Intelligent Handover Decision Scheme Using Double Deep Reinforcement Learning

Michael S.Mollel*†, Attai Ibrahim Abubakar† , Metin Ozturk†, Shubi Kaijage*, Michael Kisangiri*,
Ahmed Zoha†, Muhammad Ali Imran†, and Qammer H.Abbasi†
*The Nelson Mandela African Institution of Science and Technology (NM-AIST)
{michaelm, shubi.kaijage, kisangiri.michael}@nm-aist.ac.tz
†James Watt School of Engineering, University of Glasgow
{a.abubakar.1, m.ozturk.1}@research.gla.ac.uk,{ahmed.zoha,muhammad.imran,qammer.abbasi}@glasgow.ac.uk

***Abstract*—Handovers (HOs) have been envisioned to be more challenging in 5G networks due to the inclusion of millimetre wave (mm-wave) frequencies, resulting in more intense base station (BS) deployments. This, by its turn, increases the number of HOs taken due to smaller footprints of mm-wave BSs thereby making HO management a more crucial task as reduced quality of service (QoS) and quality of experience (QoE) along with higher signalling overhead are more likely with the growing number of HOs. In this paper, we propose an offline scheme based on double deep reinforcement learning (DDRL) to minimize the frequency of HOs in mm-wave networks, which subsequently mitigates the adverse QoS. Due to continuous and substantial state spaces arising from the inherent characteristics of the considered 5G environment, DDRL is preferred over conventional Q-learning algorithm. Furthermore, in order to alleviate the negative impacts of online learning policies in terms of computational costs, an offline learning framework is adopted in this study, a known trajectory is considered in a simulation environment while ray-tracing is used to estimate channel characteristics. The number of HO occurrence during the trajectory and the system throughput are taken as performance metrics. The results obtained reveal that the proposed method largely outperform conventional and other artificial intelligence (AI)-based models.**

***Index Terms*—double deep reinforcement learning; handover management; millimeter-wave Communication.**

I. Introduction

The enormous demand for high-speed communication for mobile devices require high data rate broadband connections. Ultra-reliable low latency communication (URLLC) and enhanced mobile broadband (eMBB) scenarios defined in the fifth generation (5G) New Radio (NR) require high reliability for mission-critical communication and high data rate across a wide coverage area [1]–[3]. On the other hand, emerging technologies, such as tactile internet, remote surgery, and augmented reality, create unprecedented challenges that need to be adequately addressed as well as being bandwidth-hungry.

Currently, the global spectrum bandwidth allocation for all cellular technologies does not exceed 780 MHz [4], which is insufficient for the future generation of mobile networks due to the enormous services that are placed on them. Furthermore recent advancement in technology has made the practical usage of frequency bands above 6 GHz (mm-wave frequencies) in the next generation of mobile networks, also known as 5G systems. The mm-wave band offers enormous potentials services to 5G networks owing to the substantial amount of available bandwidth it contains.

Albeit having a considerable amount of gain in terms of bandwidth, the mm-wave band has severe limitations when it comes to the applicability, given that the mm-wave band has distinctive characteristics compared to sub 6 GHz band. As such, the attenuation and reflection of signals, for example, are more severe compared to that of sub 6 GHz band which subsequently results in increased non-line of sight (NLOS) regions especially for outdoor environment [5]. In addition, even though both line-of-sight (LOS) and NLOS links experience a high level of attenuation, the signal in the NLOS link is weaker than that of LOS link by a margin of 10 dB+ [6]. In this regard, rain attenuation, fixed and random obstacles, and high propagation losses limit the coverage range of mm-wave frequency [7].

Overcoming the challenges mentioned above requires a joint deployment of a large number of small cells (SCs) operating at the mm-wave frequency to increase the high date rate coverage alongside macro cells (MCs) operating at sub 6 GHz frequency to provide broad coverage. This kind of network deployment is often referred to as homogeneous ultra-dense network (UDN) [8], [9], when UDN involves only mm-wave SCs and heterogeneous UDN when there is the presence of more than one-tier. Although the ultra-dense deployment of SCs will enhance the data rate, it would also result in more frequent switching of user connection from one base station (BS) to another. The process of switching user connection from one BS to another or associating and re-associating UEs is known as handover (HO) [10]. It has a significant effect on highly mobile UEs because the period spent within the coverage area of a SC (dwell time) reduces with increasing user velocity.

One of the most common metrics for choosing the best serving BS for user HO is the received signal strength (RSS). In this case, the UE switches connection from its serving BS to a target/candidate BS that provides higher RSS than the current serving BS along the user trajectory. However, the escalating level of heterogeneity in mobile communication networks requires that other metrics such as user throughput, service delay, and load balancing should also be considered in determining the best serving BS.

Furthermore, since the user would always have to switch

BS connection in the course of movement, the frequency of HO increases with an increasing number of SCs. Hence, the HO cost is usually considered alongside the aforementioned metrics [11]. The process of HO usually involves the exchange of signalling between the UE, serving BS, network controller, and the target BS. This, in turn, causes interruptions in data transmission, thereby resulting in the reduction of the UE throughput.

The rate at which these interruptions in data transmission occur is proportional to the user velocity as well as the BS density, particularly for mm-wave SCs. A recent study in [12] reveals that the interval between successful HOs could reach as low as 0.75 sec for a practical mm-wave network deployment scenario. Moreover, the authors in [13] have shown that more than 60% of HOs are unnecessary. Hence, in UDN with high mobility users, there is a need to reduce the number of HO interruptions in order to minimise the HO cost, maintain an excellent level of user throughput, and reduce the service delays associated with HO process.

Various HO optimization techniques have been proposed in the literature in order to tackle the challenges as mentioned earlier in UDN. The most popular techniques include the Markov decision process (MDP) and stochastic geometry. The authors in [11], [14], [15] applied stochastic geometry techniques to derive an analytical model for HO optimization. MDP based models for HO optimization in mm-wave networks is also proposed in [16]–[19]. However, it would become very difficult and computationally demanding to derive accurate analytical models using MDP and stochastic geometry techniques when network complexity increases and dimensions becomes very large, also these models often involves certain assumptions which might not be obtainable in real wireless networks [17], [20]. Therefore, these techniques might be impractical and inefficient for the case of UDN.

In this paper, we propose an intelligent HO decision algorithm using DDRL to select the optimal BS that maximise the user-BS connection duration in order to reduce HO cost while guaranteeing user QoS. To make such a decision definitely involves having information regarding user trajectory and network topology. A trajectory-aware HO optimization approach is developed such that instead of using exact user location i.e. Geo-coordinates of the user's location (since it is difficult to obtain), we correlate the user location to SNR values received from all BSs at any particular point. The mapping of the exact location to SNR takes into consideration various kinds of obstacles in a typical network environment, such as building, trees, vehicles and human-beings using wireless Insite software^{TM} (WI). In particular, we develop an offline learning framework; such that, first, the environment is simulated for data collection purposes. Then, the DDRL algorithm is trained with the collected data, followed by testing the developed model with a new data set generated by altering some of the points along the trajectory to test the robustness and generalisation ability of the model.

The rest of the paper is organized as follows. A review of the state-of-the-art on ML-based HO management in UDN is presented in Section II. The system model is introduced and discussed in Section III. Section IV details HO event and cell selection criteria while in Section V we present the proposed DDRL based HO optimization framework. The performance of the proposed framework is evaluated in Section VI while Section VII concludes the paper.

II. Related Works

Recently, Machine learning (ML)-based approaches have proven to be one of the most prominent tools used for solving HO optimisation problems in UDN [20]. ML techniques rely on data generated from cellular networks. They have an advantage over analytical optimisation techniques because they can learn hidden patterns and structures in a system that is difficult to derive analytically [20]. Getting the data is easy since cellular networks always generate a massive amount of data such as channel matrix, SNR and other information stored in Channel State Information CSI. Thus, ML can leverage these used and unused data to enhance network performance, including HO optimisation problem. Also, the ML-based approach can learn and predict network parameters including sojourn time, and BS traffic in advance by utilising historical data, thereby enabling proactive optimisation of network performances and user QoS [21]–[23]. Furthermore, HO optimisation in mm-wave UDN deployment using ML techniques is still yet to be fully exploited, and in this section, we present most of the ML-based state-of-the-art HO management schemes that are used in UDN.

A. Machine Learning based Handover Management in Microwave Networks

The authors in [24] proposed a Reinforcement Learning (RL) framework for HO management in heterogeneous networks (HetNets). In particular, the traffic load and optimal cell range expansion of both the macro and small cells are jointly learnt, and users are scheduled according to their velocities and previous HO rates in order to enhance user throughput. A two-tier ML-based model for HO management in vehicular networks was developed in [25] using the Recurrent Neural Networks (RNN) to predict the RSS required to initiate HO after which a stochastic Markov model was used for BS selection. The authors in [26] investigated the use of Artificial Neural Network (ANN) for vertical HO decision in HetNets using data rate, user velocity, and RSSI parameters. In [27], a data-driven mechanism was proposed using multi-layer perceptron to optimise HO parameters, such as time to trigger (TTT) and HO margin, by leveraging data generated from the network. Similarly, an ML-based scheme for selecting best serving BS while ensuring seamless connectivity was developed in [28]. An ANN model which selects the best access network based on specific criteria, such as application requirements, user demands, and UE capabilities in order to enhance the QoS of both voice and data services was introduced in [29] by jointly considering the effects of obstacles and previous Quality of Experience (QoE) of the users. A two-layer framework based on asynchronous multi-agent Deep Reinforcement Learning (DRL) for optimal HO control was presented in [30]. The authors clustered the UEs based on their mobility pattern, followed by the application of

DRL to each cluster to minimise the number of HOs while maintaining the system throughput.

B. Machine Learning based Handover Management in mm-wave Networks

The authors in [31] proposed an RL framework, which they call SMART, to minimise the number of HOs by taking into account the mm-wave channel conditions and user QoS. As such, two different RL based BS selection algorithms were proposed for single user and multiple user HO triggering in sparse and dense user distribution scenarios, respectively. The limitation of SMART is that it disregards the user's location and pays no attention to the repercussion caused by neglecting such vital information. Secondly, the model learns online, that means there is a computation cost incurred during the training phase. In addition, we select this model as a benchmark to validate the performance of the proposed model in the results section, since SMART and our proposed models share a similar system model architecture. Another reason is that SMART's authors have made few assumptions similar to this paper regarding the environment setup compared to other studies.

An ANN-based proactive HO mechanism was developed in [32], where historical data comprising user connection beams was exploited to predict the future occurrence of mm-wave link failure due to obstruction in order proactively trigger HO. A major advantage of this model and any other model which involve time series problem is that it can tell in advance when HO will occur and to which target BS to connect. However, these models do not guarantee that the target BS is the optimal BS to connect to, and the reason is the same for any supervised learning algorithm since the input features need labels in the training phase. Meanwhile inaccurate labels for the input features leads to inaccurate model since the historical data of UE does not provide any information regarding BS optimality, except the mapping of input features to specific BSs. Hence, the limitation of the model is that it can only solve proactive HO related problems based on the UE historical data but it does not consider issues regarding how to optimise the labels of the training set to reduce HO. In [33], the authors use a discrete state vector containing discrete space positions, space velocity, and BS index for HO prediction while considering pedestrian as obstacles. The proposed solution is, however, limited to a particular kind of blockage, and it does not consider other types of obstacles, including baggage, vegetation, buildings, and cars. Moreover, the problem of using Q-table limits the application of the model due to lack of scalability since Q-learning only works in the environments with discrete and finite state and action space.

In order to adequately capture the different type of obstacles in mm-wave links and to proactively trigger HO, the authors in [34] proposed a DRL assisted proactive HO policy using camera images to optimise HO decision. The proposed technique does not require any information about the location of the obstacles but instead maps the camera images as the input feature to the action value. The proposed method, nevertheless, depends on the resolution of the camera to adequately capture the images and may not be useful if camera vision is impaired.

Unlike the techniques discussed earlier, in this paper, we present a user trajectory-based HO decision optimisation strategy. The idea is to extend Q-learning in a more productive way by applying a function approximator, which for our case is Deep Neural Networks to learn the value function, taking states as inputs instead of storing the full state-action table. SNR, which is part of the state and usually available in CSI, is obtained directly from Wireless Insite software^{TM} (WI). Furthermore, we consider SNR as a crucial parameter since it represents the location of UE at a particular point and takes into consideration various environmental parameter such as obstacles and other terrain information.

C. Contributions

In this paper, our contributions are summarized as follows.

- We exploit the capabilities of the Wireless Insite software^{TM} (WI) to model a typical mm-wave UDN scenario. By mapping exact user location — which is geo-coordinate of user's location since it is not always available/practical in reality — to SNR values along user trajectory—while considering the effect of obstacles that cannot be captured by localisation devices.
- We propose an offline intelligent HO learning framework in mm-wave networks by considering the HO cost (definition will be clear shortly) associated during the HO process.
- Finally, we present the optimal BS selection policy that maximises UE-BS connection time from the offline DDRL algorithms. The policy also considers the trade-off between instantaneous received SNR and HO cost to reduce unnecessary HOs due to the frequent short-term LOS blockage by obstacles.

III. System Model

A. Network Model

Consider a cellular network environment, as shown in Fig. 1, comprising a large number of mm-wave SCs and UEs where the network considers the presence of both LOS signal, blockages, and building reflectors. The motive is to depict a real urban environment with entirely distinct obstacles. Each mm-wave SC is equipped with X antennas, and all SCs are assumed to be connected to a central controller (CC). Multiple overlapping mm-wave SCs are randomly distributed in the network to provide high throughput by LOS links. Additionally, macro BSs (MBSs) exist in the network and is for two reasons; first, to ensure reliable communication whenever no LOS link is available from the mm-wave SCs and secondly, to facilitate the transmission of control signals to the CC that acts as decision node in the network. Each mobile UE is assumed to be equipped with a single antenna.

The employed channel, beam-forming, and SINR models are elaborated in the following paragraphs.

Fig. 1: The system model of mm-wave UDN.

B. Channel Model

In the channel model, the ray-tracing model is used in this study. The ray-tracing model is based on the superposition principle whereby the sum of all reflected waves generated by a transmitter and LOS waves between the transmitter and the receiver are aggregated. More specifically, a geometric wide band mm-wave channel model [4] with N clusters is adopted, where each cluster $n \in N$ is assumed to produce a single ray with a finite time delay $\tau_n \in \mathbb{R}$ and angle of arrival (AoA) for elevation/azimuth ϕ_n, θ_n . We assume each user has a single antenna and the path loss between UE and m^{th} BS is λ_m. The time delay channel between the m^{th} BS and the UE, $s_{d,m}$, can be expressed as:

$$s_{d,m} = \sqrt{\frac{X}{\lambda_m}} \sum_{n=1}^{N} \alpha_n p\left(dT_{\mathrm{p}} - \tau_n\right) a_m\left(\theta_n, \phi_n\right), \tag{1}$$

where X symbolizes the number of antenna in the BS, $a_m\left(\theta_n, \phi_n\right)$ represents the array response vector of the m^{th} BS at AoA $\left(\theta_n, \phi_n\right)$, and $p\left(dT_{\mathrm{r}} - \tau_n\right)$ denotes the pulse shaping function of the spacing signalling, T_{p} obtained at τ seconds [35]. From the time delay channel in (1), the subcarrier, k, the frequency-domain channel model, $s_{k,n}$ can be expressed as:

$$\mathbf{s}_{k,m} = \sum_{d=0}^{D-1} \boldsymbol{s}_{d,m} e^{-j\frac{2\pi k}{K}d}. \tag{2}$$

It is assumed that the block-fading channel model, $\{s_{k,m}\}_{k=1}^{K}$ is constant within the coherence time of the channel, T_c, [4] which is dependent on the velocity of the user as well as the multi-path components of the channel.

C. Beamforming

We assume that the BSs are equipped with a highly directional antennas with the sectorized gain pattern. Due to the high-frequency mm-wave BS has, it is easier to exploit beamforming techniques. The antenna gain from the main lobe is Ω while that of the side-lobes is ω. The directional gain of the antenna is the function of θ as well as the steering angle and is given as:

$$G(\theta) = \begin{cases} \Omega & \text{if } \mid \theta \mid \leq \theta_b \\ \omega & \text{otherwise,} \end{cases} \tag{3}$$

with θ_b denoting the beam-width of the antenna and θ is the angle between UE and mm-wave BS. Highly sophisticated beam-tracking is deployed in WI$^{\text{TM}}$ to ensure the definitive association between UE and mm-wave BS are successful. Therefore, the UE is always in the main lobe with main lobe gain connected in a given antenna, and UE experiences no interference or any signal from other antennas.

D. SINR model

One of the key metrics for the wireless communication channel is the SINR, which can be expressed as:

$$SINR = \frac{S}{I + \sigma_u}, \tag{4}$$

where S represents the received signal power, σ_u denotes the noise component, and I denotes the interference power from all surrounding BSs except the serving BS. Equation 4 can be further expressed as

$$SINR = \frac{P_s \Omega_s L_s}{\sigma_u + \sum_{i=1}^{M} P_{(i=t \,\cap\, t \neq s)} \omega_i L_i}, \tag{5}$$

where P_t and P_s represent the power transmitted by surrounding and serving BSs respectively, ω_i=ω_t and Ω_s depict antenna gain of surrounding and serving BSs, and L_i=L_t and L_s is the path loss gain of the surrounding and serving BSs respectively. Because mm-wave antennas form directional beams (Eqn. 3), the contribution of inter-cell interference can be assumed to be negligible. Hence, SNR is considered in this paper instead of SINR.

IV. Handover Events and BS Selection

In this section, we present the HO triggering criteria for the proposed framework as well as the policy for selecting the target BS. The HO is initiated from criteria defined in 3GPP [36] and [37]. The aim is to avoid a short-sighted decision by selecting a BS whose contemporary has higher SNR, but the UE-BS connection is lost in few seconds after HO.

A. Condition for Initiating Handover

According to 3GPP [36], six events are defined for initiating conventional HO, where events A2 and A3 are for intra radio access technology (intra-RAT) HO. In contrast, inter-RAT HO is described in event B2. These are the criteria to initiate and terminate HO.

The condition for entering event A2 is when the serving BS becomes worse than the stipulated threshold in terms of RSRP or SNR value and the opposite for the leaving state. The authors in [37] use the SINR value as the threshold value and analyze its effect on HO in mm-wave networks. Their study concluded that different use cases (service-aware) could have different HO rates in the same trajectory. In this study, we

adopt the SNR as the triggering criteria, and we define event A2 as:

$$\begin{cases} \gamma_s < \gamma_{th} & \text{initiate HO} \\ \gamma_s \geq \gamma_{th} & \text{otherwise,} \end{cases} \tag{6}$$

where γ_s is the SNR from the serving BS and γ_{th} is minimum SNR required by UE to maintain connection based on the service type. We remove the hysteresis margin since the parameter itself needs optimization. However, removing the hysteresis margin might lead to an increase in the number of ping pong and other unnecessary HOs, but the proposed model ensures HOs are reduced without hysteresis margin. The details of how the proposed model works will be explained in a later section. Therefore, if the condition of HO is met, the HO process commences and UE needs to select a potential target BS.

The condition for entering event A3 is defined when the neighbour BS becomes offset better than the serving BS. Regardless of how much power the UE receives from the serving BS, the event looks for the offset value between serving and neighbour BS, and if the condition is met, then the HO process is initiated. In the proposed solution, we neglect this event as it can sometimes lead to unnecessary HO. Lastly, the condition for event B2 is the same as that of event A2 except that it involves inter-RAT HO while A2 involves intra-RAT HO.

B. Handover Cost

In LTE there is only hard HO [38] in use, and we assume the same applies to 5G. The author in [38] explain the complexity and the waste of resources in soft and softer HO. Therefore, in this study, hard HO is the focus unless stated otherwise. Because hard HO is considered in this study, then the optimal BS selection during HO becomes more critical since the connection is broken before a new connection is established. After the HO process is initialized in Eqn. 6, it takes time to complete the HO process for UE to switch connection from serving BS to target BS. During the HO process, nothing is transmitted between UE and either the serving BS or target BS. The time spent for completing a successful event, where the UE is switched from the serving BS to the target BS without data transmission, is known as HO delay time t_d, and the accumulation of t_d has a significant effect on the average throughput of the UE. The cumulative t_d and total number of HO in a given trajectory is known as HO cost (β_c). HO cost is the function of the number of HOs and HO delay time, and is expresses as [14]

$$\beta_c = \min(\mathcal{H}_l \times \mathrm{v} \times t_d, 1), \tag{7}$$

where $\mathcal{H}_l$ is the total number of HOs per unit length (m), v is the velocity in (ms^{-1}), and t_d is the HO time delay in (sec).

The factor β_c is evaluated as the total time wasted without useful data transmission due to HO operations such as signalling and radio link switching between serving BS and target BS. The network performance becomes zero if $(\mathcal{H}_l \times \mathrm{v} \times t_d) \geq 1$ because the UE spend the entire time transmitting HO signalling. The importance of β_c is observed in the average

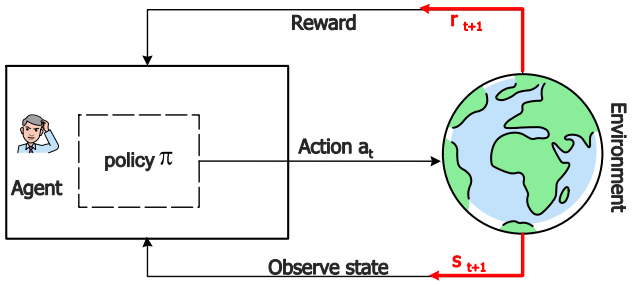


Fig. 2: Overview of generic RL algorithm

throughput equation derived as follows from the Shannon capacity formula:

$$\mathcal{T} = B \times \log_2(1+\gamma) \times (1-\beta_c), \tag{8}$$

where B is the overall bandwidth, γ is the average experienced SNR, and $\mathcal{T}$ is the system throughput. Hence, to obtain a high overall system throughput in a given trajectory, the factor β_c and γ play a vital role if B remains constant. Therefore, our objective is to maximize system throughput and avoid taking unnecessary HO by intelligently selecting BSs with a longer duration of unobstructed LOS link. At the same time, to reduce redundant HOs, we occasionally sacrifice a connection to a BS with the highest SNR that would potentially result in HO after few seconds. Yet, the average SNR should be above the threshold value to maintain the QoS of the user.

C. Trajectory and Service Aware Handover

Equation 8 has two parameters that can be varied as a trade-off to achieve the maximum average throughput. We introduce a service-aware concept whereby the UE maintains a connection to the serving BS if the γ_s experienced is above or equal to the threshold γ_{th}. This is to guarantee the QoS of UE, and we call it a service-aware strategy which also agrees with event A2. That means, in HO event, the UE doesn't need to choose the BS that has the highest γ_s instead the UE can select any BS that has γ_s above γ_{th}.

In addition, we introduce the trajectory-aware strategy, meaning the complete path is known to the UE. By knowing the trajectory of UE and service type, UE can carefully and intelligently select the BS that guarantees long connection duration with the γ_s above γ_{th}, and this far-sighted view helps to minimize the possibility of incurring multiple HOs by selecting the optimal BSs. Nevertheless, the combined strategy comes with the cost of sometimes sacrificing connection to the BS, providing maximum SNR during HO (ignoring current instantaneous SNR), and connecting to BS that can maintain extended connectivity along the UE trajectory.

V. Proposed DDRL Framework

Our problem mainly focuses on which BS during HO event to connect to in order to maintain longer connectivity along the user trajectory. The problem of selecting target BS is modelled as a multi-armed bandit problem in RL domain. Whenever the HO event is initiated, potentially, there is more

than one possible optimal BS that UE can choose to establish a connection, and reduce the chances of the UE entering into HO event again. In this section, we present the intelligent BS selection policy, which also facilitates proactive HO using RL. We start by introducing the basics of RL in subsection V-A1 followed by DDRL subsection V-B and lastly associate RL algorithm to our problem.

A. Reinforcement Learning

In this subsection, we briefly present an overview of RL. First, we elaborate on the RL framework and provide the vital components of RL.

1) RL Framework: The authors in [39] explain the relationship between agent, action and the environment, and further illustrate clearly and concisely how an agent learns the best policy through multiple interactions with the environment, as shown in the Fig. 2. Here, we first define the main elements of RL. At time t, the agent observes the state of the environment, $s_t \in S$, where S is the set of possible states. After observing state s_t, agent takes an action, $a_t \in A(s_t)$ where $A(s_t)$ is the set of possible actions at state s_t. Subsequent to the action a_t selected from state s_t, agent receive the immediate reward r_{t+1} from state-action pair (s_t, a_t). The selected action in state s_t moves agent to state s_{t+1} at time $t+1$. It is important for environment to have state transition dynamics such that P($s_{t+1}|s_t, a_t$) exists, also time (t) is the arbitrary successive stages of decision making and acting, which represent the situation and not interval of real time in second.

RL has two phases: the learning phase and the execution phase. The strategy for action selection by an agent in a given state is known as policy π. As shown in Fig. 2, the agent learns the optimal policy π_*; by first observing the current state s_t, and then taking action a_t following the current policy π, the environment state changes from s_t to s_{t+1}, and the agent gets an immediate reward r(s_t, a_t). The agent repeatedly updates policy π until it reaches optimal policy π_*. The agent's goal is to achieve the optimal policy π_* by maximizing the cumulative reward, and mathematically the cumulative reward includes the sum of immediate reward and future reward [39] and is given as:

$$\mathbf{G_t} \triangleq R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k}, \qquad (9)$$

where R_t is the immediate reward per episode and k is the total number of the episodes the agent navigate to acquire the full understanding of the environment. $\gamma \in (0, 1]$ is a discount factor for weighting future rewards, and the purpose of γ is to make the sum of reward to be finite. The episode is the complete sequence of states visited by an agent from the initial state to the terminal state. With sufficient experience through episodes iterations, the agent can learn an optimal decision policy π_* that would maximize the long-term accumulated reward.

Our problem for selecting BS during HO that can minimize the number of HO and maximizing average throughput falls into model-free learning where agents discover its environment by trail-and error [40]. And one of the most common RL algorithms in the model-free method is Q-learning which is an off-policy algorithm. In Q-learning, action-value function $Q(s,\ a)$ is defined as the long-term reward and is given as:

$$\begin{aligned} Q_\pi(s,a) &\triangleq \mathbf{E}_\pi \left[G_t | s_t = s, a_t = a\right] \\ &\triangleq \mathbf{E} \left[R_t + \gamma G_{t+1} | s_t = s, a_t = a\right] \\ &\triangleq \sum_{r,s} \left[r + \gamma \sum_{a'} G_{t+1} | s_t = s, a_t = a \right], \end{aligned} \qquad (10)$$

where optimal action-value function, $Q(s,a) \triangleq \max_\pi Q^\pi(s,a)$, obeys the Bellman optimally equation as shown in equation 11.

$$Q^*(s,a) = \mathbf{E}_{s'} \left[r_{t+1} + \gamma \max_a Q^*(s', a') | s, a \right] \qquad (11)$$

The Q-learning algorithm updates the corresponding element in Q-Table episodically according to the equation,

$$Q(s,a) = (1-\alpha) Q(s,a) + \alpha (R_{t+1} + \gamma \max_a Q_\pi(, A)) \quad (12)$$

where α is the learning rate, and s' is the next state after the agent follows policy π in state s.

B. DDQL-based optimal BS selection

In this section, we present the challenges of applying Q-learning in our problem, and then we explain in detail how the double deep Q-learning helps to solve these challenges. It is known that if the environment has many states such that the number of states is of the order of hundred's of states and hundred's of actions per state, this would result in a Q-table with ten thousand cells, hence the learning process quickly get out of control. Infinite number of states and actions create two problems. The first problem is that the amount of memory required to store and update the state action table increases as the number of states increases, and secondly, the time spent to explore each state in order to populate the Q-table accurately [39] becomes significantly high. Another limitation of Q-learning is that it only works in the environments with discrete and finite state and action spaces, meaning Q-learning is unable to estimate Q-value for any unlearned state.

The authors in [41] showed that RL could be implemented in a different way to operate efficiently with a large number of actions and continuous states. The new architecture leverages Artificial Neural Networks (ANN) to store the states and state-action value. The state is given as the input, and the state-action value, which is the Q-value of all possible actions is generated as the output for a given observed state. In this paper, we consider DDRL over DRL for the two main reasons. Firstly, the authors in [42] show that DRL suffers from a substantial over-estimations problem in some games. Secondly, after running both algorithms based on our environment design, we came to the same conclusion as [42]. We include the comparison result between DDRL and DRL in the results sections. To generalize the problem in this paper, Double Deep Q-networks (DDQN), which is DDRL, is implemented. The architecture works by exploiting the advantages of all previous models of Q-learning and Deep Q-learning.

The DDRL is the RL algorithm that uses and maintains two separate Deep Q-Networks (DQN). DQN is the multi-layer

Algorithm 1: Double Deep Q -learning Algorithm

Input : $\mathbb{D}$ - empty experience replay buffer; θ - initialize the training network for action - value function Q(s,a;θ); θ^- - initialize the target network Q(s,a; θ^-)

Input : N_b - training batch size; N_f - target network replacement frequency

for *episode = 1 to M* **do**
- Initialize the start state to S1;
- Update θ from parameters θ^+;
- **for** *i = 1 to end of trajectory* **do**
  - Agent observes the state and Service type (γ_{th});
  - **if** $\gamma_s \geq \gamma_{th}$ **then**
    - Action : $\leftarrow$ Index of serving BS;
  - **else**
    - Action : With probability ϵ take a random action $\mathbf{a}_i$ or else $a_i \leftarrow \arg\max_a Q(s, a; \theta)$;
  - **end**
  - Execute the action in the environment and observe the reward r_i and the next state (s_i'). Store (s_i, a_i, r_i, s_i') in $\mathbb{D}$.
  - Sample random mini-batch N_b from $\mathbb{D}$.
  - Construct target value , one for each of the N_r tuples:
  - Define $a^{max}(s', \theta) = \arg\max_{a'} Q(s', a'; \theta)$
  - set $y_j =$ $\begin{cases} r, & \text{if } s' \text{ is terminal} \\ r + \gamma Q\left(s', a^{\max}\left(s'; \theta\right); \theta^-\right), & \text{otherwise} \end{cases}$ Performs a gradient descent step on $\|y_j - Q(s, a; \theta)\|^2$
  - Replace target parameter $\theta^- \leftarrow \theta$ every N_f steps
- **end**

end

perceptron neural network that estimates output action values $Q(s, \cdot; \theta)$ for a given input state s, where θ are parameters of the networks. According to [42], the two separate networks for DDRL are target network and online network. The target network with parameter θ^- is the same as the online network except that its parameters are updated from the online network at every τ steps, such that θ_t^- = θ_t, and kept fixed on all other steps. DDRL reduces overestimation by decomposing the max operation in the target network into action selection and action evaluation. Therefore, the greedy policy is evaluated according to the online network, and values are estimated in the target network. The pseudo-code of the DDRL algorithms and how it works in relation to our proposed solution is given in Algorithm. 1, where vital elements in the Algorithm. 1 are explained as follow:

1) Action: The action is defined as which BS to connect to if A2 event occurs. We define action in action space ($a \in A(s)$) as the scalar representation of the serving BS index at state s. The set $A(s)$ comprises all BS in the environment.

2) State Vector: Traditionally, mobility management and other BS association strategy usually consider the location of UE to associate it to the serving BS. This study, however, considers the combination of the SNR received by UE from all surrounding BSs to represent the location of interest instead of the exact location of UE (i.e. geo-coordinates of UE's location). Getting the exact location of UE is impractical in reality; hence, we consider γ from all BSs along the UE trajectory as the representative of a point of interest instead of geo-coordinates.

Therefore, we correlate the current position of UE to the situation information from all surrounding BSs. We also assume the constant SNR values are realised from all BSs at a particular point based on the assumption that average SNR is used, and the combination of average SNR is uniquely sufficient to represent the geo-location coordinates at the specific position.

Hence, at the point, p with a total number of M BS, the state vector for UE is given as

$$s_p = \{\gamma_{\mathbf{1}}, \gamma_{\mathbf{2}}, \cdots, \gamma_{\mathrm{i}}, \cdots, \gamma_{\mathrm{M}}, \mathrm{BS_i}\}, \tag{13}$$

where $s_p \in S$ is the state at point p, γ_n is the SNR value from n^{th} BS, and BS_i is the index of the serving BS at the point p. The BS index is in one-hot encoded vector. One-hot encoding [43] is the vector representation of the integer variable into the binary value of all zero expect the index of the integer. For instance, if BS_i is assigned as three, and there are total of five BSs, then the equivalent one-hot encoding vector becomes $BS_i = [0, 0, 1, 0, 0]$.

3) Reward Design: The reward design is to motivate the agent to take actions that would maximize the cumulative reward in the long run, and since our objective is to achieve maximum system throughput ($\mathcal{T}$) for a given trajectory. Equation 8 shows that we can maximize $\mathcal{T}$ by minimizing β_c. To minimize β_c, Eqn. (7) shows for a given velocity (v) and HO time delay (t_d), the parameter $\mathcal{H}_l$ should be minimal as possible. The parameter, $\mathcal{H}_l$ can be controlled by implementing the HO skipping policy. Technically the agent initiates indirect TTT without setting a constant value, and this should be done intelligently to ensure UE achieve maximum throughput regardless of skipping some of necessary HO. This method has been used for micro and macro BS in the generation before 4G, and the TTT parameter was manually determined. Additionally, to minimise the value of β_c while maximising $\mathcal{T}$, during HO, the agent can select BS that has few numbers of event A2 in the future, known as far-sighted HO decision, provided the constraint $\gamma_s \geq \gamma_{\text{th}}$ is met.

In reward design, we avoid delayed rewards due to the problem of credit assignment [30], [39]. Therefore, we introduce the immediate reward function as the instantaneous throughput and evaluate the immediate effect of the action taken to achieve the agent's goal. Such a reward is given as follows:

$$r(s_{p+1}, a, s_p) = \begin{cases} B \times \mathcal{R} \times (1 - \beta_{\mathrm{c}}), & \text{if HO occurs} \\ B \times \mathcal{R}, & \text{otherwise,} \end{cases}$$

where B is the maximum bandwidth allocated, $\mathbb{R} = log_2(1 + \gamma_s)$ is the spectral efficiency, γ_s is the average instantaneous SNR the UE experiences from serving BS at position p obtained from the simulated environment. For the proposed

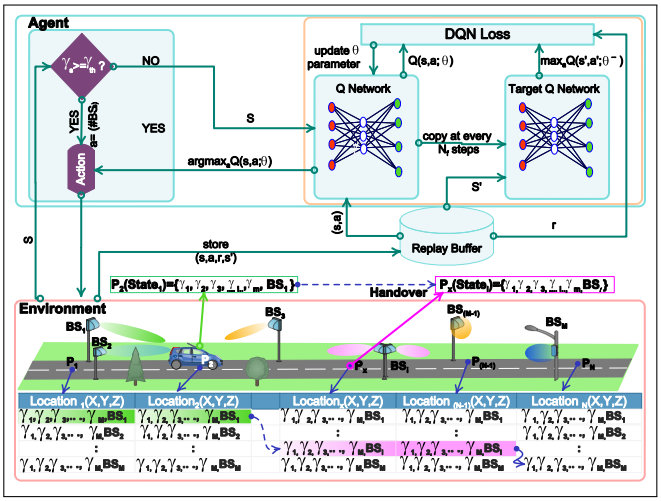


Fig. 3: The structure of the proposed DDRL - based on HO management scheme comprising environment, states, action, reward and double Q- networks

model to work accurately, information should be collected, and the agent uses accurate collected information for decision making.

4) Experience replay: The objective of experience replay is to overcome the instability of the learning algorithm. Experience replay is used to update the deep Q-network in such a way that both current and previous experiences are considered in the supervised learning based update process. This means that not only samples (s,a,r,s′) obtained from current online learning network but also old experience tuples (s,a,r,s′) are considered in the training process. Hence, experience replay store observed transitions for some time and sample uniformly from this memory bank to update the network. Using experience replay improves the performance of the algorithm [42].

5) Learning Algorithm: Fig.3 summarize how the agent interacts with the environment. For the mm-wave environment, there exists a substantial number of BSs meaning for a UE, the presence of obstacles principally initializes the event A2. Therefore, our proposed solution ensures that once event A2 is initiated, the UE switches to the BS that has a longer unobstructed time for its LOS connection or intelligently skip the HO. The proposed solution involves two phases: the learning phase and the execution phase.

In the learning phase, we use the offline learning whereby the agent gathers necessary information by simulating the UE trajectory in the environment, as shown in Fig. 3. Agent simulates the trajectory from starting point to the end point of the UE's path, and the agent performs the HO in a trial-error fashion. It is worth noting that we assume trajectory aware HO, therefore, the path that the UE takes is clearly known, and also during HO the agent can select the same BS that provides less than γ_{th} if skipping HO leads to maximum cumulative rewards. By doing trial and error, the agent can know two things in parallel: first, in HO event which BS is the best to connect UE and secondly, if HO is to happen, how long should UE remain connected to the BS with $\gamma_s \leq \gamma_{th}$ before HO to target BS. The latter information can be used to develop a proactive HO decision.

The summary of the agent learning process is presented in algorithms 1. The algorithm starts with the agent observing the type of service, which is in the state of the environment. Specifically, in the training phase, the UE takes action a according to one of the conditions stated: if condition $\gamma_s \geq \gamma_{th}$ is satisfied, UE continues with serving BS else UE uses ϵ-greedy policy with probability of less than exploration rate (ϵ) to randomly select BS else it uses policy $\arg\max_a Q(s,a;\theta)$ to choose BS. The UE receives the reward r and moves to next location $p+1$. In the new location, the UE generates the state s_{p+1} as the current state and same process starts all over and is steered by same rules which have been mentioned. The experience transition (s,a,r,s') is stored into the replay memory buffer $\mathbb{D}$ for experience replay. The process continues until it reaches the terminal state, and another iteration starts until the learning ends. The ϵ is set to decline from 1 to 0.1 after some learning steps.

In the execution phase, an agent takes action a according to the rules mentioned above. However, ϵ is set to 0.002 that means the agent use 0.2% of the time to explore, and the rest of the time uses policy with $\arg\max_a Q(s,a;\theta)$. During the evaluation phase of the model, we use the same environment, but we alter the point representing UE location in the trajectory to test model robustness and generalization behaviour. It worth noting there is no learning update in the execution phase. However, to keep the controller updated with the new data-set, UE continuously sends the observation states to the controller to update online policy. The synchronisation between UE and controller is beyond the scope of this work.

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed DDRL-based HO decision framework with other two existing HO policies, namely Rate Based HO (RBH) and SMART [31].

A. Simulation Setups

In both experiments, we consider the system model as described in Section III, where the environment, comprising an area of size $1000(\text{m}) \times 1000(\text{m})$ square area with the BSs randomly deployed, is considered. We then utilise random waypoint model [44] to generate user trajectory with the average velocity 8ms^{-1} for experiment 1 and for other experiments we vary velocity with constant threshold SNR (γ_{th}) of 20 dB. Using the random way-point model; specifically, we generated 10 trajectories and the assigned probability distribution of the four directions,which are North, South, East, and West, as follow; for trajectory 1 - 5 the distribution is $[0.25, 0.25, 025, 025]$ and $[0.6, 0.2, 0.1, 0.1]$ for other trajectories. The cumulative HO time delay during HO process t_d is set to $[0.5, 0.75, 1, 2]$ sec [14]. However, many studies mention 1 sec as the average HO time delay between mm-wave BS. In this study, we deliberately use more than one value of HO time delay to evaluate and quantify the effect of HO time delay associated during HO event. The observation time for UE mobility for all trajectories is 10000 secs, and the status of CSI is recorded at every 100 ms for training

TABLE I: Simulation Parameters

Parameter	**Value**
BS intensity	10 - 70 (BS/km^2)
mm-wave frequency	28 GHz
mm-wave bandwidth	1 GHz
BS transmit power	30 dBm
Thermal noise density	$-174\ dBm/Hz$
Delay without data transmission t_d	0.75, 1, 2, 3 sec

TABLE II: Parameters for Design for the developed DDRL Model

Parameter	**Value**
Hidden layers, Neuron size	4, 256 X 64 X 128 X64
Activation function hidden layers	relu
Activation function output layer	linear
Target network replacement frequencyN_f	40000
Initial exploration training	1
Final exploration training	0.1
Exploration in evaluation	0.0002
Optimizer	adam
Learning rate, α and Discount Factor, γ	0.001 , 0.96
Mini-batch size N_r	32
Replay memory size	100000

dataset and 10 ms with altered UE location along trajectory for evaluation dataset. All states fed to the DDRL are generated within this time period. Ultimately, we ignore the effect of interference with the reason described in Section III-C hence we consider SNR as the parameter of interest. We consider hyperparameters for DDRL from [34] since systematic grid search owing to the high computational cost. The complete parameters for the radio network environment as well as that of the DDRL are given in Table I and Table II, respectively.

B. Results

We start by analysing the performance of DDRL and comparing with DRL in the same setup, and the main reason is to show why we choose DDRL over DRL in this paper. The average cumulative reward against training episode for each training algorithm is shown in Fig. 4. From Fig. 4, it can be observed that the learning trend for DRL is quite similar to DDRL for some few episodes, which are from the start of training step to nearly 30×10^6 training steps. This means for both algorithms, an agent equally learns and improves its policy at least for some ranges of episodes. However, after certain episodes, the learning curve for DRL, shown in red, starts to drop while the learning curve for DDRL, shown in blue, ends up much higher and trend kept constant compared to DRL. The main reason for the decline of cumulative reward is because DRL mostly tends to over-estimate value estimates during the training phase. The problem, caused by DRL over-estimation, is thoroughly investigated in [45].

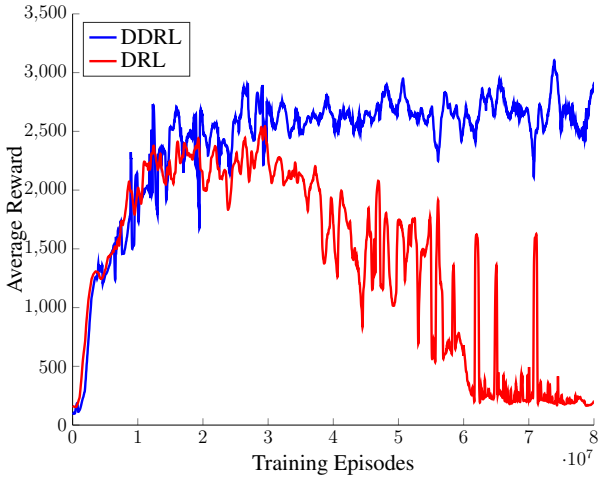
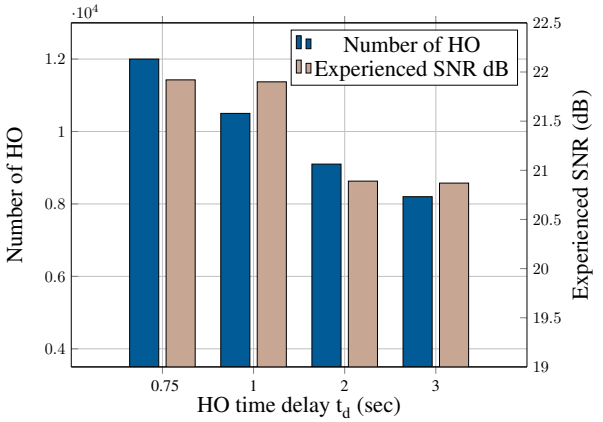


Fig. 4: Average score achieved by the agent as it is evaluated during training

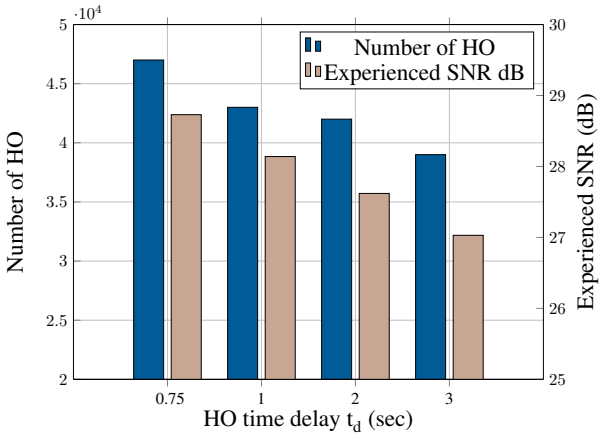
Therefore, Judging from both curves, it is without a doubt that DDRL produces more accurate value estimates, and better policies because of higher return and stable learning throughout the training process.

In the second experiment, we consider the velocity of each user to be 8 ms^{-1} for all trajectories. We analyse the relationship between the number of HOs and the experienced average SNR received by the user for the proposed model. Fig. 5 shows the relationship between the HO rate, the experienced average received SNR againts HO processing time delay t_d. From Fig. 5(a) and Fig. 5(b), It can be seen clearly that as t_d becomes higher, the number of HO decreases considerably, meanwhile the experienced average SNR decreases with small margin. This behavior illustrates how the agent sacrifices experienced SNR to reduce HO. It worth noting that in real environment, t_d varies randomly, and this is mainly due to different factors such as how sophisticated network equipment are, the channel response, and how easy it is to discover target BSs. Since t_d is part of reward, it has great influence in determining the learning and selection of BS that will maximise the cumulative reward. Also, it should be noted that t_d is a random parameter that is not constrained by design but dictated by the real environment. However, for the sake of RL design, one can select t_d as the critical parameter in order to control the trade-off. It can be clearly seen from Fig. 5 that changing the reward signal encourages the agent to select actions that substantially reduce the number of HOs by sacrificing the experienced SNR.

In the third experiment, we investigate the relationship between number of HO and SNR threshold (γ_{th}) for the different values of t_d for the proposed model, as shown in Fig. 6. In this experiment, we fix average velocity for UE to 8 ms^{-1} while other parameters remain unchanged. Fig. 6(a) illustrates the number of HO against γ_{th} when λ

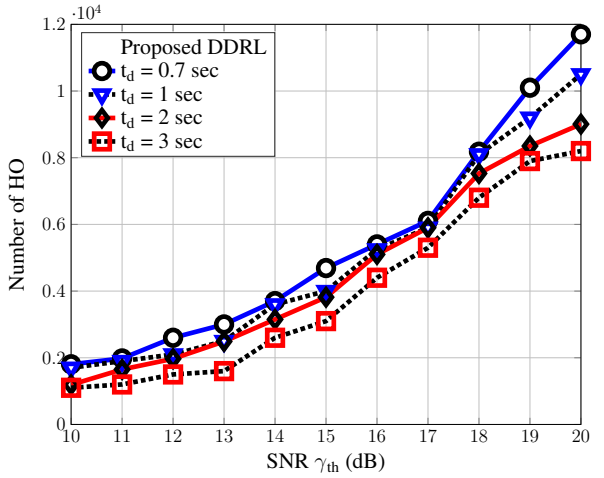


(a) The number of HO and Experienced SNR against different values of HO delay time t_d, for γ_{th} = 20 dB, and λ=10 BSKm^{-2}

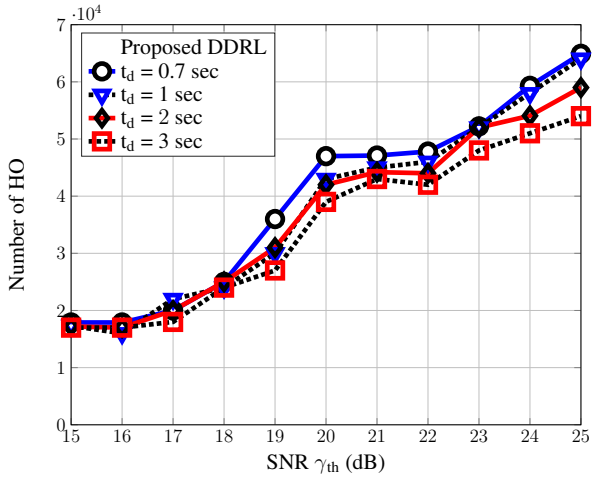


(b) The number of HO and Experienced SNR against different values of HO delay time t_d, for γ_{th} = 20 dB, and λ=50BSKm^{-2}

Fig. 5: The relation between number of HO, and Experienced SNR against HO processing time



(a) The number of HO against SNR threshold when HO average delay time of system t_d are $1, 3$ sec respectively for λ = 10 BSkm^{-2}



(b) The number of HO against SNR threshold when HO delay time of system t_d are $1, 3$ sec respectively for λ = 50 BS/km^2

Fig. 6: The relationship between numer of HO's for different service type(γ)

= 10 BSkm^{-2} while Fig. 6(b) shows the same relationship as Fig. 6(a) except that the value of λ = 50 BSkm^{-2}. From both figures, it can be understood that different value of t_d have a different contribution towards reinforcing the agent to reduce the number of HOs. However, for some values of γ_{th}, We can see subtle difference in the number of HO for value of γ_{th} from 10 to 18 when λ = 10 BSKm^{-2} and from 15 dB to 23 when λ = 50 BSKm^{-2}. We conclude the analysis by inferring that as λ increases for a given γ_{th} different values of t_d ranging between 0.75 and 3 sec have minor contribution towards reinforcing the agent to reduce number of HOs. Another interesting point from the both figures that can be seen is fluctuation of all curves, meaning for different values of t_d and same value of γ_{th}, we can have the same or different number of HOs. This is because during the evaluation we set the value of exploration ϵ to 0.0002, although the trend can also be understood from explanation given earlier. Therefore, an agent may ignore selecting BS that can result in maximum reward and choose to explore the environment by selecting the BS with maximum SNR. This effect can be neglected if the agent can set the exploration value to zero after learning phase. The factors for selecting t_d may vary depend on the agent's objective. Factor such as network response is not for agent to decide, and intuitively the network with the lower t_d provides good QoS and QoE. However the agent can decide to use higher t_d in training phase if the objective is to minimize HO due to other advantages the agent could gain by doing so.

For the forth experiment, we assess the performance of proposed model by comparing it with other benchmarks solution in term of number of HO and system throughput. The parameters for the experiment set are as follow: t_d = 3 sec, γ_{th} = 20 dB and λ = 50 BSkm^{-2} while other parameters remain unchanged. Fig. 8 demonstrate the average system throughput and number of HO for the three HO management policy against UE velocity. From Fig. 8(a) it can be seen that the proposed DDRL outperform other policies, and general trend shows slight and gradual increase in number of HO for all three model. The most interesting result is shown in Fig. 8(b) where proposed model outperform other model by far for most of high mobility UE. Although it can be observed that as UE velocity increases, the system throughput decreases, and this

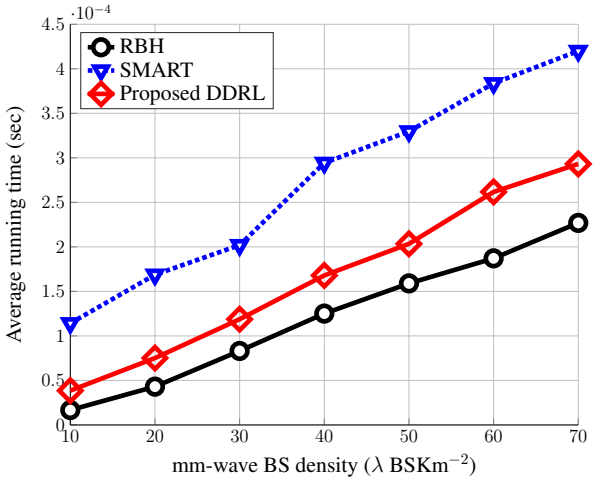


Fig. 7: Average running time as a function of number of mm-wave BS.

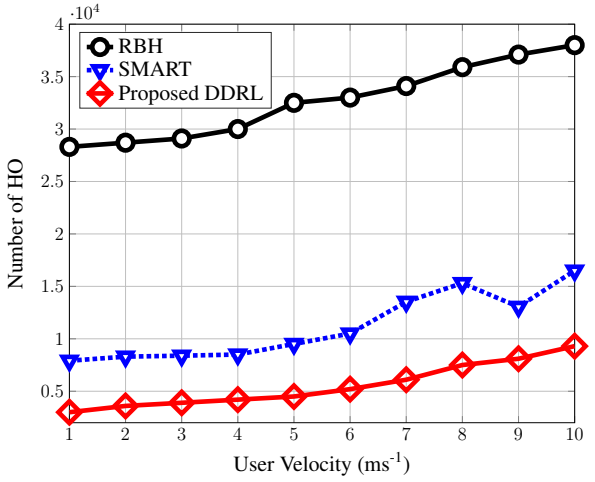
is due to rapid change in channel quality.

In the fifth experiment, the average running time per HO is evaluated against the number of mm-wave BS for the three HO strategies. The simulation parameters are as follows: t_d = 3 sec, UE velocity = 8 ms^{-1}, and γ_{th} = 20 dB. Fig. 7 shows that all the policies follow a similar trend. It can be observed that our proposed model takes more time to make HO decision compared to RBH; whereas it takes a significantly lower time compared to SMART. In addition, there is linear relationship between the increase in the number SCs and the running time for all policies.

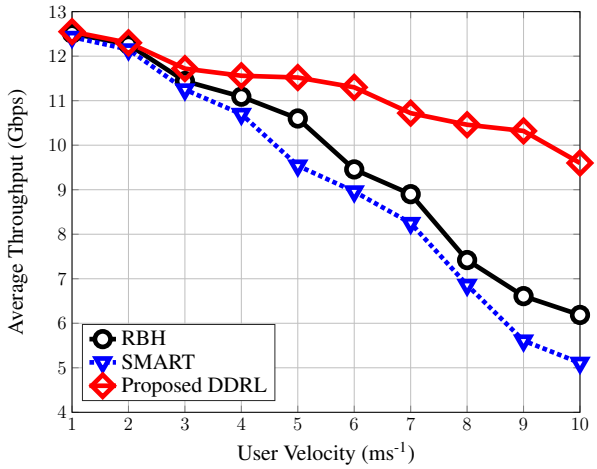
In the last experiment, we compare the performance of the three HO policies by varying the number of mm-wave BSs. We set the experiment parameter as follow: UE velocity = 8 ms^{-1}, γ_{th} = 20 dB while other parameters remain unchanged. The evaluation results are shown in Fig. 9, and it can be clearly observed that the proposed DDRL solution outperform the other two HO policies. It can be seen in Fig. 9(a) that the proposed HO policy has outstanding performance on reducing the number HOs by 20% - 69% and 7% - 49% compared to RBH and SMART, respectively. As expected Fig. 9(b) shows that the proposed model outperforms other models in terms of system throughput for RBH and SMART by 19% - 40% and 24% - 37%, respectively.

VII. Conclusion

In this study, we present an intelligent HO management framework for mm-wave communications in UDN scenario to minimize the frequency of HO occurrence, which in turn enhances the QoS of the users. In particular, we propose a DDRL algorithm with offline learning framework, such that historical user trajectory information is leveraged in order to develop a policy that ensures the selection of the optimal BS during HOs by considering both the number of HOs and system throughput. In this context, the UE is supposed to achieve the maximum possible throughput from the selected optimal BS, with which a more extended UE-BS association is constructed. Besides, instead of the actual geo-location coordinates, the proposed model exploits the combination of different SNR values received at a point to represent the locations of UEs, since the exact locations would not always be available. This enhances the feasibility, efficiency, and effectiveness of the developed BS selection policy, and the numerical results demonstrate that the designed DDRL algorithm significantly outperformed both the conventional and existing AI-based HO policy in different scenarios.



(a) The number of HO



(b) Average system throughput

Fig. 8: Relationship between HO performance and UE velocity

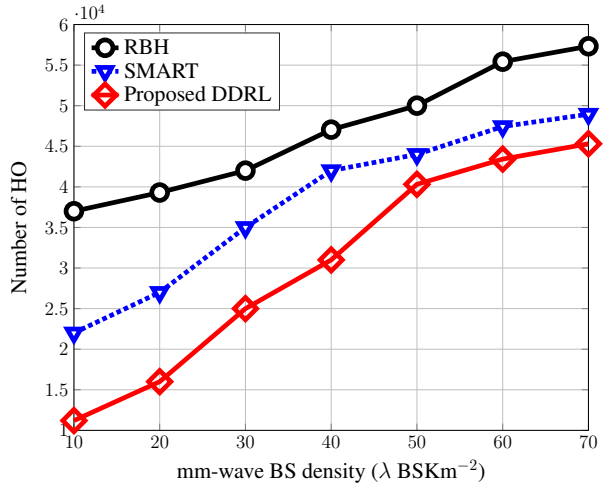
References

[1] S. A. Busari, S. Mumtaz, S. Al-Rubaye, and J. Rodriguez, "5G millimeter-wave mobile broadband: Performance and challenges," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 137–143, Jun. 2018.

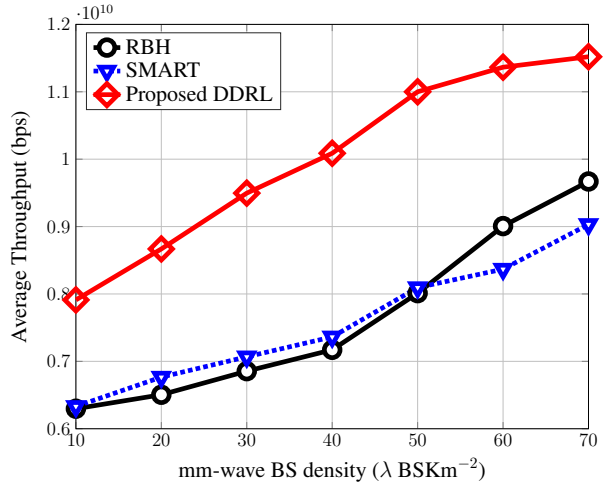
[2] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[3] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 64–71, June 2017.

[4] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave



(a) The BS ratio HO



(b) Average system throughput

Fig. 9: Comparison of performance with different number mm-wave BS

mobile communications for 5g cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.

[5] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, Jan 2017.

[6] E. Ben-Dor, T. S. Rappaport, Y. Qiao, and S. J. Lauffenburger, "Millimeter-wave 60 GHz outdoor and vehicle AOA propagation measurements using a broadband channel sounder," pp. 1–6, 2011.

[7] C. Fiandrino, H. Assasa, P. Casari, and J. Widmer, "Scaling millimeter-wave networks to dense deployments and dynamic environments," *Proceedings of the IEEE*, vol. 107, no. 4, pp. 732–745, 2019.

[8] R. Baldemair, T. Irnich, K. Balachandran, E. Dahlman, G. Mildh, Y. Selén, S. Parkvall, M. Meyer, and A. Osseiran, "Ultra-dense networks in millimeter-wave frequencies," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 202–208, Jan. 2015.

[9] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[10] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in 5G and beyond: A topology aware skipping approach," *IEEE Access*, vol. 4, pp. 9073–9081, 2016.

[11] ——, "Velocity-aware handover management in two-tier cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1851–1867, 2017.

[12] A. Talukdar, M. Cudak, and A. Ghosh, "Handoff rates for millimeter-wave 5G systems," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–5.

[13] B. Van Quang, R. V. Prasad, and I. Niemegeers, "A survey on handoffs—lessons for 60 GHz based wireless systems," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 1, pp. 64–86, 2012.

[14] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Handover management in dense cellular networks: A stochastic geometry approach," in *2016 ieee international conference on communications (icc)*. IEEE, 2016, pp. 1–7.

[15] E. Demarchou, C. Psomas, and I. Krikidis, "Mobility management in ultra-dense networks: Handover skipping techniques," *IEEE Access*, vol. 6, pp. 11 921–11 930, 2018.

[16] M. Mezzavilla, S. Goyal, S. Panwar, S. Rangan, and M. Zorzi, "An MDP model for optimal handover decisions in mmwave cellular networks," in *2016 European conference on networks and communications (EuCNC)*. IEEE, 2016, pp. 100–105.

[17] S. Goyal, M. Mezzavilla, S. Rangan, S. Panwar, and M. Zorzi, "User association in 5G mmwave networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[18] S. Zang, W. Bao, P. L. Yeoh, H. Chen, Z. Lin, B. Vucetic, and Y. Li, "Mobility handover optimization in millimeter wave heterogeneous networks," in *2017 17th International symposium on communications and information technologies (ISCIT)*. IEEE, 2017, pp. 1–6.

[19] C. Chaieb, Z. Mlika, F. Abdelkefi, and W. Ajib, "Mobility-aware user association in hetnets with millimeter wave base stations," in *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 2018, pp. 153–157.

[20] H. Tabassum, M. Salehi, and E. Hossain, "Mobility-aware analysis of 5G and B5G cellular networks: A tutorial," *arXiv preprint arXiv:1805.02719*, 2018.

[21] H. Zhang and L. Dai, "Mobility prediction: A survey on state-of-the-art schemes and future applications," *IEEE Access*, vol. 7, pp. 802–822, 2018.

[22] M. Ozturk, P. V. Klaine, and M. A. Imran, "Introducing a novel minimum accuracy concept for predictive mobility management schemes," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[23] M. Ozturk, M. Gogate, O. Onireti, A. Adeel, A. Hussain, and M. A. Imran, "A novel deep learning driven, low-cost mobility prediction approach for 5G cellular networks: The case of the control/data separation architecture (cdsa)," *Neurocomputing*, vol. 358, pp. 479 – 489, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231219300438

[24] M. Simsek, M. Bennis, and I. Güvenç, "Context-aware mobility management in hetnets: A reinforcement learning approach," in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2015, pp. 1536–1541.

[25] N. Aljeri and A. Boukerche, "A two-tier machine learning-based handover management scheme for intelligent vehicular networks," *Ad Hoc Networks*, p. 101930, 2019.

[26] A. G. Mahira and M. S. Subhedar, "Handover decision in wireless heterogeneous networks based on feedforward artificial neural network," in *Computational Intelligence in Data Mining*. Springer, 2017, pp. 663–669.

[27] S. Kumari and B. Singh, "Data-driven handover optimization in small cell networks," *Wireless Networks*, pp. 1–9.

[28] N. M. Alotaibi and S. S. Alwakeel, "A neural network based handover management strategy for heterogeneous networks," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 1210–1214.

[29] Z. Ali, N. Baldo, J. Mangues-Bafalluy, and L. Giupponi, "Machine learning based handover management for improved QoE in LTE," in *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2016, pp. 794–798.

[30] Z. Wang, L. Li, Y. Xu, H. Tian, and S. Cui, "Handover control in wireless systems via asynchronous multiuser deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4296–4307, 2018.

[31] Y. Sun, G. Feng, S. Qin, Y. Liang, and T. P. Yum, "The smart handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1456–1468, June 2018.

[32] A. Alkhateeb, I. Beltagy, and S. Alex, "Machine learning for reliable mmwave systems: Blockage prediction and proactive handoff," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 1055–1059.

[33] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Reinforcement learning based predictive handover for pedestrian-aware mmwave

networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2018, pp. 692–697.

[34] Y. Koda, K. Nakashima, K. Yamamoto, T. Nishio, and M. Morikura, "End-to-end learning of proactive handover policy for camera-assisted mmwave networks using deep reinforcement learning," *arXiv preprint arXiv:1904.04585*, 2019.

[35] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter-wave communications: The sparse way," in *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 273–277.

[36] 3GPP TS 38.331, "NR; Radio resource control (RRC); Protocol specification (Release 15)," 2018.

[37] M. Mollel, M. Ozturk, M. Kisangiri, S. Kaijage, O. Onireti, M. A. Imran, and Q. H. Abbasi, "Handover management in dense networks with coverage prediction from sparse networks," in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, April 2019, pp. 1–6.

[38] B. Christensen and O. Knape, "Optimization of algorithms for mobility in cellular systems," 2016, student Paper.

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

[40] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, England, 1994, vol. 37.

[41] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," *arXiv preprint arXiv:1512.07679*, 2015.

[42] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," *CoRR*, vol. abs/1509.06461, 2015. [Online]. Available: http://arxiv.org/abs/1509.06461

[43] J. Brownlee, "Why one-hot encode data in machine learning," 2017.

[44] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, "Stochastic properties of the random waypoint mobility model," *Wireless Networks*, vol. 10, no. 5, pp. 555–567, Sep 2004. [Online]. Available: https://doi.org/10.1023/B:WINE.0000036458.88990.e5

[45] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 2094–2100. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016191