Multi-Armed Bandit Learning for Full-Duplex UAV Relay Positioning for Vehicular Communications

Pouya Pourbaba, Samad Ali, K. B. Shashika Manosha, and Nandana Rajatheva
Center for Wireless Communications (CWC), University of Oulu, Oulu, Finland
pouya.pourbaba@student.oulu.fi, {samad.ali, manosha.kapuruhamybadalge, nandana.rajatheva}@oulu.fi

***Abstract*—Utilizing unmanned aerial vehicles (UAVs) in wireless communications can help to improve the capacity of terrestrial networks. In this paper, a novel method is proposed to position a UAV in an optimal location to relay the information from a vehicle to a base station (BS). The proposed method uses predefined locations for the UAV and treats them as the actions for a multi-armed bandit (MAB) framework. The upper confidence bound (UCB) algorithm is used to solve the MAB problem. The results show that this method can identify an optimal location for the UAV to maximize the sum rate of the network.**

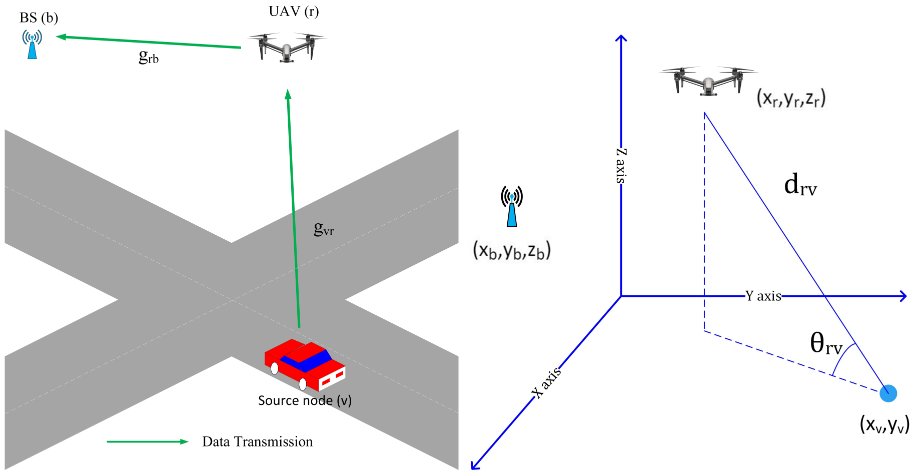
I. Introduction

Next generation of wireless networks will have native support for Internet of Things (IoT) [1], vehicle-to-vehicle (V2V) communications and unmanned aerial vehicles (UAV) [2]. Recently, there has been a surge of research literature on UAV communications, due to their ability in quick movement, low budget deployment, and the large domain of applications they provide [3]. The UAVs can be used as standalone aerial base stations (BSs) or wireless relay nodes to increase the capacity of the network. The UAVs can easily move towards the ground users and establish a reliable and low power transmission link [4]. Moreover, in the case of natural disasters such as earthquakes and floods where the terrestrial wireless BSs are damaged and out of service an aerial BS can quickly be deployed and used in the process of helping the injured ones or finding the missing people. The main challenges of UAV communication are the 3D positioning and path planning and air-to-ground channel modeling.

A statistical channel model for an air-to-ground link is proposed in [5], which defines the path loss of the link from the UAV to the ground user as a function of the elevation angle and the environment characteristics. The air-to-ground channel model is studied in [6] as well where only the line of sight (LoS) and non-line of sight (NLoS) links between the UAV and the ground users are considered. The study on the channel model in [7] concludes that UAVs can be deployed as wireless relay nodes due to their ability to have LoS links and less shadowing compared to the terrestrial wireless links. The studies in [4], [8], focus on the path planning and trajectory of UAVs. In [4], a new approach is proposed for planning an efficient path for multiple UAVs which are used as aerial BSs to collect data from the ground users. UAVs are used in [9] to relay the messages of the onboard units (OBUs) to overcome the problem of smart jamming in vehicular ad-hoc networks (VANETs) by using reinforcement learning. The work in [10] studied the application of reinforcement learning in enabling the UAVs to navigate in unknown environments autonomously. Recently, the use of reinforcement learning to solve wireless communication problems has increased. Some examples of wireless communications related problems which are solved using reinforcement learning are network selection problems of heterogeneous networks, channel sensing, and, energy harvesting [11]. In [12] reinforcement learning is utilized to allocate the sufficient amount of resources to the V2V link which shares the spectrum between a vehicle and the BS. The V2V link transmitter itself is considered as an agent which decides its own transmission power and finds the optimal sub-band to satisfy the V2V constraints. The authors in [13] use reinforcement learning to transmit delay-sensitive data efficiently over a fading channel. Reinforcement learning for UAV to a roadside unit (RSU) relaying is considered in [9]. The UAV acts as the agent and based on the information that it gets from the environment it decides whether or not to relay its message to another RSU. In [14] authors have used reinforcement learning for UAV path planning in a cellular network. The goal of the agents in this work is to maximize the energy efficiency and minimize the latency and the interference generated from the ground users. In [15], use of UAV as a wireless relay to help a vehicular network to communicate with a base station (BS) is considered. In this study, the UAV can operate in one of the predefined locations based on the QoS criteria which is defined as the signal-to-interference-plus-noise ratio (SINR) of the communication links. Multi-armed bandits (MAB) framework from reinforcement learning are used in [16] for optimal allocation of fast uplink grants in the IoT.

The main contribution of this paper is using MAB learning to solve the UAV positioning problem. First, a set of locations that the UAV can operate at are defined, then the MAB framework is utilized to select the best location for the UAV so that the maximum possible sum rate for the network can be achieved. Different MAB algorithms are used to solve the problem and their regret are compared to each other.

The rest of the paper is organized as follows. Section II presents the system model describing the air-to-ground and V2V channel model. In Section III, we introduce the UAV positioning problem and the MAB learning framework. Section IV presents the simulation results, and the conclusions are drawn in section V.



(a) System model

(b) Coordinate system

Figure 1: System model and the coordinate system

II. System Model

Consider a vehicular system in which a vehicle is required to communicate with a BS. We consider that because of some geographical conditions or high shadowing on this communication link, the communication link will be in deep fade and transmissions will fail. Therefore, a UAV is used as a relay to provide connectivity between the vehicle and the BS. The vehicle is denoted by v, the UAV operating as the relay is referred to by the letter r, and the letter b is used to the BS. The relay is assumed to be communicating in a full-duplex (FD) manner where it receives and transmits data simultaneously on the same frequency band. The set $\mathcal{D} = \{v, b\}$ which includes the users on the ground is defined to simplify the mathematical equations. The locations of v and b are given by $(x_i, y_i, z_i), i \in \mathcal{D}$. The predefined locations of the relay are defined in the form of a matrix $\boldsymbol{L}_r$ where each row $l_j \in \mathbb{R}^3$ of $\boldsymbol{L}_r$ represents jth location with three columns for the x, y, and z coordinates. The matrix of locations for the relay can be written as

$$\boldsymbol{L}_r = \begin{bmatrix} x_{r_1} & y_{r_1} & z_{r_1} \\ x_{r_2} & y_{r_2} & z_{r_2} \\ \vdots & \vdots & \vdots \\ x_{r_l} & y_{r_l} & z_{r_l} \end{bmatrix}. \tag{1}$$

Moreover, it is assumed that the vehicle is enabled with the GPS functionality and it sends its location periodically to the UAV where it can be stored in a location table. Figure 1 shows the system model considered for this section.

A. air-to-ground Channel Model

Defining the air-to-ground channel model is realized by using two main groups of communication links namely line of sight (LoS) and non-line of sight (NLoS) links [5]. Based on the environment these occurrence probability of the links would vary. In a suburban area with a low number of buildings with short heights the chance of having a LoS link between a user on the ground and an aerial user is higher compared to an urban area with a higher number of buildings. In addition to the physical characteristics of the environment, the distance between the aerial and the terrestrial user as well as the elevation angle between the two affect the probabilities of having either type of links. according to [17] These probabilities can be expressed as

$$P_{LoS} = \frac{1}{1 + \alpha \exp(-\beta[\frac{180}{\pi}\theta_i - \alpha])}, \tag{2}$$

$$P_{NLoS} = 1 - P_{LoS}, \tag{3}$$

where α and β define the dependency of the probabilities on the physical features of the environment, $\theta_i, i \in \mathcal{D}$ is the elevation angle created between the ground users i and the relay which depend on the Euclidean and vertical distance between them and can be calculated as

$$\theta_i = \frac{180}{\pi} \times \arcsin(\frac{h_{ij}}{d_{ij}}), \tag{4}$$

where the vertical distance between the ground user and the relay is calculated as $h_{ij} = z_r - z_i, i \in \mathcal{D}$ and the Euclidean distance between the two is given by

$$d_i = \sqrt{(x_r - x_i)^2 + (y_r - y_i)^2 + (z_r - z_i)^2}. \tag{5}$$

The path loss for the LoS and NLoS parts of the air-to-ground link depend on the distance of the two users [5]. These path loss values can be calculated as

$$L_{LoS}(dB) = \eta_{LoS}(\frac{4\pi f_c d_{ij}}{c})^{\mu}, \tag{6}$$

$$L_{NLoS}(dB) = \eta_{NLoS}(\frac{4\pi f_c d_{ij}}{c})^{\mu}, \tag{7}$$

where η_{los} and η_{nlos} represent the excessive path loss imposed on each type of propagation link depending on the environment characteristics, c is the speed of light, d_{ij} is the distance between the ground user $i \in \mathcal{D}$ and the jth location of the relay in the sky. Moreover, f_c is the carrier frequency of the transmission and μ is the path loss exponent. By using the probabilities of occurrence and the path losses associated with the LoS and NLoS links the average path loss for the air-to-ground z link can be calculated as

$$L = P_{LoS} \times \eta_{LoS}(\frac{4\pi f_c d_i}{c})^{\mu} + P_{NLoS} \times \eta_{NLoS}(\frac{4\pi f_c d_i}{c})^{\mu}. \tag{8}$$

After the path loss calculation, the gain of the air-to-ground link is computed and used to attain the SNR of that link. The SNR of the link between the vehicle the relay can be expressed as

$$\gamma_{vr} = \frac{p_v g_{vr}}{N_0}, \tag{9}$$

where p_v is the transmit power of v, g_{vr} is the gain of the link, and N_0 is the additive white Gaussian noise. Moreover, the SNR of the communication link between the relay and the BS is

$$\gamma_{rb} = \frac{p_r g_{rb}}{N_0}, \tag{10}$$

where p_r is the transmit power of r, g_{rb} is the gain of the link.

III. Problem formulation and MAB framework

In this section, we formulate the problem of relay positioning as a MAB problem. In our formulation, There is a maximum number of l predefined locations for the relay to accommodate.

A. *Problem formulation*

Consider the matrix of the locations for the relay defined in (1). Let $l_j \in \mathbb{R}^3$ be the ith location that the relay can operate at, where the first, second, and the third element of l_j are x, y, and z coordinates, respectively. Each location for the relay is considered as the arm of a bandit machine. We refer to these arms as the actions and show them by $a \in A = \{a_1, a_2, ..., a_l\}$. The relay will establish two links regardless of the location that it operates at. One of the links is from the v to r and the other one from the r to b. Each of these links will have a rate which determines if the link is proper or not. Therefore, for a given coordinate for the v on the ground we can calculate the value of the rates for each of the locations and store them in a vector. Each element in this vector of rates is considered to be the reward r_t assigned to the locations of the relay.

In order to find the rates associated with each relay location we define two vectors $s_{vr}, \in \mathbb{R}^{1\times l}$ and $s_{rb} \in \mathbb{R}^{1\times l}$ which contain the received powers at each location of r and the received powers in b from each location of the relay. The vector s_{vr} can be expressed as

$$\boldsymbol{s}_{vr} = p_v \boldsymbol{g}_{vr}. \tag{11}$$

where p_v is the transmit power of the v and $\boldsymbol{g}_{vr} \in \mathbb{R}^L$ is channel gain vector for the links between v and each of the predefined locations for r. Similarly, the vector s_{rb} is given as

$$\boldsymbol{s}_{rb} = p_r \boldsymbol{g}_{rb}. \tag{12}$$

where p_r is the transmit power of the r and $\boldsymbol{g}_{rb} \in \mathbb{R}^L$ is vector of the channel gains for the links between each of the predefined locations of r and b.

By using (11) and (12) the SNRs of the links can be written in the form of vectors. The SNR vector for the links between the v and r can be given as

$$\boldsymbol{\gamma}_{vr} = \frac{\boldsymbol{s}_{vr}}{N_0}. \tag{13}$$

Similarly, the SNR vector for the links between the different locations of r and b can be given as

$$\boldsymbol{\gamma}_{rb} = \frac{\boldsymbol{s}_{rb}}{N_0}. \tag{14}$$

Now we can calculate the rate for each link using the calculated SINR and the SNR calculated above

$$\boldsymbol{r}_{sr} = \log_2(\boldsymbol{\gamma}_{sr} + 1), \tag{15}$$

$$\boldsymbol{r}_{rb} = \log_2(\boldsymbol{\gamma}_{rb} + 1), \tag{16}$$

$$\boldsymbol{r}_{t} = \boldsymbol{r}_{sr} + \boldsymbol{r}_{rb}, \tag{17}$$

where $\boldsymbol{r}_{sr}$ is the vector of the rates between the s and all the l locations of the relay. Similarly, the $\boldsymbol{r}_{rb}$ is the vector of the rates for the links between all the possible locations of the relay and b. Moreover, $\boldsymbol{r}_t$ is a vector including the total rate for each location of the relay. The total rate is calculated by adding up the rates of the uplink and the downlink.

The goal of the relay is to find the location with the maximum sum rate. Since the MAB framework is designed to learn how to act in one specific situation, we play this game only for one particular given source node location and find the proper location for the UAV which can provide the best rate.

In the bandit problem, each time an action is selected and the reward for that action is selected from $\boldsymbol{r}_t$. The objective of the relay is to maximize the rewards that it attains by selecting the location which provides the maximum rate for the given coordination of the source node on the ground.

B. *Solution approach*

MABs are a form of reinforcement learning where there is a set of available arms (actions) for an agent to select from. When an arm A_t is selected, it generates a reward R_t from a probability distribution which is not known to the agent. The objective of the agent is to maximize the expected total reward. Since the agent does not know the distribution from which the rewards of each arm are drawn it needs a strategy to compensate for the lack of information to achieve its goal [18]. The age only observes the reward of the arm that it has played. Therefore, the agent can calculate an estimation of the value $Q_t(a)$ for action a before selecting it. The estimation of the action value prior to time t is given by

$$Q_t(a) = \frac{\text{sums of rewards when action } a \text{ is taken prior to } t}{\text{number of times action } a \text{ taken prior to } t}. \tag{18}$$

The agent can play the arm with the highest value for Q_t, which is known as the greedy action selection method.

Algorithm 1 UCB Algorithm

```
Input: $\tau$ (horizon), $\mathcal{A}$ (arms)
Play each arm (action) $a$ once
Observe the rewards of each arm $r_a$
Set $k_a = 1, \forall a \in \mathcal{A}$
Set $\hat{\mu_a} = \frac{r_a}{k_a}$
for $t = |\mathcal{A}|$ to $\tau$ do
    Play arm $\hat{a} = \arg\max_a \left(\hat{\mu_a} + \sqrt{\frac{2\ln(t)}{k_a}}\right)$
    Observe reward $r$
    $r_{\hat{a}} = r_{\hat{a}} + r$
    $k_{\hat{a}} = k_{\hat{a}} + r$
    Update $\hat{\mu_a} = \frac{r_a}{k_a}$
end for
```

This method leads to exploiting the arm with the highest estimation for the action value without exploring any other arm. The greedy method only increases the reward at the current time. However, the objective of the agent is to increase the cumulative reward. Therefore, the agent is required to have a reasonable trade-off between exploitation and exploration.

To overcome the challenge of exploration and exploitation the upper confidence bound (UCB) algorithm can be used. Therefore, the UCB algorithm monitors the potential of the non-greedy actions to be the optimal action instead of exploring the actions in a random fashion. The UCB selects an arm a_t at any given time according to the following equation [18]

$$A_t = \underset{a}{arg\,max}\left[Q_t(a) + \sqrt{\frac{c\ln t}{N_t(a)}}\right], \qquad (19)$$

where c is the degree of exploration, t is the time step, and N_t is the number of times that the arm a has been selected. The square root part in (19) acts as the variance of the estimated value of action a and it shows the level of uncertainty about the action. When an action is selected, the N_t for that action is increased, since this term resides in the denominator of (19), the whole term under the square root decreases. However, when other actions are selected, the value of t in the nominator increases, therefore, the uncertainty increases. This increment in the uncertainty is logarithmic, which means that the value of this increment will get smaller by time. This will guarantee that the actions that have a lower estimate value or that have been selected for a large number of times will not be selected frequently in the future. The UCB algorithm is summarized in Algorithm 1 [19].

One way to measure the performance of an MAB algorithm is by calculating its regret. The regret is the difference in reward of the best possible arm and the reward of the arm that was played. In order to compute the regret we assume that we know the probability distribution from which each action is selected, therefore, we can pick the optimal action by choosing the action with the highest payoff. The regret calculation can be written as [19]

$$L_t = T\mathbb{E}[R_t|A_t = a^*] - \sum_t \mathbb{E}[R_t|A_t = a_t], \qquad (20)$$

Table I: Environment parameters for A2G channel model.

Environment	η_{LoS}	η_{NLoS}	α	β
Suburban	0.1	21	5.0188	0.3511
urban	1	20	9.6101	0.1592
Dense urban	1.6	23	11.9480	0.1359
High rise urban	2.3	34	27.1562	0.1225

Table II: Simulation parameters.

Description	**Value**
Vehicle transmit power (P_v)	0.5 mW
Relay transmit power (P_r)	0.5 mW
Carrier frequency (f_c)	2 MHz
Bandwidth (BW)	1 KHz
Number of the locations of the relay (l)	400
Path loss exponent (n_1)	1.81
Path loss exponent (n_2)	2.85
Noise power spectral density (N_0)	-170 dBm
BS antenna height (h_b)	30 m

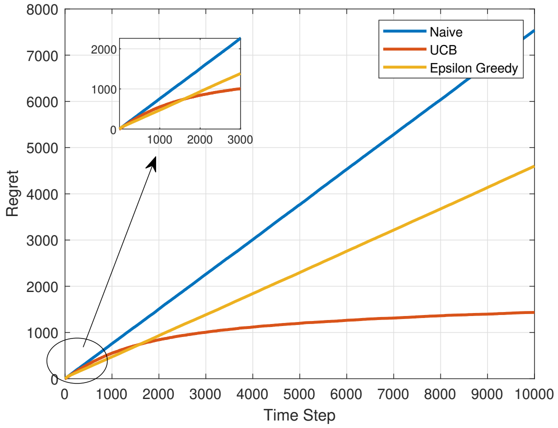


Figure 2: Regret of the MAB framework.

where a^* is the optimal action given the probability distributions of all the actions and it can be found by

$$a^* = \max_{a \in A} \mathbb{E}[R_t|A_t = a_t], \qquad (21)$$

IV. Numerical Results

We consider a cross-road in which the vehicle is located. The pre-defined locations for the relay are considered to be above this cross-road. These locations are in a square area of 226 m × 226 m with 32 m distance between them which make up for 64 locations in total. We consider the communications to be the carrier frequency of 2 GHz and the parameters used to calculate the air-to-ground channel for different environments are given in Table I [5]. Similar to section IV we assume the BS to be at the coordination of (1000,1000) and the location of the source node to be selected randomly along the streets of the length 1 Km.

Fig. 2 shows the regret of the UCB, ϵ-greedy, and a naive allocation policy which chooses the locations randomly. The regret of both naive, ϵ-greedy algorithms increases linearly. However, the regret of the UCB algorithm is logarithmic which is considered optimal in MAB problems.

Fig. 3 shows the cumulative reward for each of the locations. All of the sub-optimal locations would yield low rates and the UCB algorithm should not select them. The highest

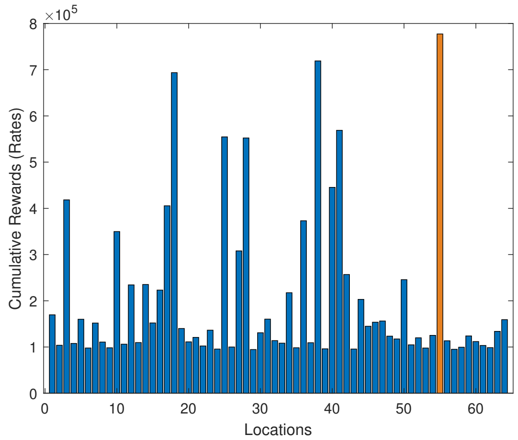


Figure 3: The cumulative reward for each location.

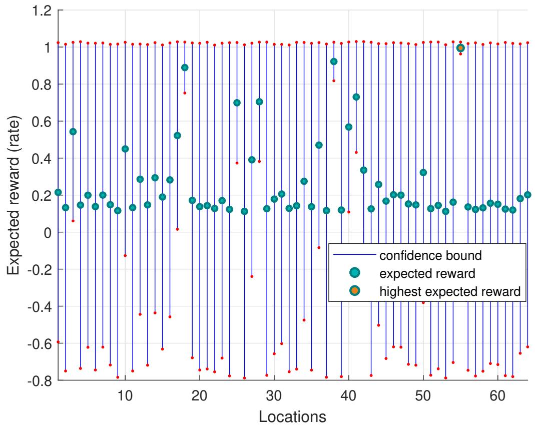


Figure 4: The normalized expected reward of each location and the confidence bound.

cumulative reward which is highlighted in Fig. 3 belongs to the optimal location. The UCB algorithm successfully identifies that arm and selects it more often, which leads to higher total throughput in the system.

Fig. 4 shows the expected values of the rewards for each arm and the confidence interval associated with each one of them calculated by the UCB algorithm. All the sub-optimal arms with a low expected value have larger confidence bound than the arms with a high expected value. This reveals that the UCB algorithm can correctly identify the optimal arm and play that arm frequently. As depicted in Figure 4, the confidence bound at the index 55 has the smallest value, indicating the index of the optimal location for the relay to operate.

V. Conclusion

In this paper, FD UAV relaying is proposed to increase wireless coverage in vehicular communication networks. First, by using a set of predefined locations for the UAV relay, and, also by considering the locations of the vehicle on the ground, we have derived the values of the sum rate for all the possible locations of the UAV. Second, to find the optimal location of the UAV, we have formulated a MAB problem. Finally, by using the UCB algorithm we have solved the problem. Simulation results have shown that by using the proposed method, the algorithm can confidently select the proper location for the UAV to operate.

References

[1] S. Ali, N. Rajatheva, and W. Saad, "Fast uplink grant for machine type communications: Challenges and opportunities," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 97–103, March 2019.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *arXiv preprint arXiv:1803.00680*, 2018.

[4] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile internet of things: Can uavs provide an energy-efficient mobile architecture?" in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[5] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 2898–2904.

[6] J. Holis and P. Pechac, "Elevation dependent shadowing model for mobile communications via high altitude platforms in built-up areas," *IEEE Transactions on Antennas and Propagation*, vol. 56, no. 4, pp. 1078–1084, 2008.

[7] Q. Feng, J. McGeehan, E. K. Tameh, and A. R. Nix, "Path loss models for air-to-ground radio channels in urban environments," in *2006 IEEE 63rd vehicular technology conference*, vol. 6. IEEE, 2006, pp. 2901–2905.

[8] J. Qi, T. Ding, and X. Lu, "Formation trajectory planning and realization of multi-uavs," in *2018 13th World Congress on Intelligent Control and Automation (WCICA)*. IEEE, 2018, pp. 280–285.

[9] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "Uav relay in vanets against smart jamming with reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4087–4097, 2018.

[10] H. X. Pham, H. M. La, D. Feil-Seifer, and L. V. Nguyen, "Autonomous uav navigation using reinforcement learning," *arXiv preprint arXiv:1801.05086*, 2018.

[11] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2017.

[12] H. Ye, Y. G. Li, and B.-H. F. Juang, "Deep reinforcement learning for resource allocation in v2v communications," *IEEE Transactions on Vehicular Technology*, 2019.

[13] N. Mastronarde and M. van der Schaar, "Fast reinforcement learning for energy-efficient wireless communication," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6262–6266, 2011.

[14] U. Challita, W. Saad, and C. Bettstetter, "Deep reinforcement learning for interference-aware path planning of cellular-connected uavs," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7.

[15] P. Pourbaba, K. S. Manosha, S. Ali, and N. Rajatheva, "Full-duplex uav relay positioning for vehicular communications with underlay v2v links," 2019.

[16] S. Ali, A. Ferdowsi, W. Saad, and N. Rajatheva, "Sleeping multi-armed bandits for fast uplink grant allocation in machine type communications," in *2018 IEEE Globecom Workshops (GC Wkshps)*, Dec 2018, pp. 1–6.

[17] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, Dec 2014.

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

[19] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.