Dynamic Antenna Control for HAPS Using Fuzzy Q-Learning in Multi-Cell Configuration

Kenshiro Wada[1], Siyuan Yang[2], Mondher Bouazizi[3] and Tomoaki Ohtsuki[4]
Yohei Shibata[5] Wataru Takabatake[6] Kenji Hoshino[7] Atsushi Nagate[8]
[1,2]Graduate School of Science and Technology, Keio University
[3,4]Department of Information and Computer Science, Faculty of Science and Technology, Keio University
[5,6,7,8] Softbank Corporation, Japan
Email: [1]wada@ohtsuki.ics.keio.ac.jp, [2]yang@ohtsuki.ics.keio.ac.jp,
[3]bouazizi@ohtsuki.ics.keio.ac.jp, [4]ohtsuki@ics.keio.ac.jp

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838271

***Abstract*—In the 5th generation mobile communications (5G) and 5G and beyond (B5G), a high altitude platform station (HAPS) is expected to serve as a flying base station (BS) to provide communications over wide areas. In the HAPS system, a multi-cell configuration with multiple beams is considered to increase system throughput. When the HAPS is subjected to wind pressure, the cell range moves accordingly, causing degradation of received signal power and handover to the user equipment (UE). To suppress such degradation and handover, beam control of HAPS is necessary. However, it is not easy to control the beam because multiple antenna parameters affect each other and determine the cell range. In this paper, we propose a beam control method for HAPS using fuzzy Q-learning in multi-cell configuration. In this type of learning, the variable states are controlled by the use of fuzzy sets, which allows multiple searches to be performed in one setup, thus reducing the cost of search, compared with conventional Q-learning. In the proposed beam control method, antenna parameters are controlled by fuzzy Q-learning so that the number of users having a received signal power larger than a predetermined threshold becomes larger in each cell. We evaluate the proposed method by computer simulation and show that the proposed method can improve the number of users having a received signal power larger than a predefined threshold and thus reduce the number of users with low throughput compared to before learning.**

I. Introduction

The 5th generation mobile communication system (5G) is attracting attention because it can provide communications with low latency, high speed, multiway connectivity, low cost, and low power consumption. To enable such communications, many providers are considering the use of high-frequency millimeter waves. In particular, the use of millimeter-wave is being considered for the high altitude platform station (HAPS) system [1]. HAPS is an aircraft operating in the stratosphere at altitudes of 17–20 km and equipped with the wireless base station (BS) functions [1]. The International Telecommunications Union (ITU) has designated 47 GHz, a high-frequency millimeter-wave band, as the HAPS band [2]. HAPS is less expensive to install than terrestrial BSs due to the miniaturization of solar cells and devices and can provide wide-area line-of-sight (LoS) communications. In such a wide area HAPS system, a multi-cell configuration with multiple beams is considered to increase the system throughput [3]. In such a multiple-cell configuration, the wind-pressure-induced movement of the HAPS causes the cell range to shift, which in turn causes degradation of received signal power and handovers to user equipment (UE) [4], [5]. For HAPS, antenna control is needed to suppress such degradation and handovers [6]. Thus, there are many studies on antenna control of HAPS [7]–[9], [12]–[18]. The antenna control methods to maximize cell coverage by suppressing interference between neighboring cells [8], and the antenna control to suppress interference with other HAPS are considered [9]. There are four beam control parameters for HAPS: horizontal tilt, vertical tilt, horizontal beam half power beam width (HPBW), and vertical beam HPBW, each of which affect each other to determine the cell range. It is not easy to control the antenna parameters appropriately so that the degradation of received signal power and handovers is reduced. Besides, the conventional beam control methods need to know the 3D coordinates of HAPS and UEs (target cells) and also the movement of the HAPS. Thus, some reinforcement learning method can be considered to solve this problem [10], [11].

Q-learning is applied to control antennas in some papers. Q-learning is a type of reinforcement learning

978-1-5386-8347-7/22/$31.00 ©2022 IEEE

in which an agent learns an optimal course of action in a search process based on rewards obtained from an evaluation formula instead of supervisory data. In Q-learning, an agent estimates the future discounted reward by executing an action from a specific state using a method that builds the Q function step by step [12]. In [13] fuzzy Q-learning is used for controlling a down tilt angle of the long term evolution (LTE) BS antennas. Fuzzy Q-learning is one of the Q-learning algorithms. It is characterized by the use of fuzzy sets for the states of the variables to be controlled, which allows multiple searches to be performed in a single training. Therefore, the search cost is lower than that of ordinary Q-learning [14].

In this paper, we propose a beam control method for HAPS based on fuzzy Q-learning in multi-cell configuration. Since the fuzzy Q-learning allows multiple searches to be performed in a single training, we expect to control the beam parameters with low cost and high convergence speed in dynamic environments. Besides, different with conventional methods which need to know the coordinates of HAPS and UEs, we predefine a received power threshold and judge the quality of communication by whether the user's SINR is greater than or equal to the threshold. We evaluate the proposed method by computer simulation and show that the proposed method can improve the number of users having a received signal power larger than a predefined threshold and reduce the number of users with low throughput compared to before learning.

The rest of the paper is organized as follows. Section II explains the HAPS model, antenna model, and introduces the fuzzy Q-learning algorithm. We present in Section III the proposed method. The simulation results are shown and discussed in section IV. Finally, the Conclusion is shown in Section V.

II. System Model

A. Model of HAPS

In this paper, we consider a standalone HAPS system in which HAPS receives signals from a gateway on the ground, which is connected to the core network, and act as a repeater to relay the signals to a wide range of users. The altitude of the HAPS is assumed to be 20 km, and the coverage area per HAPS is about 20 km in radius when the elevation angle is 45 degrees. The HAPS is equipped with multiple antennas, and the radius of 20 km is covered by multi-cells for each HAPS's antenna.

B. Model of Antenna

In this paper, we use an antenna with a gain $G(\phi)$ at an angle ϕ from the beam direction, as expressed in the ITU-R based equations [15] [16]. Here, ϕ_b is HPBW divided by 2, $\phi_1 = \phi_b\sqrt{-L_N/3}$, $\phi_2 = 3.745\phi_b$, $X = L_N + 60\log_{10}(\phi_2)$, $\phi_3 = 10^{(X-L_F)/60}$.

$$G(\phi) = \begin{cases} -3(\phi/\phi_b)^2, & (0 \leq \phi \leq \phi_1) \\ L_N, & (\phi_1 \leq \phi \leq \phi_2) \\ X - 60\log_{10}(\phi), & (\phi_2 \leq \phi \leq \phi_3) \\ L_F, & (\phi_3 \leq \phi \leq 90). \end{cases} \tag{1}$$

Since the antenna gain is obtained by combining the horizontal and vertical gains, the combined gain G can be expressed as Eq. (2) [15], where G_m is the maximum gain in the main lobe, L_N is the near-in-sidelobe level (dB) relative to the peak gain required by the system design, and L_F is the far side-lobe level (dB) relative to the peak gain required by the system design. Thus, the antenna gain can be expressed as

$$G = \max(G_v + G_h, L_F) + G_p \tag{2}$$

where G_v and G_h are the vertical and horizontal antenna gains, respectively, and G_p is the maximum antenna gain.

C. Fuzzy Q-Learning

Fuzzy Q-learning is a method of representing actions and Q functions using fuzzy inference systems (FIS). The Q function is stored in a lookup table indexed by state and action. If a high reward is obtained in a state, the reward is propagated to the states that can reach that state with each update, and learning takes place. The Q-value at the time step t is represented as $Q_t(s_t, a_t)$, s_t is the state and a_t is the selected action at the time step t. When the action a_t is selected and performed, the system transits to a new state s_{t+1} and receives a reward r_t. In this type of Q-learning, if the number of states and actions increases, the lookup table for storing Q-values becomes very large, and the search time increases. For this reason, the use of FIS has been considered for storing the Q-values in the lookup table [17]. By using the fuzzy set, approximated values can be used for states and actions, which not only reduces the lookup table size but also reduces the search time since prior knowledge can be embedded in the FIS. In fuzzy Q-learning, states are represented using the fuzzy sets. Here, $\boldsymbol{a}^{\dagger}_{i(=1,\ldots,n)}$ represents the action selected for each fuzzy set M_i. The Q-value for the state S_i and the action $\boldsymbol{a}_{j(=1,\ldots,n)}$ can be expressed as $Q(M_i, \boldsymbol{a}_j)$. The Q-value for the action $\boldsymbol{a}^{\dagger}_{i(=1,\ldots,n)}$ is then updated using the TD(λ) method [18], as in the following Eqs.

(3)-(6) where γ is the discount rate. In Eq. (5), $\boldsymbol{a}_i^*$ is the action that maximizes the Q-value in the state S_i, i.e., $\boldsymbol{a}_i^* = \arg\max_{j(=1,\ldots,n)} Q_t(S_i, \boldsymbol{a}_i)$. $e(S_i, \boldsymbol{a}_j)$ is the eligibility traces that represent a memory of which state (or state-action) values. To implement the TD(λ) algorithm, the eligibility traces $e(S_i, \boldsymbol{a}_j)$ is added to sum the temporary fuzzy membership value when the selected action $\boldsymbol{a}_j$ is the same as that in the previous state. In the fuzzy Q-learning process, an action $\boldsymbol{a}_i^\dagger$ is selected for each fuzzy set M_i using the ϵ-greedy method so that appropriate Q-values for various actions can be learned without depending on the initial Q-values. The action $\boldsymbol{a}_{i(=1,\ldots,n)}$ of each selected fuzzy set is weighted by the FIS agreement $\mu_i(x)$ and added together to obtain the action $\boldsymbol{a}$, as shown in Eq. (7).

$$\begin{aligned} Q_{t+1}(s_t, \boldsymbol{a}_t) &= Q_t(s_t, \boldsymbol{a}_t) \\ &+ \alpha[r_t + \gamma \max_{\boldsymbol{a}} Q_t(s_{t+1}, \boldsymbol{a}) - Q_t(s_t, \boldsymbol{a}_t)] \end{aligned} \tag{3}$$

$$Q_t(M_i, \boldsymbol{a}^\dagger{}_i) = Q_t(M_i, \boldsymbol{a}^\dagger{}_i)_t + \varepsilon \times \Delta Q \times e(M_i, \boldsymbol{a}^\dagger{}_i) \tag{4}$$

$$\begin{aligned} \Delta Q = r + \gamma \times \frac{\sum_{i=1}^{n} \mu_i(x) \times Q(M_i, \boldsymbol{a}^*{}_i)}{\sum_{i=1}^{n} \mu_i(x)} \\ - \frac{\sum_{i=1}^{n} \mu_i(x) \times Q(M_i, \boldsymbol{a}^\dagger{}_i)}{\sum_{i=1}^{n} \mu_i(x)} \end{aligned} \tag{5}$$

$$e(S_i, \boldsymbol{a}_j) = \begin{cases} \lambda\gamma e(M_i, \boldsymbol{a}_j) + \frac{\sum_{i=1}^{n} \mu_i(x)}{\sum_{i=1}^{n} \mu_i(x)}, & (\boldsymbol{a}_j = \boldsymbol{a}_i^\dagger) \\ \lambda\gamma e(M_i, \boldsymbol{a}_j), & \text{otherwise} \end{cases} \tag{6}$$

$$\boldsymbol{a}(x) = \frac{\sum_{i=1}^{n} \mu_i(x) \times \boldsymbol{a}^\dagger{}_i}{\sum_{i=1}^{n} \mu_i(x)} \tag{7}$$

III. Proposed Method

In this paper, we propose a beam control method based on fuzzy Q-learning.

A. State, action, and reward in Q learning

In the proposed method, the state and behavior of Q-learning are set so that each antenna beam can be controlled to cover the target cell range even when the HAPS position changes. The antenna key performance indicator (KPI) is used as the state, and the KPI is set as the coverage ratio of the beam for each antenna i in a predetermined cell C_i, as shown in Eq. (8). The KPI represents the percentage of users u whose received power $\gamma_i(u)$ from the antenna i is greater than or equal to the received power threshold Γ among the set of users U_i located in the range of C_i as shown in Eq. (8).

$$KPI_i = \frac{|\{u(\in U_i)|\gamma_i(u) \geq \Gamma\}|}{|U_i|} \tag{8}$$

In addition, we define an action $\boldsymbol{a}_t^\dagger$ at each time step t such that $\boldsymbol{a}_t^\dagger \in \mathcal{A}$. Here, $\mathcal{A} = [\Delta\phi_{3dB}, \Delta\phi_{tilt}, \Delta\theta_{3dB}, \Delta\theta_{tilt}]$, where $\Delta\phi_{3dB}$, $\Delta\phi_{tilt}$, $\Delta\theta_{3dB}$, and $\Delta\theta_{tilt}$ denote the increase or decrease of the horizontal HPBW and tilt, and the vertical HPBW and tilt, respectively. By adjusting the values of these 4 antenna parameters, we can control the received power at the users' end, thus, enhance the KPI of each cell.

We define the lower and upper bounds of KPI, KPI_{bad}, KPI_{good}, to obtain rewards. Here, a reward of 1 is given when the cell coverage of the antenna is sufficient, i.e., $KPI \geq KPI_{good}$, a reward of -1 is given when the cell coverage of the antenna is insufficient, i.e., $KPI < KPI_{bad}$, and a reward of 0 is given when $KPI_{good} > KPI \geq KPI_{bad}$. If the number of iterations for action selection and Q-table update exceeds the maximum number of iterations, -1 is given as a reward even in the case where $KPI_{good} > KPI \geq KPI_{bad}$. If -1 or 1 is given as a reward, the antenna parameters and the number of iterations are initialized and the iterations are performed again. According to rewards, the antenna parameters can be controlled so that the number of outage users is decreased. The Q-learning algorithm then learns by repeatedly updating the Q-table with rewards according to the action selection and KPI. When the Q table is sufficiently learned, each antenna can be controlled to cover the cell sufficiently with the minimum number of actions.

B. HAPS antenna control using fuzzy Q-learning

In the HAPS antenna control using fuzzy Q-learning, a fuzzy set $M_{k(=1,\ldots,n)}$ is established for the state KPI as shown in Fig. 1, and a membership function F_k is defined for each fuzzy set M_k as shown in Eq. (9). Fuzzy sets control variable states so that multiple searches can

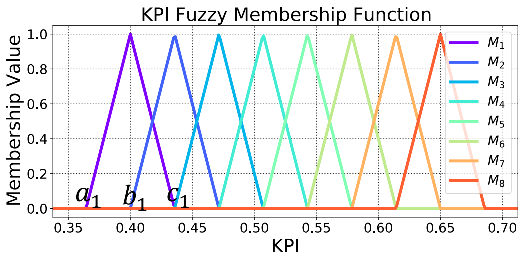


Fig. 1. KPI Fuzzy membership sets.

be performed within a single setup, reducing the cost of search in comparison with conventional Q-learning [19]. In selecting the action, for each fuzzy set M_k, the update $\boldsymbol{a}_k^\dagger$ of each antenna is selected and multiplied by the fuzzy membership value $F_k(KPI)$ and added together for all the fuzzy sets to obtain the antenna control value. In this way, all the antenna parameters can be controlled in a single action, resulting in fast convergence to the optimal solution.

$$F_k(KPI) = \begin{cases} 0, (KPI < a_k, KPI > c_k) \\ \frac{KPI - a_k}{b_k - a_k}, (a_k < KPI < b_k) \\ -\frac{KPI - b_k}{c_k - b_k}, (b_k < KPI < c_k) \end{cases} \quad (9)$$

$$(\Delta\phi_{3dB}, \Delta\phi_{tilt}, \Delta\theta_{3dB}, \Delta\theta_{tilt}) = \sum_k F_k(KPI) \times \boldsymbol{a}_k^\dagger \quad (10)$$

IV. Performance Evaluation

A. Simulation Specifications

The proposed method was evaluated by computer simulations under the environments shown in Table I to Table III. We assume that the HAPS located at (0, 0) is the control target and 18 HAPS with three antennas and a coverage radius of 20 km are placed around it. Thus, each HAPS serves three cells. The signals of these HAPSs are assumed to interfere with each other. The antenna update widths at the time of action selection for Q-learning were set to 1 deg, 10 deg, 1 deg, and 2 deg for horizontal half-width, vertical half-width, horizontal tilt, and vertical tilt, respectively, and the allowable received power Γ at KPI was set to -79 dBm. We control four antenna parameters in order: horizontal tilt, vertical tilt, horizontal HPBW, and vertical HPBW.

B. SINR Distribution

In this subsection, we evaluate the SINR distribution under HAPS with rotation and shift cases. Fig. 2 show the initial SINR distribution when the HAPS is rotated by 30 degrees and SINR distribution after antenna control by the fuzzy Q-learning. Fig. 3 show the initial SINR distributions when the HAPS has moved -5 km to the left and SINR distribution after antenna control by the fuzzy Q-learning. Comparing the SINR distributions before and after learning, it can be seen that the antenna is controlled to reduce the number of users with low SINR.

Algorithm 1 Fuzzy Q-Learning Algorithm

```
for $i = 0$ to $N$ do
    Initialize antenna parameters
    $loop = 1$
    while $loop <= Loop_{max}$ do
        for $M_1$ to $M_k$ do
            Select action $\boldsymbol{a}_k^\dagger$ based on $\varepsilon$-greedy
        end for
        Get the antenna control value:
            $(\Delta\phi_{3dB}, \Delta\phi_{tilt}, \Delta\theta_{3dB}, \Delta\theta_{tilt}) = \sum_k F_k(KPI) \times \boldsymbol{a}_k^\dagger$
        Update antenna parameters:
            $\phi_{3dB} += \Delta\phi_{3dB}$
            $\phi_{tilt} += \Delta\phi_{tilt}$
            $\theta_{3dB} += \Delta\theta_{3dB}$
            $\theta_{tilt} += \Delta\theta_{tilt})$
        $loop++$
        if $KPI < KPI_{bad}$ then
            $reward = -1$
            break
        else if $KPI > KPI_{good}$ then
            $reward = 1$
            break
        end if
        if $loop == Loop_{max}$ then
            $reward = -1$
            break
        end if
        for $M_1$ to $M_k$ do
            Update Q value:
                $Q_t(M_i, \boldsymbol{a}^\dagger{}_i) = Q_t(M_i, \boldsymbol{a}^\dagger{}_i)_t + \varepsilon \times \Delta Q \times e(S_i, \boldsymbol{a}^\dagger{}_i)$
        end for
    end while
end for
```

C. KPI Convergence Characteristics

Fig. 4(a) shows the transition of the KPI associated with learning when the HAPS is rotated by 30 degrees. Fig. 4(b) shows the transition of the KPI during learning when the HAPS moves 5 km to the left. According to Eq. (8), we can obtain the initial KPIs of three cells under HAPS with rotational motion that are $KPI_1 = 0.61, KPI_2 = 0.61, KPI_3 = 0.61$. Also, under the HAPS with translational motion, the initial KPIs are

TABLE I
HAPS PARAMETERS

HAPS Altitude	20 km
HAPS Cover Radius	20 km
Transmit Power	43 dBm
Transmission Frequency	2 GHz
Bandwidth	20 MHz
Propagation Loss	Free Space
SNR of Transmitted Signal	39 dB
Number of HAPS Causing Interference	18
Number of Users in Each Cell	10000

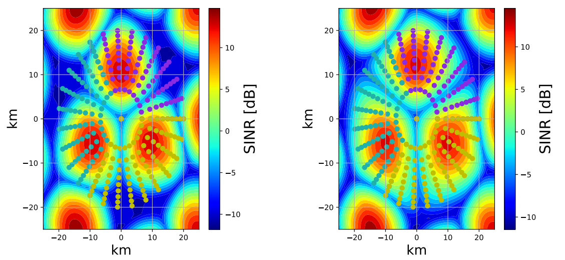
TABLE II
BEHAVIOR / REWARD PARAMETERS

Horizontal HPBW Update Value $\Delta\phi_{3dB}$	1 deg
Horizontal Tilt Update Value $\Delta\phi_{tilt}$	10 deg
Vertical HPBW Update Value $\Delta\theta_{3dB}$	1 deg
Vertical Tilt Update Value $\Delta\theta_{tilt}$	2 deg
Received Power Threshold Γ	-79 dBm
(KPI_{bad}, KPI_{good})	$(0.5, 0.75)$
Maximum Number of Repetitions $Loop_{max}$	40

$KPI_1 = 0.57, KPI_2 = 0.48, KPI_3 = 0.57$. In Fig. 4(a), we can find that after fuzzy Q-learning, the KPIs of three cells is improved to $KPI_1 = 0.78, KPI_2 = 0.8, KPI_3 = 0.82$. According to Fig. 2(a), we can find that each antenna controls the horizontal HPBW. Thus, in Fig. 4(a), after 10 steps, the three KPIs are convergent. In Fig. 4(b), we can find that after fuzzy Q-learning, the KPIs of three cells will be improved to $KPI_1 = 0.87, KPI_2 = 0.83, KPI_3 = 0.87$. According to Fig. 3(a), we can find that there are two antennas that need to control four antenna parameters and another one needs to control the horizontal HPBW. Thus, in Fig. 4(b), KPI_1 and KPI_2 are convergent after 175 steps and 650 steps, respectively. Compared with KPI_1 and KPI_3, KPI_2 converges fast. Thus, the proposed fuzzy Q-learning can effectively improve the KPI performance under both HAPS with rotational and translational motions.

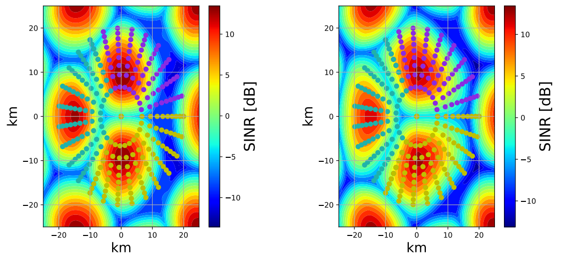
TABLE III
FUZZY Q-LEARNING PARAMETERS

ε	0.01
Learning Rate	0.1
Discount Rate	0.9
The Number of Fuzzy Sets	8



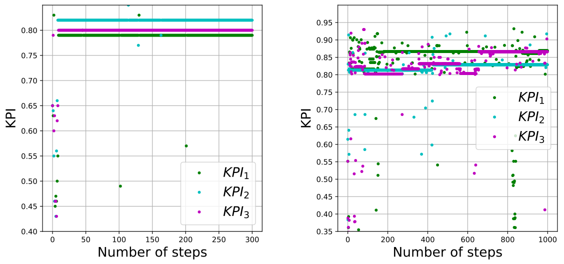
(a) Before training. (b) After fuzzy Q-learning.

Fig. 2. SINR distribution in rotational motion. (a) SINR distribution before fuzzy Q-learning. (b) SINR distribution after fuzzy Q-learning.



(a) Before learning. (b) After fuzzy Q-learning.

Fig. 3. SINR distribution in translational motion. (a) SINR distribution before fuzzy Q-learning. (b) SINR distribution after fuzzy Q-learning.

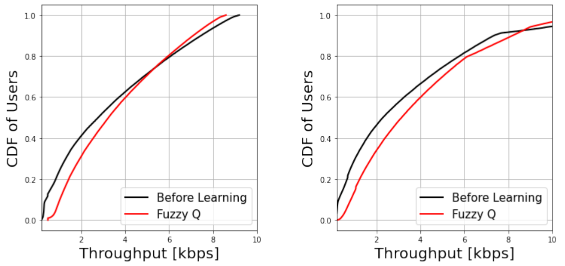


(a) Rotated 30 degrees. (b) Moved 5 km to the left.

Fig. 4. Transition of KPI in fuzzy Q-Learning. (a) In rotational motion . (b) In translational motion.

D. User Throughput

Fig. show the cumulative distribution function (CDF) of the user throughput with and without fuzzy Q-learning for rotational motion and translational motions, respectively. We can find that the number of users whose throughput is smaller than or equal to 2 kbps is reduced by about 10% in both cases. In Fig. 5(b), we can find that the number of users whose throughput is smaller than or equal to 4 kbps also are reduced by about 7%. We can confirm that after antenna control by fuzzy Q-learning,



(a) Rotated 30 degrees. (b) Moved 5 km to the left.

Fig. 5. Cumulative distribution function (CDF) of user throughput. (a) In rotational motion. (b) In translational motion.

the number of low throughput users is decreased in both cases.

V. Conclusion

In this paper, we propose an antenna control for HAPS motion using fuzzy Q-learning to compensate for the cell range shift caused by the wind pressure motion of HAPS. The proposed method uses fuzzy Q-learning because it does not require any prior supervisory data, for antenna control so that the cell range can be corrected even when the HAPS position is unknown. The reward for fuzzy Q-learning is set as KPI, which is the percentage of users whose received power is above a predetermined threshold. We evaluated the proposed method by computer simulation. We showed that the antenna is controlled to make the outage region small when the HAPS is rotated by 30 degrees and when the HAPS has moved -5 km to the left. We also showed that the CDFs of users with low throughput is improved when fuzzy Q-learning is used compared to before learning in both cases.

References

[1] J. Thornton, D. Grace, M. H. Capstick and T. C. Tozer, “Optimizing an array of antennas for cellular coverage from a high altitude platform,” IEEE Trans. on Wireless Communications, vol. 2, no. 3, pp. 484-492, May 2003.

[2] G. M. Djuknic, J. Freidenfelds and Y. Okunev, “Establishing wireless communications services via high-altitude aeronautical platforms: A concept whose time has come?”, IEEE Commun. Mag., vol. 35, pp. 128-135, Sept. 1997.

[3] B. El-Jabu and R. Steele, “Cellular communications using aerial platforms,” IEEE Trans. on Vehicular Technology, vol. 50, no. 3, pp. 686-700, May 2001.

[4] B. El-Jabu and R. Steele, “Effect of positional instability of an aerial platform on its CDMA performance,” IEEE VTS 50th Vehicular Technology Conference, Amsterdam, The Netherlands, pp. 2471-2475, vol.5, 1999.

[5] J. Thornton and D. Grace, “Effect of lateral displacement of a high-altitude platform on cellular interference and handover,” IEEE Trans. on Wireless Communications, vol. 4, no. 4, pp. 1483-1490, July 2005.

[6] K. Hoshino, S. Sudo and Y. Ohta, “A Study on Antenna Beamforming Method Considering Movement of Solar Plane in HAPS System,” IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, pp. 1-5, 2019.

[7] P. G. Sudheesh, M. Mozaffari, M. Magarini, W. Saad and P. Muthuchidambaranathan, ”Sum-Rate Analysis for High Altitude Platform (HAP) Drones With Tethered Balloon Relay,” in IEEE Communications Letters, vol. 22, no. 6, pp. 1240-1243, June 2018, doi: 10.1109/LCOMM.2017.2785847.

[8] S. C. Arum, D. Grace, P. D. Mitchell and M. D. Zakaria, “Beam-Pointing Algorithm for Contiguous High-Altitude Platform Cell Formation for Extended Coverage,” IEEE 90th Vehicular Technology Conference (VTC2019- Fall), Honolulu, HI, USA, pp. 1-5, 2019.

[9] D. Grace, J. Thornton, Guanhua Chen, G. P. White and T. C. Tozer, “Improving the system capacity of broadband services using multiple high altitude platforms,” IEEE Trans. on Wireless Communications, vol. 4, no. 2, pp. 700-709, Mar. 2005.

[10] R. Ali, Y. B. Zikria, S. Garg, A. K. Bashir, M. S. Obaidat and H. S. Kim, ”A Federated Reinforcement Learning Framework for Incumbent Technologies in Beyond 5G Networks,” in IEEE Network, vol. 35, no. 4, pp. 152-159, July/August 2021, doi: 10.1109/MNET.011.2000611.

[11] R. Ali, I. Ashraf, A. K. Bashir and Y. B. Zikria, ”Reinforcement-Learning-Enabled Massive Internet of Things for 6G Wireless Communications,” in IEEE Communications Standards Magazine, vol. 5, no. 2, pp. 126-131, June 2021, doi: 10.1109/MCOMSTD.001.2000055.

[12] C.J.C.H. Watkins,“Learning from Delayed Rewards,” Ph.D. Thesis, Cambridge University, 1989.

[13] R. Razavi, S. Klein and H. Claussen, “A Fuzzy reinforcement learning approach for self-optimization of coverage in LTE networks,” in Bell Labs Technical Journal, vol. 15, no. 3, pp. 153-175, Dec. 2010.

[14] P. Y. Glorennec and L. Jouffe, “Fuzzy Q-learning,” Proceedings of 6th International Fuzzy Systems Conference, Barcelona, vol.2., pp. 659-662, 1997.

[15] Recommendation, I. M. “Minimum Performance Characteristics and Operational Conditions for High Altitude Platform Stations providing IMT-2000 in the Bands 1885-1980 MHz 2010-2025 MHz and 2110-2170 MHz in the Regions 1 and 3 and 1885-1980 MHz and 2110 2160 MHz in Region 2,” International Telecommunications Union, Geneva, Switzerland 2000.

[16] Y. Shibata, N. Kanazawa, M. Konishi, K. Hoshino, Y. Ohta and A. Nagate, “System Design of Gigabit HAPS Mobile Communications,” in IEEE Access, vol. 8, pp. 157995-158007, 2020.

[17] P. Y. Glorennec and L. Jouffe, “Fuzzy Q-learning,” Proc. of 6th International Fuzzy Systems Conference, Barcelona, vol.2, , pp. 659-662, 1997.

[18] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, G.S. Stavrakakis, “Reinforcement learning for energy conservation and comfort in buildings,” Building and Environment, Vol. 42, Issue 7, 2007.

[19] V. François-Lavet, P. Henderson, R. Islam, M.G. Bellemare, J. Pineau, “An Introduction to Deep Reinforcement Learning”, Foundations and Trends in Machine Learning, Vol. 11, no. 3-4, pp 16-17, 2018, doi: 10.1561/2200000071.