Reinforcement-Learning Based Handover Optimization for Cellular UAVs Connectivity

MAHMOUD ALMASRI, XAVIER MARJOU, FANNY PARZYSZ
2 Avenue Pierre Marzin, Orange Lab. Lannion, FRANCE

Abstract: The demand for services provided by Unmanned Aerial Vehicles (UAVs) is increasing pervasively across several sectors including potential public safety, economic, and delivery services. As the number of applications using UAVs grows rapidly, more and more powerful, quality of service, and power efficient computing units are necessary. Recently, cellular technology draws more attention to connectivity that can ensure reliable and flexible communications services for UAVs. In cellular technology, flying with a high speed and altitude is subject to several key challenges, such as frequent handovers (HOs), high interference levels, connectivity coverage holes, etc. Additional HOs may lead to "ping-pong" between the UAVs and the serving cells resulting in a decrease of the quality of service and energy consumption. In order to optimize the number of HOs, we develop in this paper a Q-learning-based algorithm. While existing works focus on adjusting the number of HOs in a static network topology, we take into account the impact of cells deployment for three different simulation scenarios (Rural, Semi-rural and Urban areas). We also consider the impact of the decision distance, where the drone has the choice to make a switching decision on the number of HOs. Our results show that a Q-learning-based algorithm allows to significantly reduce the average number of HOs compared to a baseline case where the drone always selects the cell with the highest received signal. Moreover, we also propose which hyper-parameters have the largest impact on the number of HOs in the three tested environments, i.e. Rural, Semi-rural, or Urban.

Keywords: Drones Connectivity, Reinforcement Learning, Handovers Optimization, Decision Distance
Received: April 16, 2021. Revised: June 17, 2022. Accepted: July 16, 2022. Published: September 13, 2022.

1. Introduction

Over the last 10 years, the unmanned aerial vehicles (UAV) market has witnessed rapid evolution and exponential growth in a wide range of applications such as natural resource management, site monitoring, etc. Most existing drones these days should not be operated beyond direct line of sight and therefore use 2.4 and 5 GHz Wi-Fi for connectivity. Although current cellular networks were designed to meet the communication needs of user equipments (UEs) at low altitudes, they also represent a very promising connectivity solution for UAVs, as they offer wide coverage, quality broadband and secure connectivity. However, to further meet the UAV needs, existing cellular networks (like Long Term-Evolution (LTE) LTE and 5G networks) must integrate evolution to provide them with even better reliable, flexible and ubiquitous connectivity. To do so, the third-Generation Partnership Project ($3GPP$) has been developing new key performance indicators (KPIs) for enhanced LTE Support for Connected Drones [1]–[3]. For instance, Release-15 studied UAVs-dedicated models for Line of Sight (LoS) probability, pathloss and shadowing in order to enable robust and uninterrupted services to drones. Despite the efficiency of the proposed models in Release-15, integrating UAVs into the current cellular networks still suffer from several challenges:

- Current cellular networks are mainly designed to serve terrestrial users, thus requiring to down-tilt the antennas. Consequently, some drones may be served by side lobe of the antenna and may suffer from coverage holes in the sky due to the nulls in the antennas radiation pattern [4].
- At high altitude, the radio waves channel of BS-drones travel freely without obstacles. Subsequently, the channel of BS-drones is LoS with a high probability, and a drone may receive signals from many neighboring cells with a strong power level resulting in more interference in the down-link direction. This interference, if not properly controlled, may degrade the performance of the wireless communication network for both terrestrial and aerial users.
- Ensuring stable and robust connection for a flying drone represents a major challenge in future mobile networks. Indeed, depending on its speed and trajectory, a drone may perform unnecessary additional handovers compared to ground users that may lead to "ping-pong" between serving cell, resulting in a loss of radio connectivity, and deteriorate the Quality of Service (QoS) of BS-drone connectivity. In such scenario, managing the HOs among cells becomes an important issue to ensure robust connectivity.

In this paper, we focus on Reinforcement Learning (RL) algorithms in order to optimise the number of HOs in different network environments: Rural, Semi-rural or Urban. To this end, the drone attempts to minimize the number of HOs while maximizing the RSRP values. Therefore, the objective function contains two factors: 1-The RSRP value of the serving cell and 2- a penalty when performing a HO.

2. Related Works

ML algorithms have substantially increased in various research domains and applied fields especially in cellular technologies. They grow fast and extensively to handle the mobility management in cellular networks. In [5], the authors use RL algorithms in order to optimize handovers with user mobility under a dynamic small-cell network. In [6], the authors combine fuzzy-based function with the Q-Learning to control and optimize the HO and load balancing issue. By considering the velocities and locations of a user, the authors of [7] attempt to maximize the throughput of the terrestrial users under a given location and velocity by using RL optimal HO decision-making policy. The works of [8] implement a hidden Markov process in order to reduce latency mobile networks, learn the optimal control for HOs, and predict the next connected access point.

In [9], the authors propose a novel method to minimize the interference in a cellular network caused by the drones on the ground users using deep RL algorithms. In [10], cell selection and handover measurements are discussed for drones connected to an LTE in a suburban environment. Simulations show the increasing in the number of the HOs while increasing flight altitude. As discussed in the prior work, mobility challenges pertaining to drone communications is widely suggested in the literature. While, efficient HOs optimization for drone has received little attention. To this end, in this work, a HO mechanism based on Q-learning is investigated in different topology of cellular connected drone networks, i.e. Rural, semi-Rural, or Urban. We also suggest the impact of the hyper-parameters on the average number of HOs.

3. System Model

In this work, we consider three cellular networks topologies, each consisting of different number of geo-spatially deployed ground Base stations (BSs) in order to serve the UAVs. These latter are supposed to fly in a two-dimensional (2D) trajectory with a fixed height h_{UAV}. While flying, an UAV may operate several HOs by switching from a BS to another in order to maintain reliable connectivity. Several factors may lead to a HO process such as the BS distribution, the received signal at the UAV, its speed, height or trajectory.

3.1 Environment Generator

Let K represent the number of the base stations separated by a distance d_{BS}, and C represent the number of cells per base station. Three types of cellular network are considered, i.e. Rural, Semi-rural, and Urban, with an area of same length $L = \{-l/2, +l/2\}$ and width $W = \{-w/2, +w/2\}$ but different K and d_{BS} by taking into account the base station deployment in each environment. Propagation Path Loss (PL) estimation is an important constraint to formulate and design cellular networks. Generally, PL can be influenced by terrain contours, environment (Urban or Rural), propagation medium (dry or moist air), the distance between the transmitter and the receiver, and the height and location of antennas. We use two different definition of the PL, for Rural or Urban environment, introduced in the $3GPP$ reference as follows [1]:

$$PL_{\{Rural\}} = \max(23.9 - 1.8 * log_{10}(h_{UAV}, 20) * log_{10}(d_{3D}) + 20 * log_{10}(\frac{40 * \pi * fc}{3}) \quad (1)$$

$$PL_{\{Urban\}} = 28 + 22 * log_{10}(d_{3D}) + 20 * log_{10}(fc) \quad (2)$$

where h_{UAV} donates the height of the drone, d_{3D} represents the $3D$ distance from the drone to the base station, and fc is the transmission bandwidth. For a more realistic model, we also consider the standard deviation (σ) of the shadowing propagation in the environment defined in [1] as follows:

$$\sigma_{\{Rural\}} = 4.2 * exp(-0.0046 * h_{UAV}) \quad (3)$$

$$\sigma_{\{Urban\}} = 4.64 * exp(-0.0066 * h_{UAV}) \quad (4)$$

To evaluate the quality of the signal, we mainly focus on the Reference Signals Received Power ($RSRP$) as introduced in [11]:

$$RSRP = P_{tx} - 10 * log10(12 * fc) - PL - Sh + G_{UAV} + G_K$$

where P_{tx} represents the maximum transmit power from the base station, Sh donates the probability density function of the shadowing with a standard deviation σ. G_{UAV} and G_K respectively represent the antenna gain of UAV and the BSs.

3.2 Drone Trajectory Generator

At first, N drone trajectories are generated in order to train and test the RL algorithm: $2N/3$ are used to train the model and $N/3$ for testing.

We note that, the initial location and the destination for each trajectory are generated in the range of $\{-l/4, l/4\}$ and $\{-w/4, w/4\}$ in order to avoid border effect, dropped calls, access failures, and dead zones. We suppose that each trajectory is divided into several waypoints with a distance d_{UAV} between them. As long as the initial and final location for each trajectory are randomly generated, then each of them may have different length with different number of waypoints. When the initial location of each trajectory has been generated, the drone selects the shortest path to reach the final location. In particular, the drone selects a movement direction $\theta_s \in \{r.\pi/4, r = 0, 1, ..., 7\}$ and moves in a fixed distance d_{UAV} to get the next waypoint. This procedure is repeated until the drone reaches its final destination.

Let x_s and y_s represent the 2D drone's position, and c_s being the currently connected cell. We subsequently define $s = \{x_s, y_s, \theta_s, c_s\}$ as the state of the drone at each waypoint. Using eq. 3, we can obtain the $RSRP$ value for the k-strongest cells at each waypoint in the environment in which we define C_{k_s} that contains the k strongest cells at state s.

At each waypoint, the drone has to make an action A by selecting a serving cell among the k-strongest cells. We note that decision-making approaches are better in the long run as compared to the baseline approaches in which the

drone always selects the cell with the highest $RSRP$ value. Indeed, using the RL algorithms, especially the Q-learning, may significantly reduce the average number of HOs and prevent the "ping-pong" effect between the drone and serving cells. Moreover, it improves the Quality of service and reduce the overall energy consumption.

3.3 Q-learning

Reinforcement Learning (RL) is a popular ML algorithm for sequential decision making in which an agent interacts with its environment aiming to find the optimal action that maximizes the reward received from the environment [12]. RL is often described using a Markov decision process defined by a tuple (S, A, T, R):

- S donates a finite set of states,
- A donates a finite set of actions,
- $T : S \times A \rightarrow Pr(S)$ is referred to the transition probability over the states,
- R : is a reward function.

At each time slot, the agent observes the state $s \in S$, takes an action a, and finally receives the reward r from the environmental feedback. The main goal of the agent is to enhance its action a while maximizing the accumulated reward. With this information, the Markov decision process can be solved to get the optimal policy, i.e. the action to take at each time slot that maximizes the expected sum of discounted rewards. Q-learning [13] is a model-free RL algorithm to learn the optimal policy in a given state. Let us define the Q-value $Q^{\pi}(s, a)$ for a policy π as the expected rewards when the agent takes an action a in state s and chooses actions according to the policy π thereafter. The actions with the highest Q-values for each state provide the optimal policy [11], [13]. By selecting the action with the highest Q-value, the agent will eventually learn the optimal policy $Q^*(s, a)$ over time. Let $Q_t(s, a)$ denote the obtained Q-value at time t when the agent makes an action a in a state s. Therefore, the agent receives reward r_{t+1} and transitions to state s'. Therefore, the new Q-value can be obtained using the following expression:

$$Q_{t+1}(s,a) = (1-\alpha) * Q_t(s,a) + \alpha[r_{t+1} + \lambda * \max_{a' \in A} Q_t(s', a')] \quad (5)$$

where $\alpha \in [0, 1)$ is the learning rate, $\lambda \in [0, 1)$ gives the discount factor. Its full procedure is listed in Algorithm 1. The reward r received at each waypoint may combine between the RSRP and the HOs. We note that the main goal of the UAV is not only to reduce the average number of HOs but also maintain reliable connectivity. Then, the received reward r be the weighted combination between the RSRP and the HO cost defined as follows:

$$r = W_{RSRP} * RSRP - W_{HO} * I(HO) \quad (6)$$

where $I(HO) = 1$ if the serving cells at the current state and last one is different, and 0 otherwise. $RSRP$ represents the obtained $RSRP$ value from the serving cell.

Algorithm 1 Q-learning algorithm to optimize the HOs

Input parameters:
α, λ, ϵ, W_{RSRP}, W_{HO}
c_s: represents the currently connected cell at state s,
$_{RSRP_s}$: represents the $_{RSRP}$ value of the selected cell at s,
r_i: represents the obtained reward at the i-th waypoint,
Initialization:
while done==0 **do**
    #done = 1 indicates that the drone arrives to its
    #destination, and 0 otherwise,
    if $c_s \neq c_{s'}$ **then**
        HO=1
    else
        HO=0
    end if
    $r_i = RSRP_{s'}.W_{RSRP} - HO.W_{HO}$,
    $i = i + 1$,
end while
for Training step $\leq \frac{2.N}{3}$ **do**
    #Generate a random trajectory:
    $\mathcal{T} = \{(x_i, y_i, \theta_i) \mid i = 0, 1, ..., l-1\}$,
    State $s = \{x_s, y_s, \theta_s, c_s\}$,
    Action a: represents the selected action at state s,
    while done==0 **do**
        if $\epsilon > \zeta$ (a uniform random variable $\in [0, 1]$) **then**
            select a random action a
        else
            select the optimal action a^*:
            $a^* = \max_{a \in A} Q_i(s, a)$
        end if
        $Q_i(s,a) = (1-\alpha) * Q_i(s,a) + \alpha[r_i +$
        $\lambda * \max_{a' \in A} Q_{i+1}(s', a')]$,
        $s = s'$,
        $i = i + 1$,
    end while
end for

3.4 Experimental Processes

We evaluate the performance of the RL-based HO mechanism, with different weight, compared to the baseline case in which the drone always connects to the strongest cell. For each flight trajectory, we calculate a performance metric called HO ratio which we define as the ratio of the number of HOs using the proposed scheme to that for the baseline scheme.

At first, we generate three environments with different number of BS and distance between BSs d as follows:

- Rural: 9 BSs with a distance $d_{BS} = 3000\ m$ between BSs,
- Semi-Rural: 25 BSs with a distance $d_{BS} = 1500\ m$ between BSs,
- Urban: 100 BSs with a distance $d_{BS} = 500\ m$ between BSs.

where each BS has 3 cells.

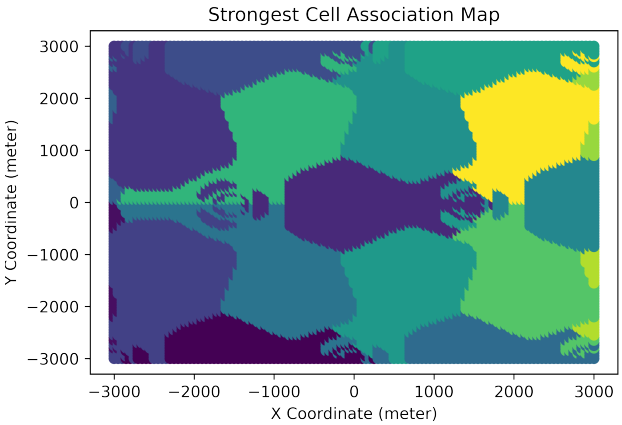

Fig. 1: Rural environment: $h_{UAV} = 120\ m$, 9 BSs, $d_{BS} = 3000\ m$

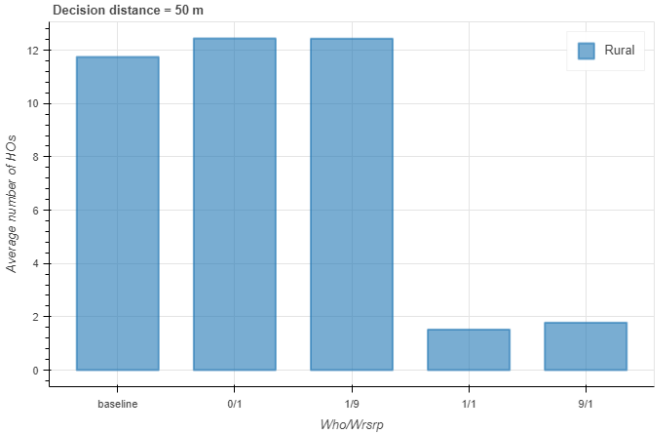

Fig. 2: Average number of HOs in a Rural environment

As an illustrative example, Fig. 1 shows the strongest cell at each waypoint in a Rural environment[1].

We compare the number of HOs using the Q-learning algorithm, with different values of W_{RSRP} and W_{HO}, to the baseline. Three main cases of weights can be considered:

- $W_{HO} = W_{RSRP}$
- $W_{HO} > W_{RSRP}$
- $W_{HO} < W_{RSRP}$

For the special case when there is no HO cost (i.e. $W_{HO} = 0$), the proposed RL-based HO scheme is equivalent to the baseline. As the ratio W_{HO}/W_{RSRP} increases, the number of HOs decreases and the HO ratio approaches zero.

We simulate the performance using 30000 runs (20000 for training and 10000 for testing) for the Q-learning. We also set the Q-learning parameters as follows λ = 0.3, α = 0.5, and ϵ = 0.2. For each run, the testing route is generated randomly as explained in Section III-B. We compare the obtained results with the baseline where the drone always selects the cell with the highest RSRP value. For each network topology, we also show the impact of the decision distance d, in which the drone has the choice to switch to another cell, on the number of HOs. Indeed, we suppose that the environment is divided into bins of size $d \times d$ m^2. For each bin, we obtain the k cells having the strongest RSRP value in that bin. In total, we collect about 15000 samples of RSRP values for different drone locations at an altitude of 120 m. RSRP samples are linearly normalized and transformed to the interval $[0,\ 1]$. As the decision distance increases, the number of HOs decreases.

4. Results and Discussion

In this section, we evaluate the performance of the Q-learning in the three environments (i.e. Rural, Semi-rural, Urban). We also investigate the impact of the decision distance on the average number of HOs.

[1]Fig. 1 shows the RSRP values in each waypoint excluding shadowing in order to clearly visualize the position of the BSs and easily show the geographical areas covered by 9 sectored BSs. However, in overall simulations, we consider the shadowing to generate the different Network Topologies: Rural, Semi-Rural or Urban.

In Fig. 2, we plot the average number of HOs for different weight combinations W_{HO}/W_{RSRP} in a rural environment. While the proposed scheme is approximately equivalent to the baseline when there is no HO cost (i.e. $W_{HO} = 0$), it can reduce the number of HOs by 85%, compared to the baseline, when $W_{HO}/W_{RSRP} \geq 1$ (see Tab. I).

In Fig. 3, we compare the average number of HOs in three different types of environment: Rural, Semi-Rural and Urban. As we can see, the average number of HOs in the Semi-Rural environment is enhanced by 77% compared to the baseline in the case of $W_{HO}/W_{RSRP} \geq 1$. As well as, in the Urban environment the average number of HOs in a flight, in the case of $W_{HO}/W_{RSRP} = 1/1$ and when $W_{HO}/W_{RSRP} = 9/1$, is respectively enhanced by 41% and 29% compared to the baseline case. As expected, the Q-learning algorithm performs more efficiently in a Rural environment where there are fewer cell candidates compared to an Urban one. However, Q-learning algorithm still have a fundamental role to decrease the average number of HOs in Rural or Urban environment.

Fig. 4 compares the average number of HOs in the Rural environment with different decision distance. While in the baseline case the average number of HOs is significantly increased with the decision distance, this later could not affect the HOs using the Q-learning. Indeed, in the case of $W_{HO}/W_{RSRP} \geq 1$, the average number of HOs is approximately the same for the three decision distance case: d = 50 m, 100 m, 150 m. Moreover, the average number of HOs is decreased by 85% compared to the baseline case. We note that, $W_{HO}/W_{RSRP} < 1$ represents the worst case in terms of the average number of HOs.

Fig. 5 shows the average number of HOs in the semi-Rural

TABLE I: Average HOs in the three environment with d = 50 m

Topology	Baseline	Q-learning		
		0/1	1/1	9/1
Rural	11.7	12.4 (**0%**)	1.5 (**87%**)	1.7 (**85%**)
Semi-rural	17.8	18,1 (**0%**)	3.5, (**80%**)	4.1 (**77%**)
Urban	24.7	24.7 (**0%**)	14.5 (**41%**)	17.6 (**29%**)

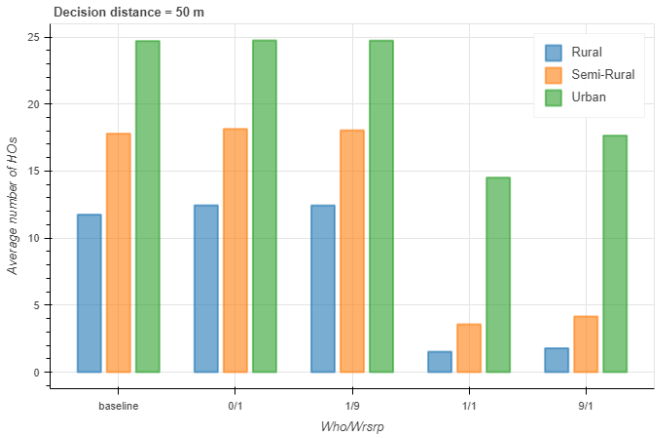

Fig. 3: Compare the average number of HOs in different networks topology: Rural, Semi-rural, Urban

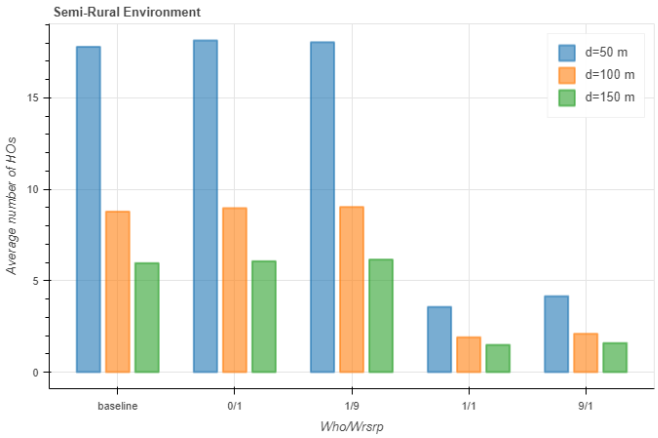

Fig. 5: Compare the average number of HOs with different decision distance in the semi-Rural environment

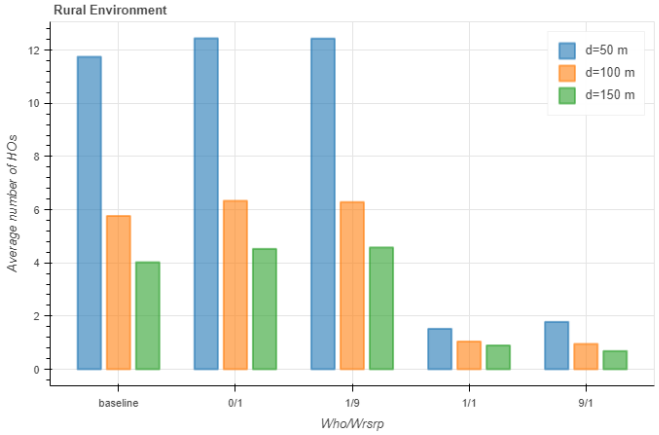

Fig. 4: Compare the average number of HOs with different decision distance in the Rural environment

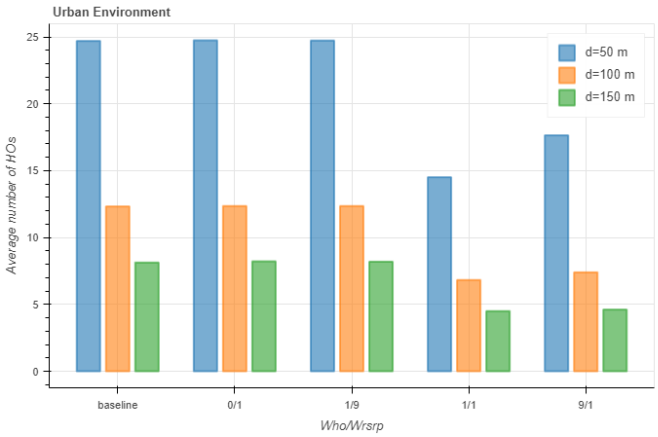

Fig. 6: Compare the average number of HOs with different decision distance in the Urban environment

environment with the three different decision distance cases. While, the average number of HOs in overall cases is increased compared to the Rural environment, it can be shown that the average number of HOs for the case of $W_{HO}/W_{RSRP} \geq 1$ is slightly increased for the three decision distance: d = 50 m, 100 m, 150 m.

Finally, Fig. 6 shows the average number of HOs in the Urban environment in which we notice a significant change in terms of the average number of HOs for the three different distance. Moreover, the case of $W_{HO}/W_{RSRP} < 1$, across the three decision distance, still represents the worse case and almost achieves the same average number of HOs as in the baseline case.

5. Conclusion

In this work, we have used Q-learning algorithm for HO decision making mechanism to achieve robust drone connectivity in a cellular-connected drone network. Using $3GPP$ formulas, we first generated three representative environments: Rural, Semi-Rural and Urban. We tested the Q-learning algorithm in the generated networks for a given flight trajectory. The simulation results have revealed that using Q-learning algorithm can significantly reduce the number of HOs in the three networks while maintaining reliable connectivity, compared to the baseline HO scheme in which the drone always connects to the strongest cell. Moreover, we investigated the performance of Q-learning in the three environments while changing the decision distance.

In future work, several points can be suggested such as considering the 3D drone mobility to attempt obtaining even more realistic simulation. Moreover, considering the case of the multi-Mobile Network Operators (MNOs) still represents an important task in order to make the model more realistic than the case of a single MNO.

References

[1] 3GPP TR 36.777, "Enhanced LTE support for aerial vehicles," 2017.
[2] 3GPP TR 22.825, "Study on remote identification of unmanned aerial systems," 2018.
[3] S. D. Muruganathan, X. Lin, H.-L. Maattanen, Z. Zou, W. A. Hapsari, and S. Yasukawa, "An overview of 3GPP Release-15 study on enhanced LTE support for connected drones," arXiv preprint arXiv:1805.00826, 2018.

[4] X. Lin, R. Wiren, S. Euler, A. Sadam, H. Maattanen, S. Muruganathan, S. Gao, Y. . E. Wang, J. Kauppi, Z. Zou, and V. Yajnanarayana, "Mobile network-connected drones: Field trials, simulations, and design insights," IEEE Veh. Technol. Mag., vol. 14, no. 3, pp. 115–125, Sep. 2019.
[5] M.T. Nguyen, S. Kwon, Machine Learning–Based Mobility Robustness Optimization Under Dynamic Cellular Networks. IEEE Access 2021, 77830–77844.
[6] S.A. Hashemi, H. Farrokhi, Mobility robustness optimization and load balancing in self-organized cellular networks: Towards cognitive network management. J. Intell. Fuzzy Syst. 2020, 38, 3285–3300.
[7] Y. Koda, K. Yamamoto, T. Nishio, M. Morikura, Reinforcement learning based predictive handover for pedestrian-aware mmWave networks. In Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu, HI, USA, 15–19 April 2018; pp. 692–697.
[8] Z. Wang, L. Li, Y. Xu, H. Tian, S. Cui, Handover control in wireless systems via asynchronous multi-user deep reinforcement learning. IEEE Internet Things J. 2018, 5, 4296–4307.
[9] U. Challita, W. Saad, C. Bettstetter, (2019). Interference management for cellular-connected UAVs: A deep reinforcement learning approach. IEEE Transactions on Wireless Communications, 18(4), 2125-2140.
[10] A. Fakhreddine, C. Bettstetter, S. Hayat, R. Muzaffar, and D. Emini, "Handover challenges for cellular-connected drones," in Proc. 5th Workshop on Micro Aerial Vehicle Networks, Systems, and Applications, 2019, pp. 9–14.
[11] Qualcomm Technologies, Inc. "LTE Unmanned Aircraft Systems Trial Report,"2017,
[12] R. S. Sutton, A. G. Barto et al., Introduction to reinforcement learning. MIT press Cambridge, 1998, vol. 2, no. 4.
[13] C. J. Watkins and P. Dayan, "Q-learning," Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.