Uplink Interference Management in Cellular-Connected UAV Networks Using Multi-Armed Bandit and NOMA

Fatemeh Banaeizadeh*, Michel Barbeau*, Joaquin Garcia-Alfaro†, Venkata Srinivas Kothapalli*,‡, Evangelos Kranakis*

* School of Computer Science, Carleton University, K1S 5B6, Ottawa, Ontario, Canada

† Télécom SudParis, Institut Polytechnique de Paris, 91120, Palaiseau, France

‡ Ericsson Canada, K2K 2V6, Ottawa, Ontario, Canada, Email: kvsrinivas@ieee.org

***Abstract*—Ground users suffer from severe uplink interference originating from high altitude Unmanned Aerial Vehicle (UAV) line-of-sight channels. Using multi-armed bandit, we propose a method aiming to find the best resource block and transmit power level for a UAV dynamically paired with a ground user using Non-Orthogonal Multiple Access (NOMA). It is done according to the UAV's location. It results in mitigating the UAV-uplink interference on its co-channel ground user and maximizing the sum of their data rate in the shared resource block. Performance is evaluated via simulating three exploration-exploitation strategies, namely, epsilon-greedy, upper confidence bound and Thompson sampling.**

***Index Terms*—Unmanned Aerial Vehicle, Uplink Interference, Multi-Armed Bandit, Non-Orthogonal Multiple Access.**

I. Introduction

Apart from the abundant benefits of Unmanned Aerial Vehicles (UAVs), their employment in cellular networks as an aerial user has created challenging issues, such as generating air-to-ground interference, increasing unsuccessful connections for UAVs due to their sequential handovers and optimizing energy consumption [1]. Compared with terrestrial users, UAVs experience less path loss, shadowing and multi-path fading effects due to their high altitude and existence of Line-of-Sight (LoS) links to Base Stations (BSs). A LoS propagation channel means a more reliable communication link between a transmitter and a receiver. These benefits have increased the use of UAVs as aerial BSs or relay nodes in Unmanned Aerial Vehicle (UAV)-cellular communication networks. Aerial BSs provide seamless connectivity, low latency and high coverage services for ground networks [2]. In contrast, integration of UAVs as an aerial user in 5G networks leads to a challenging issue known as air-to-ground interference. That is, existing LoS link between the UAVs and BSs can cause severe interference to Ground User Elements (GUEs) and thus decreases their uplink performance. In such scenario, uplink throughput of GUEs reduces. To mitigate this problem, more uplink resource has to be allocated to ground users. This solution is not effective and also takes GUEs more to transmit their signal [3]. To maintain Quality-of-Service (QoS) for terrestrial users, efficient interference management techniques are required.

We aim to mitigate the air-to-ground interference using two techniques, namely, Multi-Armed Bandit (MAB) [4], [5] and Non-Orthogonal Multiple Access (NOMA), whose capability in solving challenges such as spectral efficiency and inter-cell interference (ICI) is highlighted in [6]. Recent research has demonstrated that NOMA is a potential candidate for improving bottlenecks, such as spectral efficiency and inter-user interference in 5G networks and beyond using power management [7]. This motivates us to use NOMA for UAV-interference management by dynamically optimizing its transmission power. Since UAVs are mobile in nature, a challenging question is how to optimize its transmit power and find the best Resource Block (RB) for the UAV, based on its location and the use of NOMA. In our proposed method, the MAB learning framework is utilized to address this issue and manage the UAV-uplink interference.

The main contributions of this paper are as follows: i) An integrated UAV-Ground User Element (GUE) cellular network is designed. Orthogonal Resource Blocks (RBs) are assigned to GUEs using orthogonal multiple access. In order to suppress UAV-interference, NOMA is utilized by the Base Station (BS) to pair a UAV with a GUE as a function of location. ii) The BS runs the MAB learning approach for each location of UAV to find the best action pair (RB, transmission power level). Therefore, the effect of UAV-interference on its co-channel GUE is minimized, and the sum of the data rate of the UAV and GUE sharing RB is maximized. iii) Reward in the proposed MAB framework is calculated according to the Signal-to-Interference-Noise Ratios (SINRs) of the UAV and GUE. If the SINRs are equal or greater than a threshold, then a reward is the sum of data rates of the UAV and GUE. Otherwise, it is null. The goal is to maximize the cumulative reward during training. Different exploration-exploitation strategies are investigated to find the optimal action pair.

The rest of this paper is organized as follows. Section II presents the literature review. Section III provides the system model. Section IV describes our proposed MAB framework and NOMA technique. Section V presents the simulation results, followed by the conclusion and future work in Section VI.

II. Related Work

There are a limited number of papers which have addressed the UAV-uplink interference. The effect of transmit power and operational altitude on UAV-uplink communications is investigated by Ernest *et al.* in [8]. Experiment results show flying at higher altitudes leads to less probability of outage and more Signal-to-Interference-Noise Ratio (SINR). Air-to-ground interference management using power and trajectory optimization is proposed in [9], [10]. Challita *et al.* in [9] investigate a novel deep reinforcement learning approach, considering UAVs as an intelligent agent to learn their optimal power and trajectory. In [10], Li *et al.* propose an interference cancellation (IC) process done locally by ground BSs in UAV's range without need to share data among adjacent BSs through backhaul. The UAV-interference effect on co-channel GUEs is studied by Mei *et al.* in [2] using a low-complexity and decentralized ICIC technique. The whole cellular network is divided into clusters. The UAV using the received information obtained from the cluster head-BSs selects the optimal cell, RB and transmit power level. Mei *et al.* propose a cooperative interference cancellation (CIC) process, considering a static location for the UAV in [11]. The effect of UAV-interference on its co-channel users is cancelled by local cooperation between co-channel BS and its adjacent BSs.

The capability of NOMA in UAV-interference cancellation on its co-channel GUEs is evaluated by Pang *et al.*and Mei and Zhang in [12], [13], considering a static location for the UAV. Most prior works control the UAV-interference by considering the UAV staying in a fix location, while interference management dynamically through power or resource allocation optimization is still an open research area [14]. The main aim of this paper is to address this problem using NOMA and MAB methods.

III. System Model

We consider a single-cell network consisting of n GUEs and a UAV which are served by a ground BS. GUEs and the UAV are equipped with a single antenna. The locations of the GUEs are fixed. Orthogonal RBs are assigned to GUEs. The UAV moves randomly. The goal is to pair the UAV with a GUE in each location using MAB so as to minimize the UAV- uplink interference. NOMA is used for pairing the UAV with a GUE in a RB. The UAV-BS uplink channel in location k is modeled considering only the free-space path loss and depending only on distance [15]:

$$\mathrm{PL}_k = 20\log_{10}4\pi + 20\log_{10}d_k - 20\log_{10}\lambda \text{ dB} \tag{1}$$

where λ is wavelength (m) and d_k is the Euclidean distance (m) between the UAV and BS in location k. Therefore, in linear form the UAV-BS channel power gain at location k is $h_{UAV_k} = 10^{PL_k/10}$. The channel between GUE_j and BS considering path loss and shadowing, modeled as follows [16]:

$$\mathrm{PL}_j = PL_0 + 10\alpha\log_{10}d_j + X_\sigma \text{ dB} \tag{2}$$

where PL_0 is the path loss at reference distance one meter, d_j is the GUE$_j$-BS distance (m), PL_j is the path loss at distance d_j, α is the path loss exponent and X_σ denotes the shadowing effect (variation of received signal power) that is modeled using a Gaussian distribution with zero mean and variance one. The shadowing component captures the stochastic of the channel. In linear form, the GUE$_j$-BS channel power gain is $h_{GUE_j} = 10^{PL_j/10}$.

IV. Interference Management with MAB

NOMA is able to cancel mutual interference by allocating unique transmit power levels to users. We propose a method to mitigate the UAV-uplink interference on GUEs by dynamically finding the best RB and optimal UAV-power level according to location using MAB learning. In a NOMA system, for an uplink there are n users with transmit power levels $p_1, p_2, \ldots, p_n$ and respective channel gains $h_1, h_2, \ldots, h_n$. They use the same RB to communicate with the BS. The BS receives the superposition of all user signals. The BS employs the Successive Interference Cancellation (SIC) process to decode the user signals following their channel gain strength, from the user with the highest channel gain to the user with lowest channel gain. That is, the BS first decodes and subtracts the signal of the strongest user. It continues the SIC process sequentially by decoding and subtracting the signal from the second strongest user to the weakest user. For a given user, the signals from the weaker users are considered to be interference [14], [17]. Let us assume that channel gains are sorted from highest to lowest in the order from $h_1, h_2, \ldots, h_n$. Therefore, the SINR of user i on the shared RB is denoted by:

$$SINR_i = \frac{p_i|h_i|^2}{\sum_{j=i+1}^{n} p_j|h_j|^2 + N_0 B_i} \tag{3}$$

where N_0 is noise power density, B is the total bandwidth and $B_i = B/N$ is the bandwidth of user i on the shared RB (N is the number of RBs).

Then, the achievable data-rate of user i is expressed by:

$$R_i = B_i \log_2(1 + SINR_i) \text{ bps} \tag{4}$$

As a result, the uplink sum data-rate of all the users in the shared RB is:

$$R_T = \sum_{i=1}^{n} B_i \log_2(1 + SINR_i) \text{ bps} \tag{5}$$

In our proposed method, the problem of resource allocation to the UAV for minimizing the interference is solved by maximizing R_T while the following conditions are met:

$$\begin{aligned} C_1: &\quad SINR_i \geq SINR_{threshold} \quad \forall i \in n \\ C_2: &\quad 0 \leq P_i \leq P_{max} \quad \forall i \in n \\ C_3: &\quad R_i \geq R_{min} \quad \forall i \in n \end{aligned} \tag{6}$$

where C_1 states that the SINR of every user i on the shared RB has to be more than or equal to a threshold. C_2 is the range of transmit power of every user i, which should not be larger than P_{max}. C_3 indicates that the data rate of every user i has

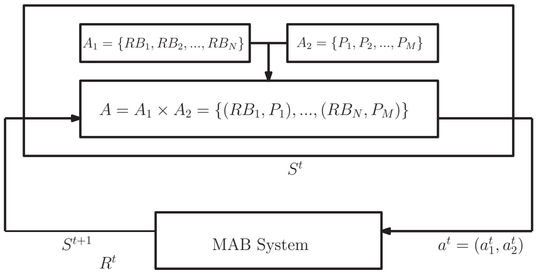

$A_1 = \{RB_1, RB_2, ..., RB_N\}$
$A_2 = \{P_1, P_2, ..., P_M\}$
$A = A_1 \times A_2 = \{(RB_1, P_1), ..., (RB_N, P_M)\}$
S^t
S^{t+1}
R^t
MAB System
$a^t = (a_1^t, a_2^t)$

Fig. 1. MAB system model. A_1: available RBs for the UAV, A_2: available transmit power levels for the UAV, A: all action pairs, and $a^t = (a_1^t, a_2^t)$: selected action pair at time t.

to be more than or equal to a minimum required data rate. It should be noted that in the proposed method, all RBs have the same bandwidth size. Therefore, meeting C_1 provides the minimum data rate for users. While if the RBs have different bandwidth sizes, the role of C_3 becomes more important and both C_1 and C_3 should be checked, i.e., using only C_1 does not guarantee the minimum required data rate for users. MAB is known as one-step reinforcement learning [4], [5], see Fig. 1. There is a set of actions. After pulling each action, the agent obtains a random reward (R_T), due to environment stochastic nature. The reward probability distribution is a priori unknown. The agent's goal is to maximize the sum of rewards obtained through choosing actions sequentially. It aims at finding the action with the highest total payout (R_T). The process of finding the best action is through exploration-exploitation. Such strategies guide the agent to achieve trade-offs between exploiting the action with the highest total reward and exploring other actions. Algorithm 1 summarizes our method. The BS is the agent. Action space A consists of two sub-action spaces, A_1 and A_2. A_1 contains all available RBs, $RB_1, RB_2, \ldots, RB_N$. A_2 contains all UAV-transmit power levels, $P_1, P_2, \ldots, P_M$, where P_M is maximum transmit power of the UAV. At each step, the agent takes an action pair, $a^t = (a_1^t, a_2^t)$, where $a_1^t \in A_1$ and $a_2^t \in A_2$. They determine RB assignment and transmit power level for the UAV. We discretize the transmit power of the UAV into M levels. The set of all available action pairs is equal to:

$$\begin{aligned} A = A_1 \times A_2 = &\{(RB_1, P_1), ..., (RB_1, P_M), \\ &(RB_2, P_1), ..., (RB_2, P_M), ..., \\ &(RB_N, P_1), ..., (RB_N, P_M)\} \end{aligned} \tag{7}$$

Reward is considered as the sum of the users' data rates in the shared RB, when the SINR of every user is greater than a threshold. Otherwise, the reward is null. The goal is to maximize the cumulative reward during the agent's training. The training process for finding the optimal action pair is done using three exploration-exploitation strategies: ϵ-greedy, Upper Confidence Bound (UCB) and Thompson Sampling (TS). In ϵ-greedy, the agent exploits the action with the highest average reward ($Q(a)$) with probability $1 - \epsilon$ and explores a random action with probability ϵ. The average reward for actions is calculated as follows:

$$Q(a_i) = \frac{\sum_{t=1}^{T} r_{t,i}}{K_i} \tag{8}$$

where $r_{t,i}$ is the reward for action a_i at time t. K_i is the number of times the action a_i is taken. Then, the action with the highest $Q(a)$ value is chosen as the optimal one:

$$a^* = \arg\max_{a \in A} Q(a) \tag{9}$$

In ϵ-greedy, all non-best actions are explored with the same probability while some of them should be given more weight in the exploration. So, It leads to uncertainty in optimal action selection. To remove the uncertainty, we also simulate UCB and TS [4]. UCB gives more chance to actions that have been taken less often. That is, it calculates UCB for every action a_i as follows:

$$UCB(a_i) = Q(a_i) + \sqrt{\frac{c \log(t)}{K(a_i)}} \tag{10}$$

where $Q(a_i)$ is the average reward of action a_i, c is the exploration degree, which is equal to 2.5 in this work, t is the total number of rounds and $K(a_i)$ is the number of times action a_i has been selected. MAB selects the action with the highest UCB as the best at each round:

$$a^* = \arg\max_{a \in A} UCB(a) \tag{11}$$

TS is a Bayesian approach that models a probability distribution from actual outputs of each action. The process is started with an initial estimate refined during training. We implement TS using the most common scenario, i.e., we assume that each action's outputs normally are distributed with mean μ_a and variance σ_a^2 (both unknown). To model the true distribution of each action's rewards, the initial estimate is modeled using the normal-inverse-gamma distribution as a prior conjugate. The normal distribution is used for the likelihood [18]–[20]. The conjugate prior and normal likelihood are defined by the following equations:

$$\begin{aligned} &\mathcal{NIG}(\mu_a, \sigma_a^2 \mid m_0, \nu_0, \alpha_0, \beta_0) = \\ &\mathcal{N}(\mu_a \mid m_0, \frac{\sigma_a^2}{\nu_0})\mathcal{IG}(\sigma_a^2 \mid \alpha_0, \beta_0) = \sqrt{\frac{\nu_0}{2\pi\sigma_a^2}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{1}{\sigma_a^2}\right)^{\alpha_0+1} \\ &\qquad \exp\left(-\frac{2\beta_0 + \nu_0 (\mu_a - m_0)^2}{2\sigma_a^2}\right) \quad \forall a \in A \end{aligned} \tag{12}$$

$$\begin{aligned} &P\left(\{X_1, X_2, \ldots, X_t\} \mid \mu_a, \sigma_a^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma_a^2}}\right)^{K_{a,t}} \\ &\qquad \exp\left(-\frac{1}{2\sigma_a^2} \sum_{L=1}^{t} (X_L - \mu_a)^2\right) = \\ &\left(\frac{1}{\sqrt{2\pi\sigma_a^2}}\right)^{K_{a,t}} \exp\left(-\frac{1}{2\sigma_a^2}\left[K_{a,t}(\bar{x}_{a,t} - \mu_a)^2 + s_{t,a}\right]\right) \end{aligned} \tag{13}$$

where m_0, ν_0, α_0 and β_0 are hyper-parameters of the conjugate prior and defined as the prior mean, count, shape and scale parameters, respectively. $\{X_1, X_2, \ldots, X_t\}$ are rewards obtained by taking an action a from starting experiment to time t and $K_{a,t}$ is the number of times action a is taken. The posterior (true) distribution is defined as follows:

$$\begin{gathered} P\left(\mu_a, \sigma_a^2 \mid \{X_1, \ldots, X_t\}\right) \propto P\left(\{X_1, \ldots, X_t\} \mid \mu_a, \sigma_a^2\right) \\ P\left(\mu_a, \sigma_a^2 \mid m_0, \nu_0, \alpha_0, \beta_0\right) \quad \forall a \in A \end{gathered}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma_a^2}}\right)^{K_{a,t}} \exp\left(-\frac{1}{2\sigma_a^2}\left[K_{a,t}\left(\bar{x}_{a,t} - \mu_a\right)^2 + s_{t,a}\right]\right)$$
$$\sqrt{\frac{\nu_0}{2\pi\sigma_a^2}} \frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)} \left(\frac{1}{\sigma_a^2}\right)^{\alpha_0+1} \exp\left(-\frac{2\beta_0 + \nu_0\left(\mu_a - m_0\right)^2}{2\sigma_a^2}\right) \tag{14}$$

The posterior distribution also follows a normal-inverse-gamma distribution, like the initial estimate. To achieve the posterior distribution of actions' output, hyper-parameters of initial estimate must be updated after taking an action using the following equations (cf. [19] for the proofs and details):

$$m_{a,t} = \frac{\nu_0 m_0 + K_{a,t}\bar{x}_{a,t}}{\nu_0 + K_{a,t}} \qquad \forall a \in A \tag{15}$$

$$\nu_{a,t} = \nu_0 + K_{a,t} \qquad \forall a \in A \tag{16}$$

$$\alpha_{a,t} = \alpha_0 + \frac{1}{2}K_{a,t} \qquad \forall a \in A \tag{17}$$

$$\beta_{a,t} = \beta_0 + \frac{1}{2}s_{a,t} + \frac{K_{a,t}\nu_0\left(\bar{x}_{a,t} - m_0\right)^2}{2\left(K_{a,t} + \nu_0\right)} \quad \forall a \in A \tag{18}$$

where $\bar{x}_{a,t}$ and $s_{a,t}$ are the average sampled rewards and the sum of squares for action a at time t, which are denoted by:

$$\bar{x}_{a,t} = \frac{1}{K_{a,t}} \sum_{L=1}^{t} X_{a,L} \qquad \forall a \in A \tag{19}$$

$$s_{a,t} = \sum_{L=1}^{t} \left(X_{a,L} - \bar{x}_{a,t}\right)^2 \quad \forall a \in A \tag{20}$$

To implement TS and take an action at time t, two steps are followed. i) Draw a sample of variance:

$$\hat{\sigma}_{a,t}^2 \sim \mathcal{IG}\left(\frac{1}{2}K_{a,t} + \alpha_0, \beta_{a,t}\right) \quad \forall a \in A \tag{21}$$

ii) Draw a sample of normal distribution:

$$\hat{\mu_{a,t}} \sim \mathcal{N}\left(m_{a,t}, \frac{\hat{\sigma}_{a,t}^2}{\nu_0 + K_{a,t}}\right) \quad \forall a \in A \tag{22}$$

where $\hat{\sigma}_{a,t}^2$ and $\hat{\mu_{a,t}}$ are sampled values from the inverse-gamma and normal distributions. After sampling, the action with the highest $\hat{\mu_{a,t}}$ is taken as the best action:

$$a_t^* = \underset{a \in \mathcal{A}}{\arg\max} \mathbb{E}\left[X_{a,t} \mid \mu_{a,t}, \sigma_{a,t}^2\right] = \underset{a \in \mathcal{A}}{\arg\max} \mu_{a,t} \tag{23}$$

where $X_{a,t}$ is the expected reward of action i at time t. Then, equations (15)–(18) are updated for chosen action. The performance associated to the exploration-exploitation strategies is evaluated via the regret, i.e., the difference between the total reward obtained from always choosing the best action and the total reward obtained from playing the selected actions, until round t. The regret is calculated as follows:

$$R = \sum_t \mathbb{E}[R_t | A_t = a^*] - \sum_t \mathbb{E}[R_t | A_t = a] \tag{24}$$

where a^* is the optimal action, a is the selected action, R_t is the expected reward from playing the best action or playing the selected action. In MAB learning, the main goal is to minimize the regret in order to maximize the obtained reward.

Algorithm 1 *ϵ-greedy, UCB and TS learning*

```
Input: Action spaces $(A_1, A_2)$
$\epsilon$-greedy strategy
$Counter(a_1^t, a_2^t) \leftarrow 0$
$SumReward(a_1^t, a_2^t) \leftarrow 0$
$AvgReward(a_1^t, a_2^t) \leftarrow 0$
for $j \leftarrow 1$ to $NumIteration$ do
    $r \leftarrow$ Generate a random number
    if $r < \epsilon$ then
        Choose a random action
    else
        Choosing the best action
    $Counter(a_1^t, a_2^t) \leftarrow Counter(a_1^t, a_2^t) + 1$
    if $SINR_{UAV}$ & $SINR_{GUE} \geq Threshold$ then
        $SumReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t) + reward$
        $AvgReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t)/Counter(a_1^t, a_2^t)$
    else
        $SumReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t) + 0$
        $AvgReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t)/Counter(a_1^t, a_2^t)$
UCB strategy
Play each action pair once
Update their $AvgReward$
for $j \leftarrow 1$ to $NumIteration$ do
    Calculate UCB for each action pair
    Select the action pair with maximum UCB
    $Counter(a_1^t, a_2^t) \leftarrow Counter(a_1^t, a_2^t) + 1$
    if $SINR_{UAV}$ & $SINR_{GUE} \geq Threshold$ then
        $SumReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t) + reward$
        $AvgReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t)/Counter(a_1^t, a_2^t)$
    else
        $SumReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t) + 0$
        $AvgReward(a_1^t, a_2^t) \leftarrow SumReward(a_1^t, a_2^t)/Counter(a_1^t, a_2^t)$
TS strategy
Initialization:
$\alpha_0 = 1/2, \beta_0 = 1, \nu_0 = 2 \quad \forall a \in A$
Play each action pair $\nu_0$ times
for $j \leftarrow 1$ to $NumIteration$ do
    for $t \leftarrow 1$ to $NumActions$ do
        Draw sample using Equation (21)
        Draw sample using Equation (22)
    Play action pair with the highest sample value $(\hat{\mu_{a,t}})$
    Update $m_{a,t}$, $\nu_{a,t}$, $\alpha_{a,t}$ and $\beta_{a,t}$ for chosen action pair
```

V. Simulation Results

A simulation is run in Matlab with parameters listed in Table I [9], [12]. We designed a single-cell connected UAV network of size 3000 m by 3000 m by 200 m, see Fig. 2. There are a ground BS at the centre of the cell, ten GUEs (the evaluation for different number of users and distributions is considered as future steps) and a UAV. The UAV follows a random walk mobility model. For simplicity, the proposed model is simulated for one UAV-location with altitude of 200 m. The results can be applied to other locations. There are ten RBs that are assigned to GUEs by the BS. It is assumed that there is no intra-cell interference between GUEs, due to the orthogonality of the RBs. The GUEs only get interference from the UAV in their uplink when they use the same RB.

TABLE I
SIMULATION PARAMETERS

Parameters	Value
Number of RBs	10
Number of GUEs	10
BS altitude	10 m
UAV altitude	200 m
SINR threshold for UAV	1 dB
SINR threshold for GUE	1 dB
α	3
Bandwidth	20 MHz
Carrier frequency (f_c)	2 GHz
GUE transmit power	1000 mW (1 W)
Maximum UAV transmit power	100 mW
N_0	-174 dBm/Hz
Number of iterations	10000
ϵ	[0,1]

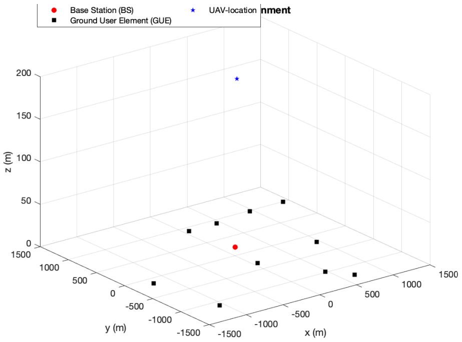


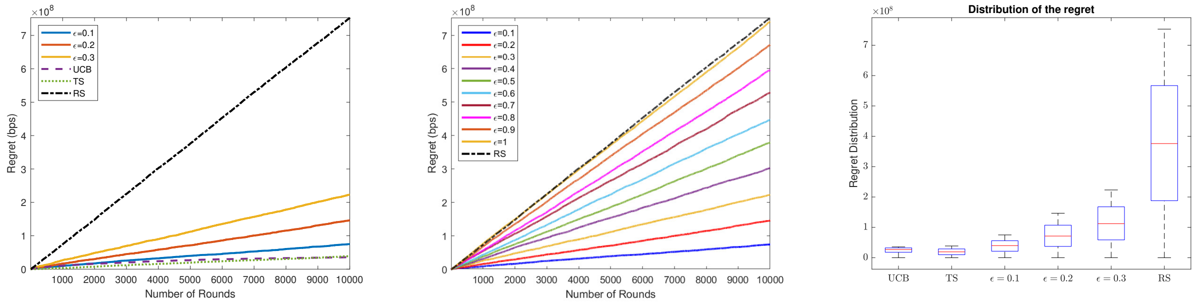
Fig. 2. Simulation Scenario with a Base station (●) at the centre of cell, ten ground users (■) and a UAV ($\star$) at the altitude 200 m.

At each location of the UAV, the agent (BS) uses MAB to determine the best action pair. The agent aims to maximize the sum of the data-rate for the UAV and paired GUE, while at the same time minimizing the uplink interference of the UAV on its co-channel GUE. Since the action space for UAV-transmit power (A_2) is considered discrete, we implement the simulation for five different levels ($60, 70, 80, 90$ and 100 mW). Thus, there are 50 action pairs (10 RBs×5 power levels) for the agent at each location of the UAV.

The performance of the proposed method is evaluated in terms of the regret. Fig. 3 compares the regret value for the aforementioned strategies. The random selection method is considered to be the base for comparison. It selects an action randomly. As it is shown in Fig. 3(a), the regret for epsilon strategy and random selection increases linearly, while UCB and TS have the lowest regret. Therefore, it shows the two latter methods have more accuracy in choosing the best action. Fig. 3(b) shows the regret for different ϵ-values. $\epsilon = 0.1$ outperforms other ϵ-values. We can see that increasing ϵ-values results in more regret due to more exploration. Thus, the performance of $\epsilon = 1$ should be similar to random selection strategy because it only explores actions without exploitation. For more analysis, we show the regret distribution for all four strategies using boxplot (i.e., a statistical way to describe scattering of numerical values in a group). Fig.3(c) indicates the distribution only for one time experiment. As shown, UCB and TS have less regret distribution in comparison to epsilon-greedy and random selection. Besides, UCB has almost better performance than TS due to more exploration of actions with high probability of being optimal action. Fig.4 and Fig.5 represent another visualization of regret. They show the regret variations for the first 100 iterations. Each iteration is repeated 50 times. From Fig.4, it can be realized that the regret for all ϵ-values enhances linearly. But, $\epsilon = 0.1$ has less increase in terms of regret because it does less exploration and searches other no optimal actions with probability only 10% while this probability is more for other epsilon values. The evaluation of results in Fig.5 shows the increasing trend in regret for UCB, TS and random selection. Thus, it can prove learning of the agent during the training process. The increasing trend for UCB illustrates that at the beginning of training process, other actions have large confidence interval. Therefore, uncertainty in selecting actions has to be removed in order to produce an accurate estimate of average reward interval of each action. As a result, UCB searches non-best actions more which results in more regret. For TS, the initial estimate of actions' distributions is explored by the agent to learn the true distribution of actions' outputs. Thus, it makes an increasing trend for regret in initial iterations.

VI. CONCLUSION

UAV-uplink interference degrades the quality of GUE-uplink communications. Learning approaches are an outstanding solution to address this problem in cellular-connected UAV networks. In this work, we utilized the NOMA and multi-armed bandit (MAB) techniques to mitigate UAV-uplink interference by choosing the best action pair for the UAV, according to its location. The simulation results showed the efficiency of the proposed method in alleviating interference of the UAV on its co-channel ground user (GUE). In future steps, the performance of the proposed method in terms of different mobility models with more UAVs will be evaluated. Besides, we plan to improve the limitations of the proposed method, which are as follows. 1) Considering the transmit power sub-action space (A_2) as discrete is challenging in networks with many users. It exponentially increases the size of the action space and computational overhead on the agent. Also, it does not evaluate all values for the transmit power, which might be more optimal. 2) Using only the MAB framework, the agent does not have the capability to learn the environment dynamics and guide the UAV to a location that generates less interference. Future perspectives include the use of continuous action spaces and further analysis of the transmit power. In the end, we plan to analyze the computational complexity of the proposed method and compare it with other UAV-interference management methods.



(a) (b) (c)

Fig. 3. (a): regret over number of rounds , (b): regret for different ϵ-values, (c): regret distribution for one experiment of all strategies.

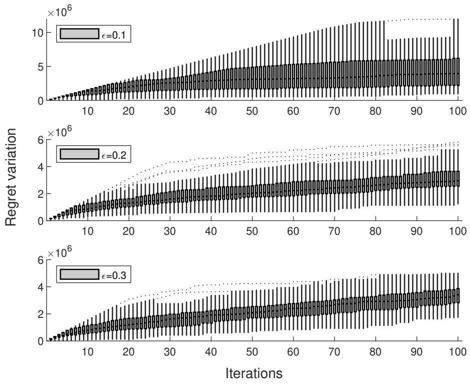


Fig. 4. Regret variation vs. 100 iterations.

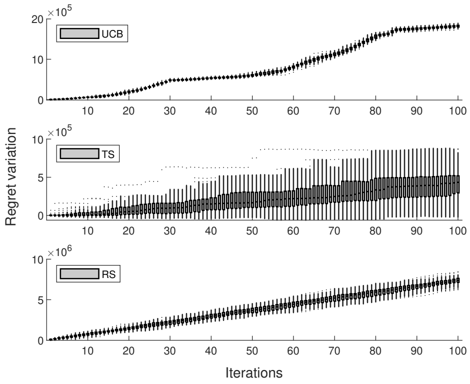


Fig. 5. Regret variation vs. 100 iterations.

Acknowledgements — We acknowledge financial support from Ericsson, Mitacs and NSERC (Natural Sciences and Engineering Research Council) of Canada.

References

[1] D. Mishra and E. Natalizio, "A Survey on Cellular-connected UAVs: Design Challenges, Enabling 5G/B5G Innovations, and Experimental Advancements," *Computer Networks*, vol. 182, pp. 1–32, 2020.

[2] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: Uplink association, power control and interference coordination," *IEEE Transactions on wireless communications*, vol. 18, no. 11, pp. 5380–5393, 2019.

[3] S. D. Muruganathan, X. Lin, H. Maattanen, Z. Zou, W. A. Hapsari, and S. Yasukawa, "An overview of 3GPP release-15 study on enhanced LTE support for connected drones," *arXiv:1805.00826*, 2018.

[4] S. Ravichandiran, *Deep Reinforcement Learning with Python*. Packt Publishing, 2020.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[6] M.-J. Youssef, V. V. Veeravalli, J. Farah, C. A. Nour, and C. Douillard, "Resource allocation in NOMA-based self-organizing networks using stochastic multi-armed bandits," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6003–6017, 2021.

[7] A. Akbar, S. Jangsher, and F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques," *Computer Networks*, vol. 190, p. 107950, 2021.

[8] T. Z. H. Ernest, A. S. Madhukumar, R. P. Sirigina, and A. K. Krishna, "Impact of cellular interference on uplink UAV communications," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–5, 2020.

[9] U. Challita, W. Saad, and C. Bettstetter, "Interference Management for Cellular-Connected UAVs: A Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, 2019.

[10] P. Li, L. Xie, J. Yao, and J. Xu, "Cellular-connected UAV with adaptive air-to-ground interference cancellation and trajectory optimization," *IEEE Communications Letters*, vol. 26, no. 6, pp. 1368–1372, 2022.

[11] W. Mei and R. Zhang, "Uplink cooperative interference cancellation for cellular-connected UAV: A quantize-and-forward approach," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1567–1571, 2020.

[12] X. Pang, G. Gui, N. Zhao, W. Zhang, Y. Chen, Z. Ding, and F. Adachi, "Uplink precoding optimization for NOMA cellular-connected UAV networks," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 1271–1283, 2020.

[13] W. Mei and R. Zhang, "Uplink cooperative NOMA for cellular-connected UAV," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 644–656, 2019.

[14] W. K. New, C. Y. Leow, K. Navaie, Y. Sun, and Z. Ding, "Application of NOMA for cellular-connected UAVs: Opportunities and challenges," *Science China Information Sciences*, vol. 64, no. 4, pp. 1–14, 2021.

[15] W. Saad, M. Bennis, M. Mozaffari, and X. Lin, *Wireless Communications and Networking for Unmanned Aerial Vehicles*. Cambridge University Press, 2020.

[16] J. B. Andersen, T. S. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *IEEE Communications Magazine*, vol. 33, no. 1, pp. 42–49, 1995.

[17] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 185–191, 2017.

[18] J. Honda and A. Takemura, "Optimality of Thompson sampling for gaussian bandits depends on priors," in *Artificial Intelligence and Statistics*, pp. 375–383, 2014.

[19] K. P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," *Technical report*, 2007.

[20] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Harvard University, 1961.