mmWave Uplink UAVs

Praneeth Susarla*, Yansha Deng†, Giuseppe Destino*†, Jani Saloranta*,
Tōktām Mahmoodi†, Markku Juntti*, Olli Silven*
*University of Oulu, Finland
†King’s College London, United Kingdom
email: praneeth.susarla@oulu.fi

Abstract—A Connectivity-constrained based path planning
for unmanned aerial vehicles (UAVs) is proposed within
the coverage area of a 5G NR Base Station (BS) that uses mmWave
technology. We consider an uplink communication between UAV
and BS under multipath channel conditions for this problem.
The objective is to guide a UAV, starting from a random location
and reaching its destination within the BS coverage area, by
learning a trajectory alongside achieving better connectivity.
We propose simultaneous learning-based path planning of UAV
and beam tracking at the BS side under urban macro-cellular
(UMa) pathloss conditions, to reduce its sweeping time with
apriori computational overhead using the deep reinforcement
learning method such as Deep Q-Network (DQN). Our results
show that our proposed learning-based joint path planning and
beam tracking method is on par with the learning-based shortest
path planning, besides beam tracking comparable to heuristic
exhaustive beam searching method.

Index Terms—5G, mmWave, reinforcement learning, UAVs,
path planning, beamforming

I. INTRODUCTION

Over the last few years, there is tremendous research
interests towards integrating unmanned aerial vehicles (UAVs)
into cellular networks using fifth Generation (5G) and beyond
wireless communications [1], [2]. The deployment of cellular-
enabled UAV-User Equipment (UE)s (hereafter addressed as
UAVs) adds an additional dimension of challenges in the design
and planning of these communication systems for high altitude
coverages in a myriad of civil applications, such as traffic
surveillance, mineral exploration, internet drone delivery sys-
tems, etc. [3], [4]. For instance, Amazon is projected to deploy
a commercial fleet of 450,000 drones under “Amazon Prime
Air” worldwide delivery operating service by 2020 [5].

Millimeter wave (mmWave) beamforming communication
is considered as one of the key innovations for 5G networks
[6], [7]. The wide spectrum (ranging from 30 GHz to 300
GHz) of mmWave frequencies with beamforming has provided
a promising way for sustaining a real-time ultra-high speed
transmissions for UAVs. Besides, mmWave provides high-
capacity and line-of-sight (LoS) dominated connectivity for
UAVs, it also suffers from many challenges, such as rapid
channel variation, blockage effects, atmospheric attenuations,
rage limits etc. [8], [9], leading to substantial interference
problems especially in multi-cell scenarios. The recent studies
have even reported nearly 150 UAV crashes occurred due
to loss of communication with ground base stations [10].

As there is an expected increase of UAVs in near future,
it is of significant importance to support high-performance
communications between UAVs and Base Station (BS).

Connectivity-constrained trajectory optimization for UAV-
mounted relays and UAV-mounted base stations in cellular
networks has been widely investigated in [11]–[14]. In
[11], a minimal delay trajectory design for UAV relays was
proposed to ferry data from multiple sources to destination
using reinforcement learning (RL) algorithm. In [13], a multi-
UAV enabled multi user communication system was jointly
optimized with UAV trajectory and communication resource
allocation to maximize throughput over all users for downlink
transmission scenario. The authors in [14], proposed a deep
RL framework based on echo state network (ESN) cells for
optimizing the trajectories in a multi-cellular UAV scenario.
This approach allowed UAVs act as individual RL agents to
minimize their interference on ground network under latency.

In this paper, we consider a connectivity-constrained based
trajectory optimization for any cellular-enabled UAV within
the coverage area of mmWave BS, using deep RL framework.
Unlike in previous works, here the mmWave BS acts as RL
agent to learn communication-aware optimal trajectory and
optimal beam tracking for UAV. Some literature has recently
studied the use of RL for mmWave beam learning in an online
manner. [15], [16]. The deep RL framework offers a generic
and scalable solution, where BS acts as a commander for
all UAVs in a multi-UAV scenario. Besides, the smart BS
approach for cellular-enabled UAVs can comply well with
current UAV battery standards [17].

The main contribution of this paper is that we propose
a deep RL based generic framework at BS side, to jointly
optimize path planning and beam tracking for any uplink
mmWave cellular-enabled UAV. This communication is sim-
ulated by considering a multi-path channel model, Multiple
Inputs Multiple Outputs (MIMO) beamformers and UAV-BS
environment. The simulated environment is used for offline
training of BS RL agent before deployment similar to prior
works [11]–[14]. The trained agent is used to benchmark
our results against learning-based shortest path planning with
heuristic exhaustive beam searching method.

The rest of the paper is organized as follows. Section II
presents the system and communication model with the prob-
lem formulation. Section III describes the implementation of
our deep RL framework and UAV environment. Section IV elaborates our simulations and benchmark results. Section V summarizes our conclusion and future work.

II. PROBLEM FORMULATION AND SYSTEM MODEL

As illustrated in Fig 1, we consider a mmWave uplink multi-path (both LoS and Non-Line-of Sight (NLoS)) radio beamforming communication between a fixed BS and a moving unmanned aerial vehicle (UAV), following 5G protocol standards as mentioned in [18]. UAV starts from a random source location with the goal to reach destination by assuming all the locations hovered by UAV (as UE) are within the coverage area of BS. The objective of this problem is to guide the UAV via BS in reaching destination, by predicting the next best UE location as well as BS radio frequency (RF) beam direction based on its connectivity-constraints.

A. System Model

We consider a multi-antenna UE, multi-antenna BS scenario and also conceptualize the coverage area U around BS into a grid as shown in Fig 1(a). The BS acts as fixed serving node located at $O(0,0)$. UE transmits a radio signal in multiple beam directions following a codebook $B$ defined as

$$b_i = (i - 1) \frac{\pi}{N-1}, 1 \leq i \leq N,$$  

(1)

where $b_i$ represents a RF radio beam direction with a fixed narrow beam width($\frac{\pi}{N-1}$), $N$ represents the number of codebook directions in $B$. BS receives the radio signal through one of its multiple beam directions each time, following the same codebook set $B$. UE starts from source location $UE_s$ and moves with velocity $v = (v_x, v_y, v_z)$ ($|v_x|, |v_y|, |v_z| \in V$, where $V$ denote the set of UE pre-defined speed values) towards a defined target location $UE_t$, following a certain path $p$. Meanwhile, the RX radio unit starts with a random beam $b_p \in B$ at time $t = 0$ and learns to choose the beam direction $b_k$ at every time, until reaching the target location at along the path $p$. Let $M \in \{ 'L' , 'R' , 'U' , 'D' \}$ represents the set of possible UE move directions under the coverage area $U$. If $UE_s$ is denoted as $(x_1, y_1, z_1)$, then the position of UE via BS at any time instant $t$ is given by

$$UE_t = (x_1 + v_x t, y_1 + v_y t, z_1 + v_z t), \forall 1 \leq t \leq t_p,$$  

(2)

where $(x_1 + v_x t, y_1 + v_y t) \in U$ and $(z_1 + v_z t) \in Z$, depending on UE move direction and $t_p$ is the maximum time step limit. Here, $U$ and $Z$ is the coverage area of the the serving Node BS and altitude range of UAV respectively.

B. Communication Model

We consider an uplink MIMO mmWave communication between UE at location $UE_i \in \mathbb{R}^2$ and BS at location $O \in \mathbb{R}^2$, as shown in Fig. 1(b). UE acts as transmitter (TX), BS as receiver (RX) and are equipped with Uniform Linear Array (ULA) structure of $N_t$ and $N_r$ antennas respectively. We assume $UE_0$ is unknown and estimated using (2) while BS location is known as $O(0,0)$. The radio channel considered in this problem is a multi-ray link (LoS and NLoS) with propagation delay $\tau_m \forall 0 \leq m \leq M$, where $M$ represents the number of reflection points between UE and BS. Let $\alpha \in [0, 2\pi]$ be the angle of rotation of UE antenna array with respect to y-axis. $\theta_{tx,m}, \theta_{rx,m}$ are the Angle of Departure (AoD) and Angle of Arrival (AoA) of $m^{th}$ communication link between BS and UE respectively. Using these notations, we can now define $\alpha = \pi + \theta_{tx,m} - \theta_{rx,m}$ and propagation delay $\tau_m$ with velocity of light $c$ is given by $\tau_m = \frac{\|UE_{t,m}\|}{c}$.

UE transmits radio signals in all codebook directions at a carrier frequency $f_c$ (or wavelength $\lambda$) with bandwidth $W$. The BS receives carrier $f_c$ signal through one of its multiple beams defined under same direction set $B$ using (1). Based on this, we employ a narrow band receiver signal model given by

$$y(t) = \sum_{m=0}^{M} \sqrt{P_{tx}/\beta} \ w^H \ a_R(\theta_{tx,m}) a_T^H(\theta_{tx,m}) f_x(t - \tau_m) + w^H \ n(t),$$  

(3)

where $P_{tx}$ is transmission power, $\beta$ is the antenna channel gain with UMa environment multi-path loss conditions [19], $w \in \mathbb{C}^{N_r}$, $f \in \mathbb{C}^{N_t}$ are the transmit and receive unit-norm beamforming vectors, $n(t) \in \mathbb{C}^{N_r}$ is a Gaussian noise vector with zero mean and two-sided power spectral density $\frac{N_0}{2}$, $x(t)$ represents one Orthogonal Frequency Division Multiple access (OFDM) symbol of the time-domain transmitted signal with bandwidth $W$ and time period $T_{sym}$ with $\frac{1}{T_{sym}} \int_0^{T_{sym}} \|x(t)\|^2 dt = 1$, $a_R(\theta_{tx,m}) \in \mathbb{C}^{N_r}$, $a_T(\theta_{tx,m}) \in \mathbb{C}^{N_t}$ are the array response vectors for $\theta_{tx,m}$ and $\theta_{tx,m}$ along $m^{th}$ communication path. Here, $a(\theta)]_{n=1}^{N-1} = \frac{1}{\sqrt{N}} \ exp(i \frac{2\pi Nld}{\lambda} \ sin(\theta))$, where $\theta = \theta_{tx,m}, N = N_r$ and $\theta =$
\( \theta_{tx,m}, N = N_t \) for \( a_{RT}(\theta_{tx,m}) \) and \( a_{RT}(\theta_{tx,m}) \), respectively. If \( N_{FFT} \) represents the number of OFDM subcarriers, then signal-to-noise ratio (SNR) can be defined as

\[
\text{SNR} = \sum_{n=0}^{N_{FFT}} \sum_{m=0}^{M} \frac{|\beta|^2 R_{tx} |w^H a_{RT}^*(\theta_{tx,m})|^2 |f^H a_T^*(\theta_{tx,m})|^2}{N_0 W}
\]

and overall rate measurement \( R \) is given by

\[
R = W \log(1 + \text{SNR}).
\]  

\( C. \) Problem Formulation

We now formulate the problem to guide UE along a path via BS, with the goal to reach its destination using its mobility and data rate measurements (4). At any time instant \( t \), we define the parameters, \( \alpha_t = \{R_{t-1}, \alpha_{t-2}, \ldots, \alpha_1, R_1 \} \) that denote observation history of previous data rate measurements \( R_t \) and \( \alpha_t \) are the action parameters \( \alpha_t = (\eta_t, b_t) \) where \( \eta_t \in \mathcal{M} \) represents UE move direction and \( b_t \in \mathcal{B} \) represent receiver beam direction from (1).

The BS aims at maximizing the long-term average of both connectivity and path guidance to UE by following a stochastic policy \( \pi \), mapping current observation history \( \alpha_t \) to selected action probabilities \( \alpha_t \). The goal of BS is to serve UAV with an optimal path guidance as well as provide better connectivity at every instant along the path. We consider the connectivity and path constraints in terms of data rate measurements between UAV-BS and distance measurements with respect to destination location respectively. Therefore, in order to achieve the desired goal, we need to find an optimal policy that can maximize the data rate measurements by moving closer towards destination at every instant along the path. Based on this, optimization problem can now be formulated as

\[
(P1): \max_{\{\pi(\alpha_t | \alpha_t)\}} \sum_{k=t}^{\infty} \gamma^{k-t} E_{\pi}(\frac{R_t}{\delta_t}),
\]

where \( \gamma \in [0, 1) \) is the discount factor for future constraints along the path reaching destination. \( R_t \) corresponds to the data rate measurements of UE \( d_t \) while \( \delta_t \) is the UE \( d_t \) distance from destination. Here, the expectation operator \( E \) helps in providing the long-term average both connectivity and path constraint measurements as \( t \to \infty \). Since, the dynamics of the system is Markovian over time and is defined by DQN flowchart to be further discussed below as shown in Fig. 2, this is considered as a Partially Observable Markov Decision Process (POMDP) problem [20] that is generally intractable. Approximate solution using deep RL method will be discussed in Section. III.

\( \text{III. IMPLEMENTATION} \)

As shown in Fig. 2, we apply a DQN [21] learning framework for the uplink radio communication with BS as a learning Q-agent and UE as the environment. The objective is to predict the immediate UE direction along with its RX beam direction using its distance from target and current data rate measurement information.

We consider \( s_t \in S, \alpha_t \in \mathcal{O}, \alpha_t \in \mathcal{A} \) and \( r_t \in \mathcal{R} \) as any state, observation, action and reward at time instant \( t \), from their corresponding sets, respectively. The BS-agent observes the current state \( s_t \) corresponding to the observations history \( \alpha_t \) and selects a specific action \( \alpha_t \in \mathcal{A}(s_t) \) following standard DQN procedure. Here, \( s_t \) represents set of indices mapping to UE \( d_{t-1} \) location and \( \alpha_t \) corresponds to (\( \eta_t, b_t \)) pair described under Section II-C. Once an action \( \alpha_t \) is performed on the environment, the Q-agent will receive a scalar reward \( r_{t+1} \) observing a new state \( s_{t+1} \). As our goal is to optimize the UE path guidance, the reward function should be defined according to the considered constraints such data rate measurements \( R_t \) and distance from destination \( \delta_t \) as

\[
(r(t + 1) = \log_{10}(\frac{R_t}{\delta_t})).
\]

In this section, we introduce the Neural Networks (NN) architecture, used as the DQN-Agent at BS and learn the desired goal using the optimization function described in P1. At every iteration, the UAV environment computes the subsequent states and rewards (6) for DQN-Agent based on the UE target distance and data rate measurements (4).

\( \text{A. UAV Environment} \)

We implement a UAV custom environment (denoted as UAVEnv) using python framework and OpenAI gym interface [22]. UAVEnv consists of defined coverage area \( U \) of the BS using 2D discrete state space \( S \) (2), radio beam direction set \( \mathcal{B} \) (1), possible UE directions \( \mathcal{M} \) and discrete action space \( \mathcal{A} \). The BS and target location (UE\(_{d_t}\)), multi-path channel model \( h \) with UMa conditions, received signal (3) and data rate measurements (4) are also implemented under UAVEnv class interface. At each episode, UAVEnv resets to random source location UE\(_s\), computes new observations and rewards based on the received action until it reaches destination.

\( \text{B. DQN Architecture} \)

DQN is a value-based RL approach [21], learning an optimal approximated policy of states mapping to actions \( \pi(s) = a \) by parameterizing and estimating state-action value function \( Q(s, a; \theta) \) using Deep Neural Networks (DNN). We denote the primary DNN network weight matrix and target DNN network weight matrix as \( \theta \) and \( \theta^\ast \), respectively [21]. We consider a fully connected DNN for both the networks where \( \theta^\ast \) is updated with primary network parameters \( \theta \), after every \( K \) 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) by using the function \( f(x) = \max(0, x) \) while the output layer is composed of linear units, which are in one-one corresponding relationship with all available actions in \( \mathcal{A} \). We consider initialization of bias and weights of these layers with zeros and Kaiming normalization [23], respectively.

At a time instant \( t \), \( \alpha_t \) selects either a random action from \( \mathcal{A} \) or perform forward propagation of \( Q(s_t, a_t; \theta) \) following \( \epsilon \)-greedy policy [20]. A memory buffer of experiences \( D_t = \)
\{e_1, e_2, e_3, \ldots, e_t\}, e_t = (s_t, a_t, r_{t+1}, s_{t+1}) are collected where a mini batch of them \(U(D)\) are randomly sampled and sent into DQN [21]. During back propagation, a Mean Squared Error (MSE) loss function is computed between primary, target networks and \(\theta\) is updated using Stochastic Gradient Descent (SGD) [24] and Adam Optimizer [25] as

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

(7)

where \(\xi_{\text{Adam}}\) is the learning rate. \(\nabla \mathcal{L}^\text{DQN}(\theta_t)\) is the gradient of the DQN loss function, given as

\[\nabla \mathcal{L}^\text{DQN}(\theta_t) = E_{(s_t, a_t, r_{t+1}, s_{t+1})} \left[ (R_{t+1} + \gamma \max_a Q(s_{t+1}, a; \theta_t) - Q(s_t, a_t; \theta_t)) \nabla \theta Q(s_t, a_t; \theta_t) \right],\]  

(8)

where \(\theta_t\) is used to estimate future value of Q-function inside \(\mathcal{L}^\text{DQN}\). Complete steps followed by DQN for every episode of our path planning problem is shown in Algorithm 1.

**IV. SIMULATION RESULTS**

In this section, we evaluate the performance of the proposed DQN algorithm for our path planning problem. We compare our learning-based joint path planning and beam tracking method against learning-based path planning with heuristic exhaustive beam searching method [26]. We adopt standard radio channel modelling parameters for mmWave communications as well as grid environment parameters listed together in Table I. The hyper-parameters used for DQN learning algorithm are also listed in Table II.

As our state space dimension is smaller (\(U_{x}, U_{y}\)), for now we consider a three layer network in our implementation. However, with increase in action space \(A\) and state space \(S\) for more generalization, more layers in DNN’s have to be used. Thus, our DQN architecture is set with two 128 hidden layers and one 64 hidden layer ReLU units, respectively. Also, we assume a 2D uplink beamforming communication between UAV and BS by considering the altitude of UAV fixed. Fig. 3 plots the received episodic reward over training episodes, smoothened for every 200 samples (given as \(E\{r\} = 1/200 \sum_{t=0}^{200} 1/N_{\text{step}} \sum_{t=0}^{N_{\text{step}}} r_t\) i.e. one episode consists of duration \(N_{\text{step}}\)). It can be observed that DQN converge to a optimal value function \(Q^*(s, a)\) over 2500 training episodes. The increase in average reward over training episodes in all the observations demonstrate that DQN can successfully learn the optimal trajectory for UAV along with BS beam directions.

Red colored path along the grid in Fig. 4 shows the UAV trajectory of proposed DQN-agent based joint path planning and beam tracking method for a random test episode starting...
from source (UE$_1$=(-500, -500)) and reaching towards destination (UE$_2$=(450, 450)). Similarly, black colored path along the grid represents learnt shortest path planning UAV trajectory using a different RL BS agent under same simulation conditions. The BS for this scenario is located at (0, 0) as shown in Fig. 4. We consider a multi-path communication model scenario with 3 reflection points at rp$_1$=(0, 150), rp$_2$=(250, 50) and rp$_3$=(-200, -150). The DQN-agent used for the proposed path is trained over 2500 episodes. For the shortest path planning method, we retrained the same DQN-agent by computing rewards without rate conditions. A heuristic exhaustive beam search method is separately performed later at every location along the learnt path, computing the best possible data rate measurements.

We observe that DQN approach follows a different path compared to shortest path in reaching the same target location. Despite different optimization criteria, both the methods are able to reach the destination in an equal number of steps. Thus, the DQN approach can offer as a generic framework to learn the optimal trajectories for UAV under different constraint conditions. Also, the computed data rate measurements (in Gbps) are labelled in a bounding box (following color coding) at every location along the path, for both the methods. Rate measurements along the red path are learnt directly from the joint path-planning and beam tracking DQN approach while the measurements along black path are computed using heuristic exhaustive beam search method following multipath channel conditions. We observe that red labelled learnt data rate values are comparable to that of heuristic grey labelled values, justifying the beam learning under multi-path conditions. The learnt beam directions at each location along the path, prevents frequent beam scanning at BS, reducing its communication overhead to a great extent. This can be very energy efficient especially when the BS commands multiple UAVs at the same time. BS considers a prior computational overhead by learning optimized UAV trajectories within its coverage area, based on its constraints and prevent frequent beam sweeping overhead all the time.

However, the accuracy of the learnt data rate values still needs to be analyzed. This can be better understood by considering learnt and heuristic data rate values along red and blue line respectively, starting from same random source location as shown in Fig. 5. We observe that increase in cumulative rates along the blue curve is higher compared to that of cumulative rates in red plot. This indicates that the learnt rate values are comparable to heuristic approach but the

### Algorithm 1: DQN Based Path Planning

**Input:** The set of UAV x,y location coordinates and training iterations M

1. Initialize replay memory M to capacity C, the primary Q-network with parameters $\theta_1$, the target Q-network with parameters $\theta_2$
2. Sample random minibatch of transitions $D$
3. for Iteration ← 1 to M // for each episode
   4. Initialization of $s_1$ by executing a random action $a_0$
5. n, N = 0, Episode Limit
6. while True do
   7. if $p_a < \epsilon$ then
      8. select $a_t = \text{argmax}_{a \in A} Q(s_t, a, \theta)$
   9. else
      10. select $a_t = \text{random action from } A$
11. BS applies $a_t$ over the channel, receive signal for $(t+1)$th iteration during uplink communication
12. UE observes $S_{t+1}$ and calculate the reward using eqs. (4)-(6)
13. Store transition $e = (s_t, a_t, r_{t+1}, s_{t+1})$ in replay memory $D$
14. Sample random minibatch of transitions $U(D)$
15. Compute Loss and Perform gradient descent for $Q(s, a; \theta)$ using eqs. (7),(8)
16. Every K steps update the target network parameters $\theta_2 = \theta_1$
17. $n = n + 1$ // Increment episode time
18. if done or (n = N) then
   19. break // End episode
20. end while

---

### Table I

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmWave Channel</td>
<td>UMa</td>
</tr>
<tr>
<td>UMa-LoS Pathloss coefficients</td>
<td>$\alpha = 2.8, \beta = 11.4, \gamma = 2.3, \sigma = 4.1$</td>
</tr>
<tr>
<td>UMa-NLoS Pathloss coefficients</td>
<td>$\alpha = 3.3, \beta = 17.6, \gamma = 2.0, \sigma = 9.9$</td>
</tr>
<tr>
<td>mmWave freq</td>
<td>30 GHz</td>
</tr>
<tr>
<td>carrier spacing freq df</td>
<td>60 kHz</td>
</tr>
<tr>
<td>Num of subspace carriers NFFT</td>
<td>1200</td>
</tr>
<tr>
<td>antenna element spacing d</td>
<td>0.5</td>
</tr>
<tr>
<td>Transmit power $P_{tx}$</td>
<td>30 dBm</td>
</tr>
<tr>
<td>Transmit antenna elements $N_t$</td>
<td>8</td>
</tr>
<tr>
<td>Receiving antenna elements $N_r$</td>
<td>8</td>
</tr>
<tr>
<td>Noise Level $N_0$</td>
<td>-174 dBm</td>
</tr>
<tr>
<td>BS location</td>
<td>[0, 0, 0]</td>
</tr>
<tr>
<td>coverage xloc $U_{xloc}$</td>
<td>$[-500, 500, 50]$ m</td>
</tr>
<tr>
<td>coverage yloc $U_{yloc}$</td>
<td>$[-500, 500, 50]$ m</td>
</tr>
<tr>
<td>Cardinality of Beamset $M$</td>
<td>8</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate $\lambda_{Adam}$</td>
<td>$5 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\epsilon$-start</td>
<td>1.0</td>
</tr>
<tr>
<td>$\epsilon$-end</td>
<td>$0.01$</td>
</tr>
<tr>
<td>$\epsilon$-decay</td>
<td>0.9995</td>
</tr>
<tr>
<td>Soft-Update rate $\tau$</td>
<td>0.001</td>
</tr>
<tr>
<td>Target Q-Network Update Frequency $K$</td>
<td>10</td>
</tr>
<tr>
<td>Minibatch Size $U(D)$</td>
<td>64</td>
</tr>
<tr>
<td>Replay Memory Size $</td>
<td>D</td>
</tr>
<tr>
<td>Discount rate $\gamma$</td>
<td>0.999</td>
</tr>
</tbody>
</table>

---

### Figure

Figure 4: 3D multi-path geometry of multi-antenna setup.

Figure 5: Cumulative data rate over 2500 episodes for shortest path planning and joint path-planning and beam tracking DQN approach.
average learnt rate measurement for a location is slightly lower than its average exhaustive rate measurement along the similar joint path planning and beam tracking path. The difference in measurements is due to joint optimization of target distance and receiver beam direction at BS, during path planning. This justifies that the proposed approach jointly optimize the trajectory as well as achieve mmWave data rates comparable to that of exhaustive method, using beam tracking.

![Graph showing UAV Distance (m) w.r.t destination (450, 450) vs Cumsum Rate (Gbps)](image)

**Fig. 4.** Joint path planning and beam tracking trajectory as well as shortest path planning trajectory for a random DQN test episode of source (450, 450) and destination (−500, −500)

V. CONCLUSION AND FUTURE WORK

We developed a deep RL based uplink trajectory optimization framework for cellular-enabled UAVs within the coverage area of BS. We confined our UE constraints to target distance and data rate measurements while designing the generic framework using DQN algorithm. We initially shown the converge performance of DQN algorithm and then analyzed the results of joint path planning and beam tracking based trajectory against learnt shortest path planning towards destination. We also compare the performance of learnt rate measurements against the heuristic exhaustive beam search method. Thus, the proposed approach provides a generic framework for jointly optimizing multiple practical constraints to the path planning problem via BS. Having demonstrated some promising results, we would like to consider other constraints such as energy, fading and other complex terrestrial channel conditions in a multi-UAV scenario around BS, as extension to this research in future.

ACKNOWLEDGEMENT

The research leading to these results has received funding from European-Union project Primo-5G, Academy of Finland projects 6Genesis Flagship (Grant No. 318927), IIoT

![Graph showing Cumulative exhaustive and learnt rates along joint path planning and beam tracking trajectory across UAV distance with respect to its destination location](image)

**Fig. 5.** Cumulative exhaustive and learnt rates along joint path planning and beam tracking trajectory across UAV distance with respect to its destination location

CONnectivity for mechanICAL systems (ICONICAL) and Positioning-aided Reliably-connected Industrial Systems with Mobile mmWave Access (PRISMA).

REFERENCES


