Multi-Agent Learning Approach for UAVs Enabled Wireless Networks

Lorenzo De Simone, Yongxu Zhu, Wenchao Xia, Tasos Dagiuklas, Kai Kit Wong

***Abstract*—The unmanned aerial vehicle (UAV) technology provides a potential solution to scalable wireless edge networks. This paper uses two UAVs, with accelerated motions and fixed altitudes, to realize a wireless edge network, where one UAV forwards the downlink signal to user terminals (UTs) distributed over an area where another UAV collects uplink data. Both downlink and uplink transmissions consider the active user probability and the queue structure as well as the hovering times of UAVs. Specifically, we develop a novel joint Q-Learning multi-agent (JQ-LMA) algorithm to maximize the overall energy efficiency of the edge networks, through optimizing the UAVs trajectories, transmit powers, and the resistant distance between UAVs. The simulation results demonstrate that the proposed algorithm achieves much higher energy efficiency than other benchmark schemes.**

***Index Terms*—UAV swarm, energy efficiency, trajectory optimization, multi-agent reinforcement learning, queue theory.**

I. Introduction

Mobile devices and data traffic in the edge networks [1] will grow exponentially over the next few years. To meet these demands and provide the holographic coverage for the future edge network, it is necessary to develop dynamic, scalable, and self-organized networks. In the last decade, the technology related to autonomous drones, also called Unmanned Aerial Vehicles (UAVs), has been rapidly developed. It is commonly regarded as an effective technology in future wireless networks. However, due to the physical limitations of the UAVs [2], such as short battery life, it would be difficult to rely on a single UAV to complete complex tasks. Hence, in many applications, multiple UAVs are required to cooperate with each other to improve the energy efficiency of wireless networks [3]. The energy efficiency of UAVs enabled wireless networks has been addressed in various works. In [4], a UAV-enabled wireless communication system with energy harvesting has been investigated, where the total energy consumption of the UAV is minimized while satisfying the minimal data transmission requests of the users. In [5], the energy efficiency is maximized by optimally planning the trajectory of the UAV collecting sensor data from devices scattered around.

RL has been widely used in the field of UAVs too. In [6], an adaptive federated RL-based jamming attack defense strategy has been developed. In [7], a deep RL algorithm has been used to compute the optimal trajectories. In [8], a RL algorithm has been proposed to control the transmission power and to manage interference.

UAV-enabled technology in wireless communications has been widely used. In [9], an analytical framework has been developed to evaluate the performance of a finite, three dimensional (3D) UAV network in the presence of interference. The framework is based on stochastic geometry tools and uses the Binomial Point Process (BPP) to model the spatial distribution of the UAVs. In [10], radio interference is analyzed using stochastic geometry and 3D grid-based designs of a primary exclusive region is presented for a UAV network with spectrum sharing.

The existing literature only addresses the downlink [4] or uplink [11] of UAV-based wireless networks, but not both, and also without a realistic energy consumption model and a multi-UAV system. In order to fill this gap, this paper studies a multi-UAV enabled wireless communication system, where the uplink and the downlink are jointly optimized for energy efficiency, considering both moving and hovering energy consumption models, with the final aim to maximize the system energy efficiency. To the best of our knowledge, this has not been investigated in the existing literature. The contribution of this paper lies in several directions.

- **Multiple modes collaborative edge network.** We design a multi-agent-based edge network, with two UAVs in charge of downlink and uplink transmission, where the UAVs are with accelerating. Two interference sources are considered.
- **Service time managed by queue theory.** We have considered an M/M/1 queue formed by two queues in both downlink and uplink for existing UTs and newly coming UTs.
- **Dynamic time scheme and corresponding energy consumption.** We consider a dynamic time scheme to use a realistic fly and stop scheme for both UAVs.
- **Multi-agent synergy approach for energy efficiency maximization.** We develop a novel RL-based JQ-LMA algorithm to maximize the energy efficiency of the system, with a dynamic learning rate and dynamic probability of action choice.

II. System Model

This paper considers a two UAV-enabled BSs and multiple UTs time-division duplex (TDD) edge network as in Fig. 1. One UAV, $\mathrm{U_A}$, is responsible for downlink transmission while

L. De Simone, Y. Zhu and T. Dagiuklas are with the Division of Computer Science and Informatics, London South Bank University, London, UK (Email: {desimol2, yongxu.zhu, tdagiuklas}@lsbu.ac.uk).

W. Xia is with the Department of Wireless Communication Key Lab of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: xiawenchao@njupt.edu.cn).

K. K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, United Kingdom (e-mail: kai-kit.wong@ucl.ac.uk).

another UAV, $\mathrm{U_B}$, is responsible for data collection from the UTs in the uplink. Both are equipped with single-antenna and elevated at fixed altitude h_{A} and h_{B}, respectively. Note that all UTs are also with single-antenna and are geographically distributed with Poisson Point Process (PPP) $\Phi_{\mathtt{G}}$ with density $\lambda_{\mathtt{G}}$. Therefore, the number of active UTs will be $\widetilde{\lambda}_{\mathtt{G}} = p_{\mathtt{G}}\lambda_{\mathtt{G}}$, if the active probability for UTs is $p_{\mathtt{G}}$. We assume that both $\mathrm{U_A}$ and $\mathrm{U_B}$ travel between the Macro Base Station (MBS) and a certain destination area $\mathcal{S}$ for information transfer, and no direct communication links are present between the MBSs and the edge UTs.

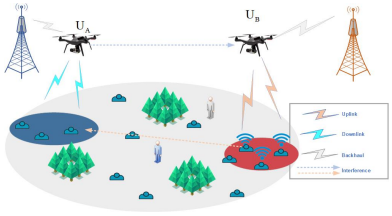


Fig. 1: System model. Two UAV-enabled BS $\mathrm{U_A}$ and $\mathrm{U_B}$ are equipped with single-antenna each and UTs distributed in destination area $\mathcal{S}$.

Area $\mathcal{S}$ is divided into smaller regions. Each is entirely covered by the footprint of the respective UAVs. The number of regions depends on their radii. The footprint has a hexagonal shape as we approximate the analysis of UTs on the edges of the map in a low-density scenario. $\mathrm{U_A}$ has a radius r_{A}, the number of sub-areas $K = \frac{2\mathcal{S}}{3\sqrt{3}r_{\mathrm{A}}^2}$, and the set of all sub-areas is denoted as $\mathcal{K} = \{\widetilde{s}_1, ..., \widetilde{s}_k, ..., \widetilde{s}_K\}$. Similarly, $\mathrm{U_B}$ will have $J = \frac{2\mathcal{S}}{3\sqrt{3}r_{\mathrm{B}}^2}$ sub-ares and the set $\mathcal{J} = \{\widehat{s}_1, ..., \widehat{s}_j, .., \widehat{s}_J\}$. Without loss of generality, the two UAV-enabled BSs serve all the UTs located in $\mathcal{S}$, and therefore we tag the starting sub-area as $\widetilde{s}_{o,k} \in \mathcal{K}$ and $\widehat{s}_{o,j} \in \mathcal{J}$, respectively. Once they reach the next sub-area, they stop and change their status in the hovering slots and start fulfilling requests. The time for UAVs to carry out their respective actions may differ. The flight time depends mainly on the distance travelled and the flight speed. The hovering time depends on the UTs' bit rate request, their density, and the level of the achievable rate. $\mathrm{U_A}$ and $\mathrm{U_B}$ can communicate with one UT at a time. We denote T_m^{A} and T_m^{B} as the total transition times from one sub-area to another for $\mathrm{U_A}$ and $\mathrm{U_B}$, respectively. Furthermore, T_h^{A} and T_h^{B} are the sums of all the time spent on hovering for $\mathrm{U_A}$ and $\mathrm{U_B}$, respectively.

We assume that the two UAVs can serve area $\mathcal{S}$ simultaneously. The coverage radii for $\mathrm{U_A}$ and $\mathrm{U_B}$ are r_{A} and r_{B} respectively. Based on the PPP model, the average active user numbers in the coverage areas of $\mathrm{U_A}$ and $\mathrm{U_B}$ are $\mathcal{S}_{\mathrm{A}} = \widetilde{\lambda}_{\mathtt{G}}\frac{3\sqrt{3}}{2}r_{\mathrm{A}}^2$ and $\mathcal{S}_{\mathrm{B}} = \widetilde{\lambda}_{\mathtt{G}}\frac{3\sqrt{3}}{2}r_{\mathrm{B}}^2$, respectively. We denote $\Phi_{\mathtt{A},k}$ as the UT sub-set receiving the downlink signal from $\mathrm{U_A}$, $\Phi_{\mathtt{B},j}$ as the uplink signal UT sub-set to $\mathrm{U_B}$.

1) LoS Probability: The communication links between two nodes can be modeled by a probabilistic path loss model, where both the line-of-sight (LoS) and non-LoS (NLoS) links can be considered separately with different probabilities of occurrences. The probability of having a LoS connection between the two nodes with distance X is given by [12]

$$p_{\mathrm{LoS}}(X) = \frac{1}{1 + a\exp(-b\tan^{-1}(\frac{h_i}{X}) - a)}, \tag{1}$$

where a and b are two constants that depend on the environment, h_i is the height of the correspondent i-th UAV-enabled BS, where $i \in \{\mathrm{A}, \mathrm{B}\}$. The probability of NLoS links is

$$p_{\mathrm{NLoS}}(X) = 1 - p_{\mathrm{LoS}}(X). \tag{2}$$

2) Downlink Transmission: For downlink transmission, the received signal to interference plus noise ratio (SINR) from UAV $\mathrm{U_A}$ to the associated UT in the $\Phi_{\mathtt{A}}$ is [13]

$$\mathrm{SINR_A}(X_{\mathtt{A},o}) = \frac{P_{\mathtt{A}}\hbar_a\beta d_{\mathtt{A},o}^{-\alpha_o}}{\delta\overline{\mathcal{I}}_u + \sigma^2}, \tag{3}$$

where $P_{\mathtt{A}}$ is the transmit power from $\mathrm{U_A}$, σ^2 is the additive noise power, $\hbar_a \sim \Gamma(1,1)$ is the equivalent small-scale fading channel power gain between $\mathrm{U_A}$ and the UTs, and $\Gamma(k_1, k_2)$ is the Gamma distribution with a shape parameter k_1 and a scale parameter k_2. In (3), β denotes a frequency dependent constant parameter, $d_{\mathtt{A},o}$ is the distance between the $\mathrm{U_A}$ to the typical user o in $\Phi_{\mathtt{A}}$ and can be computed by

$$d_{\mathtt{A},o} = \sqrt{X_{\mathtt{A},o}^2 + {h_{\mathrm{A}}}^2}, \tag{4}$$

where $X_{\mathtt{A},o}$ is the projection distance from $\mathrm{U_A}$ to the typical user o. Note that δ in (3) is the index of beaconing refereed to the interference state from $\mathrm{U_A}$ to $\mathrm{U_B}$, given by

$$\delta = \begin{cases} 1, & \text{if both UAVs hovering,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The inter-cell interference $\overline{\mathcal{I}}_u$ from $\Phi_{\mathtt{B}}$ is given by

$$\overline{\mathcal{I}}_u = P_u g_u \beta \left|\overline{X}_{\mathtt{A}}\right|^{-\alpha_{\mathrm{N}}}, \tag{6}$$

where $\overline{X}_{\mathtt{A}}$ is the average distance between a downlink typical user o to the active UT set $\Phi_{\mathtt{B}}$. Also, $g_u \sim \exp(1)$ is the small scale fading channel power gain. The average distance between a downlink typical user o to $\mathrm{U_B}$ is given by

$$\overline{X}_{\mathtt{A}} = \frac{\sum_{u\in\Phi_{\mathtt{B}}} \left|X_{B_u,A_d}^t\right|^{-\alpha_o}}{r_{\mathrm{B}}^2\frac{3\sqrt{3}}{2}\lambda_{\mathtt{G}}}. \tag{7}$$

3) Uplink Transmission: The received SINR from typical UTs in $\Phi_{\mathtt{B}}$ to the serving UAV $\mathrm{U_{Bi}}$ is defined as

$$\mathrm{SINR_B}(X_{\mathtt{B},\dot{o}}) = \frac{P_u\hbar_b\beta d_{\mathtt{B}}^{-\alpha_o}}{\delta\mathcal{I}_d + \sigma^2}, \tag{8}$$

where $d_{\mathtt{B}} = \sqrt{X_{\mathtt{B},\dot{o}}^2 + {h_{\mathrm{B}}}^2}$ is the distance between typical UTs to $\mathrm{U_B}$. The inter-cell interference is given by

$$\mathcal{I}_d = P_{\mathtt{A}} g_a \beta |d_{\mathtt{A},\mathtt{B}}|^{-\alpha_{\mathrm{L}}}, \tag{9}$$

where $d_{\mathtt{A},\mathtt{B}} = \sqrt{X_{\mathtt{A},\mathtt{B}}^2 + |h_{\mathrm{A}} - h_{\mathrm{B}}|^2}$, and with $g_a \sim \exp(1)$ being the small scale fading channel power gain from the interference $\mathrm{U_A}$.

4) Queuing: In our system model, we consider the existing queue in the target sub-area with L_a^i the UTs already in the area below $\mathrm{U_i}$ and the second queue formed by the new arrivals L_n^i during the hovering time. The arrival rate follows a Poisson distribution with parameter ν_i and the service time rate t_{μ_i} follows an exponential distribution with parameter μ_i. Since the proposed system can be considered as a finite $M/M/1$ with an initial queue length L_a^i, we can compute the expectation of service time for each UT as

$$\mu_i = \frac{1}{t_{\mu_i}} = \frac{\overline{\mathcal{R}}_i}{\mathtt{Q}_i}, \tag{10}$$

where $\mathtt{Q}_i$ is the number of the approximate data requested, for $i \in \{\mathrm{A}, \mathrm{B}\}$. From [14], the number of new arrivals over the service time will be $L_n^i = \frac{\nu_i}{\mu_i - \nu_i}$. We can easily get the average waiting time for each user in the queue under $\mathrm{U_i}$ as $T_n^i = \frac{L_n^i}{\nu_i}$. Hence, the hovering time for $\mathrm{U_i}$ will be

$$t_h^i = \frac{L_n^i}{\nu_i} + \frac{L_a^i}{\mu_i} = \frac{1}{\mu_i - \nu_i} + \frac{L_a^i}{\mu_i}. \tag{11}$$

Moreover, the total length in the queue needs to be less than the total number of UTs in the target area while

$$r_i \geq \sqrt{\frac{2(\frac{\nu_i}{\mu_i - \nu_i} + L_a^i)}{\mathcal{S}_i}}. \tag{12}$$

Furthermore, we have $\nu_i < \mu_i$ in order to respect the existence conditions of stationarity of the queue.

III. Performance Evaluation

The summarize tractable lower bound for the conditional average downlink achievable rate between UEs and their serving $\mathrm{U_A}$ can be computed as

$$\mathcal{R}_{\mathrm{A}}^{\mathrm{Low}}(\theta_1, u_1) = \frac{\widetilde{\lambda}_{\mathtt{G}}}{\mathcal{S}_{\mathrm{A}}} \int_0^{2\pi} \int_0^{r_{\mathrm{A}}} \log_2\left[1 + \Delta_{\mathrm{A}}\right] d\theta_1 du_1, \tag{13}$$

where Δ_{A} is expressed as (14) shown at the top of the next page. Note that $\Xi(x_1, x_2, \vartheta) = \left({x_1}^2 + {x_2}^2 - 2x_1 x_2 \cos\vartheta\right)^{1/2}$. For what concern the data rate lower bound in the uplink scenario, it can be computed as

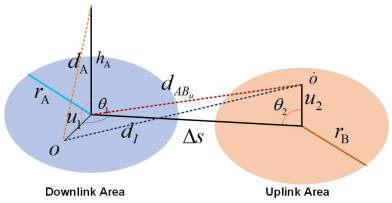


Fig. 2: Distance diagram between $\mathrm{U_A}$ and $\mathrm{U_B}$

$$\overline{\mathcal{R}}_{\mathrm{B}}^{\mathrm{Low}} = \frac{1}{\mathcal{S}_{\mathrm{B}}} \int_0^{2\pi} \int_0^{r_{\mathrm{B}}} \log_2(1 + \frac{1}{\Pi(u_2)}) u_2 du_2 d\theta, \tag{15}$$

with

$$\Pi(u_2) = \frac{\delta P_{\mathtt{A}} \beta \sqrt{\Delta s^2 + (h_{\mathrm{A}} - h_{\mathrm{B}})^2}^{-\alpha_{\mathrm{L}}} + \sigma^2}{P_u \beta \left\{ \frac{p_{\mathrm{LoS}}(u_2)}{\sqrt{{u_2}^2 + {h_{\mathrm{B}}}^2}^{\alpha_{\mathrm{L}}}} + \frac{p_{\mathrm{NLoS}}(u_2)}{\sqrt{{u_2}^2 + {h_{\mathrm{B}}}^2}^{\alpha_{\mathrm{N}}}} \right\}}. \tag{16}$$

IV. Energy Consumption Analysis

In this section, we analyze the energy consumption in the whole process. We consider the energy consumption model as rotary wings UAV based BSs, which includes communication-related energy and propulsion power consumption. The non-straight flight refers to the uniform accelerated movement in the horizontal plane. For a rotary-wing UAV, its speed can be computed as

$$v(t) = 2\left(\frac{3}{t^2}\mathrm{d_s}\right)\mathcal{T}_{f,f+1} + 3\left(\frac{2}{t^2}\mathrm{d_s}\right)\mathcal{T}_{f,f+1}^2, \tag{17}$$

where $\mathrm{d_s}$ is the space travelled and the power consumption can be expressed as

$$\mathcal{E}_{(m)}^i(t) = \sum_{s_f \in \mathcal{F}} \Bigg[\int_0^{\mathcal{T}_{f,f+1}} c_1[1 + c_2|v(t)|^2]dt + \int_0^{\mathcal{T}_{f,f+1}} c_5|v(t)|^3 dt + \int_0^{\mathcal{T}_{f,f+1}} c_3 \sqrt{1 + \frac{a_{\mathrm{U}}^2(t)}{g^2}} \cdot \left(\sqrt{1 + \frac{a_{\mathrm{U}}^2(t)}{g^2} + \frac{|v(t)|^4}{c_4}} - \frac{|v(t)|^2}{c_4} \right)^{1/2} dt + \Delta K \Bigg] dt, \tag{18}$$

where $\mathcal{T}_{f,f+1}$ is the transition time from one sub-area to the next with $f \in \{k, j\}$ and $\mathcal{F} \in \{\mathcal{K}, \mathcal{J}\}$, g is the gravitational acceleration, a_{U} is the UAV acceleration given by

$$a_{\mathrm{U}}(t) = 2\left(\frac{3}{t^2}\mathrm{d_s}\right) + 6\left(\frac{2}{t^3}\mathrm{d_s}\right)\mathcal{T}_{f,f+1}, \tag{19}$$

and ΔK is the change in kinetic energy

$$\Delta K = \frac{1}{2} m \left(|v(\mathcal{T}_{f,f+1})|^2 - |v(0)|^2 \right), \tag{20}$$

with $v(T_m^i)$ the final and $v(0)$ the initial $\mathrm{U_i}$ speeds. In the above, the UAV acceleration and speed functions have been computed using an interpolation technique as in [15]. The polynomial function used is cubic as we are in a system with constant acceleration/deceleration. Moreover, $c_i \in [1, 2, ..., 5]$ are the modelling parameters that depend on the UAV weight, air density and rotor disc area, as specified in [16]. The communication-related energy for the UAVs can be expressed as [17]

$$\mathcal{E}_{(p)} = (T_h^{\mathrm{A}} \delta P_{\mathtt{A}}) + (T_h^{\mathrm{B}} \mathcal{S}_{\mathrm{B}} P_u). \tag{21}$$

***Corollary** 1:* For the static speed $v(t) \to 0$, the power consumption corresponding to the hovering UAV at the fixed location is asymptotically derived as

$$\mathcal{E}_{(h)}^i = T_h^i \{c_1 + c_3\}, \tag{22}$$

where T_h^i is the total time spent hovering, defined as

$$T_h^{\mathrm{A}} = K \cdot t_h^{\mathrm{A}}, \tag{23}$$

$$T_h^{\mathrm{B}} = J \cdot t_h^{\mathrm{B}}. \tag{24}$$

$$\Delta_{\mathrm{A}} = P_{\mathtt{A}}\beta\left(\frac{p_{\mathrm{LoS}}\left(u_1\right)}{\sqrt{u_1^2+h_{\mathrm{A}}^2}^{\alpha_{\mathrm{L}}}}+\frac{p_{\mathrm{NLoS}}\left(u_1\right)}{\sqrt{u_1^2+h_{\mathrm{A}}^2}^{\alpha_{\mathrm{N}}}}\right)\Bigg/\delta\frac{\widetilde{\lambda}_G P_u\beta}{\left[\frac{\int_0^{2\pi}\int_0^{r_{\mathrm{A}}}\Xi(d_{\mathrm{AB_u}},u_1,\theta_1)u_1du_1d\theta_1}{\mathcal{S}_{\mathrm{A}}}\right]^{\alpha_N}}+\sigma^2, \tag{14}$$

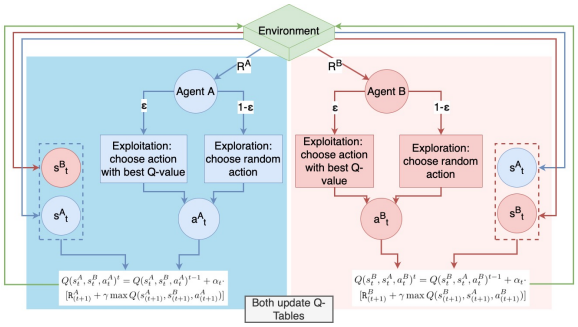


Fig. 3: JQ-LMA algorithm framework.

V. Problem Statement and JQ-LMA

In this section, we present our objective and details for the JQ-LMA algorithm. The energy efficiency of the whole system is maximized and computed as

$$\mathtt{EE} = \frac{K\mathcal{S}_{\mathrm{A}}\overline{\mathcal{R}}_{\mathrm{A}} + J\mathcal{S}_{\mathrm{B}}\overline{\mathcal{R}}_{\mathrm{B}}}{\sum_{i\in\{\mathrm{A,B}\}}\mathcal{E}^i_{(m)}+\sum_{i\in\{\mathrm{A,B}\}}\mathcal{E}^i_{(h)}+\mathcal{E}_{(p)}}, \tag{25}$$

where $\mathcal{E}^i_{(m)}$ denotes the mechanical power consumption of $\mathrm{U_A}$ and $\mathrm{U_B}$ computed as (18), $\mathcal{E}^i_{(h)}$ denotes the hovering power consumption (22), and $\mathcal{E}_{(p)}$ is the transmit power consumption of $\mathrm{U_A}$ and UTs in area $\mathcal{S}$ during the total period (21). The trajectories are computed by solving the Travelling Sales Problem (TSP) task jointly with transmit power management and interference reduction problems in a 3D environment. To achieve the maximum energy efficiency of the whole system, the optimization problem can be formulated as

$$\mathbb{P}_1: \begin{cases} \max\limits_{\Theta(t)} & \mathbb{E}\left[\mathtt{EE}\middle|V^{\pi}_{\Theta(t)}\right] \\ \text{s.t.} & \mathrm{C1}: \delta\in\{0,1\}, \\ & \mathrm{C2}: v(t)\le V_{\max}, \end{cases} \tag{26}$$

where $V^{\pi}\left(a\,|s,\Theta\right)$, $\Theta = (\mathbf{c}_{\mathrm{A}},\mathbf{c}_{\mathrm{B}},P_{\mathtt{A}})$, $\mathbf{c}_{\mathrm{A}} = \{\widetilde{s}_{o,k}\}$ and $\mathbf{c}_{\mathrm{B}} = \{\widehat{s}_{o,j}\}$ are the optimized subarea sequence of UAVs trajectories, and $\overline{P}^{\mathtt{A}}$ is the transmit power constraint for $\mathrm{U_A}$. Overall, $\mathbb{P}_1$ can provide the best trajectories and transmit power and achieve maximum energy efficiency of the whole network. To maximize the energy efficiency of the whole destination edge area, we use multiple agent reinforcement model $(\mathrm{U_A},\mathrm{U_B})$ to achieve the minimum mutual interference and energy consumption with two sets of states S_{A} and S_{B}, and two actions sets C_{A} and C_{B} for $\mathrm{U_A}$ and $\mathrm{U_B}$, respectively. By carrying out an action $a^i_t\in C_i$, for $i\in\{\mathrm{A,B}\}$, at the t-th iteration of the algorithm, the agent moves from one state to another state. In our proposed multiple agent algorithm, $\mathrm{U_A}$ sends position related information to agent $\mathrm{U_B}$, which on the other side will choose an action based on this information and vice versa (Fig. 3).

Reward: The reward function of agent $\mathrm{U_A}$ at each iteration t is defined as

$$\mathtt{R}^{\mathrm{A}}_t = \frac{\mathcal{S}_{\mathrm{A}}\overline{\mathcal{R}}_{\mathrm{A}}\cdot\left[1+\left(K\mathcal{S}_{\mathrm{A}}-\mathtt{G}^{\mathrm{A}}_t\right)\right]^{-1}}{\mathcal{E}^{\mathrm{A}}_{(m)}+\mathcal{E}^{\mathrm{A}}_{(h)}+\mathcal{E}_{(p)}}, \tag{27}$$

where $\mathtt{G}^{\mathrm{A}}_t$ is the number of users already served by corresponding agent $\mathrm{U_A}$ in the previous iterations, computed as

$$\mathtt{G}^{\mathrm{A}}_t = k_t(L^{\mathrm{A}}_a + L^{\mathrm{A}}_n), \quad k_t\le K, \tag{28}$$

with k_t as the number of sub-area already covered by $\mathrm{U_A}$ previously. Similarly, the $\mathrm{U_B}$ reward at each algorithm iteration t can be defined as

$$\mathtt{R}^{\mathrm{B}}_t = \frac{\mathcal{S}_{\mathrm{B}}\overline{\mathcal{R}}_{\mathrm{B}}\cdot\left[1+\left(J\mathcal{S}_{\mathrm{B}}-\mathtt{G}^{\mathrm{B}}_t\right)\right]^{-1}}{\mathcal{E}^{\mathrm{B}}_{(m)}+\mathcal{E}^{\mathrm{B}}_{(h)}+\mathcal{E}_{(p)}}, \tag{29}$$

where $\mathtt{G}^{\mathrm{B}}_t = j_t(L^{\mathrm{B}}_a + L^{\mathrm{B}}_n)$, with $j_t\le J$. The proposed reward function in (27) and (29) enables to achieve three objectives: maximizing the average achievable rate for downlink and uplink, maximizing the coverage area for the active UTs, and minimizing the energy consumption. The terms $(K\mathcal{S}_{\mathrm{A}}-\mathtt{G}^{\mathrm{A}}_t)$ in (27) and $(J\mathcal{S}_{\mathrm{B}}-\mathtt{G}^{\mathrm{B}}_t)$ in (29) can be considered as the effective incremental coverage, which adds a penalty to the actual value. Maximizing the cumulative reward is equivalent to maximizing the energy efficiency.

Agents, States and Actions: We assume two agents A and B, corresponding to $\mathrm{U_A}$ and $\mathrm{U_B}$. The states of the agents are the three-dimensional location of the UAVs at the t-iteration. The action set $\mathbf{C}_{\mathrm{A}}$ is the output of agent $\mathrm{U_A}$ in the network.

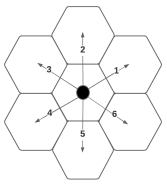


Fig. 4: Actions direction.

Each action $\mathbf{a}^{\mathrm{A}}_t = [m^{\mathrm{A}}_t, P_{\mathtt{A}_k}] \in \mathbf{C}_{\mathrm{A}}$ by $\mathrm{U_A}$ contains the movement action and the transmit power action. Note that m^{A}_t represents the movement direction on a 2D surface as $m^{\mathrm{A}}_t\in\{0,1,2,3,4,5,6\}$, where value 0 indicates the hovering action at the same position, as shown in Fig. 4. Different as agent $\mathrm{U_A}$, $\mathbf{a}^{\mathrm{B}}_t = [m^{\mathrm{B}}_t]\in\mathbf{C}_{\mathrm{B}}$, and $\mathbf{C}_{\mathrm{B}}$ does not have to consider the transmit powers because P_u are managed by the UTs.

Learning Rate: The learning rate factor $\alpha_t\in[0,1]$ can control the speed of updated information at which the model learns. In particular, $\alpha_t = 0$ means that the agent has stopped learning, which uses the last episode ignoring all the steps before. When the environment is deterministic the optimal

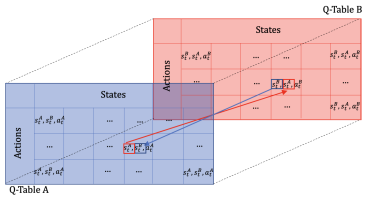


Fig. 5: Q-Tables interaction at the t-th iteration.

learning rate, α_t, will approach to 1. The step size used in the JQ-LMA is adaptive and can be expressed as

$$\alpha_t = \left[\alpha_{\mathrm{M}} + \sum_{t=1}^{t/\tau} \frac{(\alpha_{\mathrm{M}}/\alpha_{\mathrm{m}})\tau}{\mathtt{T}}\right]^{-1}, \tag{30}$$

with $\alpha_{\mathrm{m}} \leq \alpha_t \leq \alpha_{\mathrm{M}}$, and where $\mathtt{T}$ is the expected training iterations value. We assume that α_t decreases every τ iterations following the degrowth factor $\frac{(\alpha_{\mathrm{M}}/\alpha_{\mathrm{m}})\tau}{\mathtt{T}}$ in (30). The dynamic learning rate speeds up the training time and guarantees the convergence of the algorithm, as proved in [18].

Discount Factor: The agents choose their policy according to the discount factor $\gamma \in (0, 1)$ which is fixed during the training process. The value of the discount factor will change the behavior of agents, and a higher discount factor will make the agents more greedy to look for future rewards.

Probability of action choice: In a classic Q-Learning approach, when the states and actions begin to grow exponentially, the probability (ϵ_t) that the agent will be able to visit all the cells by performing all possible actions decreases considerably. In JQ-LMA, we overcome this issue by employing a dynamic ϵ_t. The value of ϵ_t is used by the agent when choosing what action to use. In particular, ϵ_t is the probability, that the agent chooses the highest value of the available states in the Q-Table, while a random action is taken to help the agent explore with a probability $1 - \epsilon_t$. That is

$$a^i_{t+1} = \begin{cases} \arg\max Q(\varrho^i_{t+1}) & \epsilon_{t+1}, \\ \text{random} & 1 - \epsilon_{t+1}, \end{cases} \tag{31}$$

where $(1 - \epsilon_{t+1})$ is the probability to take a random action instead of follow the optimizing policy π_i [19], and $\arg\max Q(\varrho^i_{t+1})$ is the max future Q-value, with $\varrho^{\mathrm{A}}_{t+1} = (s^{\mathrm{A}}_{t+1}, s^{\mathrm{B}}_{t+1}, a^{\mathrm{A}}_{t+1})$, and $\varrho^{\mathrm{B}}_{t+1} = (s^{\mathrm{B}}_{t+1}, s^{\mathrm{A}}_{t+1}, a^{\mathrm{B}}_{t+1})$. In JQ-LMA, ϵ_t is linearly dynamic and increases for every τ iterations, up to a maximum value ϵ_{M} over the algorithm training, according to the following constraints:

$$\epsilon_t = \epsilon_{\mathrm{m}} \cdot \left\{1 + \sum_{t=1}^{t/\tau} \frac{[(\epsilon_{\mathrm{M}}/\epsilon_{\mathrm{m}}) - 1]\tau}{\mathtt{T}}\right\}, \tag{32}$$

with $\epsilon_{\mathrm{m}} \leq \epsilon_t \leq \epsilon_{\mathrm{M}}$. At the iteration $t = 0$, we have $\epsilon_t = \epsilon_{\mathrm{m}}$, where ϵ_{m} is the minimum ϵ_t value used in the training. In this situation, the probability to take a random action is higher than to take an action based on $\arg\max Q(\varrho^i_{t+1})$. This initial behavior allows the agents first to explore the environment and enrich the Q-tables. An episode of the algorithm ends when the agents have served all the users, or when the agent goes out from the map. We define the value function $V^\pi : S_i \rightarrow \mathtt{R}^i$ that represents the expected value obtained by following policy π_i from each state $s^i_t \in S_i$. The value function V_i for policy π quantifies the goodness of the policy through an infinite horizon and can be expressed as follows:

$$V^\pi_i(s^i_t) = \mathbb{E}_\pi[\mathtt{R}^i_t \varrho^i_t + \gamma V^\pi_i(s^i_{t+1}|_{s^i_t}). \tag{33}$$

VI. Simulations Results

In this section, the performance of the proposed multi-UAV system is evaluated by presenting numerical results. In the simulations, we consider an area of 1 km^2. The horizontal locations of the UAV-BSs are restricted in the area. We assume that the noise power is $\sigma^2 = -174 + 10\log_{10}(\mathrm{BW}) + \mathrm{Nf}$ dBm, where BW is the mmWave bandwidth equals to 3 GHz and Nf the noise figure equals to 10 dB. Moreover, the frequency is equal to $\frac{c}{4\pi f_c}$, with $f_c = 1$ GHz, the urban environments parameters a and b respectively equal to 9.6 and 0.28. The footprint radii are 90 m for both UAVs, while the altitudes are respectively 70 m for $\mathrm{U_A}$ and 80 m for $\mathrm{U_B}$. The UTs have a 250/km^2 density and an active probability $p_{\mathtt{G}} = 0.8$. Furthermore, α_{L} and α_{N} are 2 and 3 respectively [20]. The propulsion modelling parameters are explicit in [16]. Finally, the JQ-LMA parameters are $0.1 \leq \alpha_t \leq 1$, $0.4 \leq \epsilon_t \leq 0.95$, $\gamma = 0.9$, $\tau = 10$, and the training iterations 250.

A. Dynamic Parameters Performance

Here we analyze the JQ-LMA dynamic parameters performance in terms of time processing compared with others four pairs parameters approaches: ***1)*** with a fixed learning rate and dynamic ϵ_t, ***2)*** with a linearly dynamic α_t and dynamic ϵ_t, ***3)*** with a dynamic α_t and fixed ϵ, ***4)*** with fixed α and ϵ.

Fig. 6 shows the algorithm time processing of the five parameters pairs against the footprint radii of $\mathrm{U_A}$ and $\mathrm{U_B}$. Results illustrate that the proposed approach largely decreases the algorithm time processing. Compared with the static and the only-learning rate dynamic approaches (green and grey bars), the proposed parameters decrease the time with peaks of 70% less, while compared to the linearly dynamic approach (orange bars) the time decrease reaches up to 25%, and compared to the only-ϵ_t dynamic approach (red bars) the time is slightly lower. More importantly, the minimum time processing is obtained with $r_{\mathrm{A}} = r_{\mathrm{B}} = 90$ m for all the approaches. At this value, it has reached the optimal trade-off between the computational time for the flying movement and the computational time for interference. Indeed, the number of sub-flights done by the UAVs from one group of UTs to another increases with lower radii, and statistically, the number of times the UAVs serve simultaneously two UTs groups and thus generate interference increases with higher radii, hence increasing the computational time.

B. Performance Comparison

In this sub-section, we compare our algorithm with two other approaches, a random action selection and a Zigzag trajectory approach. Fig. 7 provides the results for the overall system results in terms of achievable rate, energy consumption,

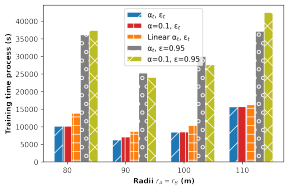
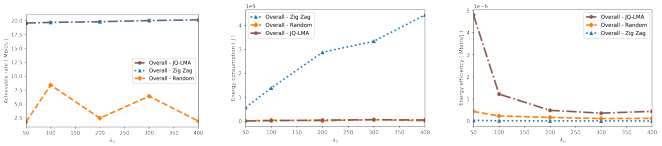


Fig. 6: Time processing against footprint radii, with $r_{\mathrm{A}} = r_{\mathrm{B}}$, $\mathtt{T} = 250$, and $\tau = 10$.



(a) Overall JQ-LMA, Random and Zigzag achievable rates.

(b) Overall JQ-LMA, Random and Zigzag energy consumption.

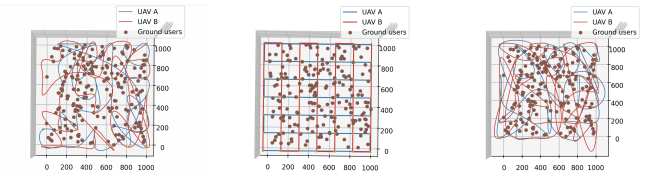
(c) Overall JQ-LMA, Random and Zigzag energy efficiency.

Fig. 7: JQ-LMA, Random and Zigzag comparison, with $r_{\mathrm{A}} = 90$ m, $r_{\mathrm{B}} = 90$, $P_{\mathtt{A}} = 46$ dBm $P_u = 30$ dBm, $h_{\mathrm{A}} = 70$ m and $h_{\mathrm{B}} = 80$ m, against UTs density.

and energy efficiency of the three approaches, against multiple densities. The achievable rate curves in Fig. 7a are obtained by the sum of (13) and (15), the energy consumption solid curves in 7b, by the sum of (18), (21), and (22), while the energy efficiency curves in Fig. 7c, by (25). Results in Fig. 7a illustrates that with any kind of UTs density the JQ-LMA and the Zigzag approach overcome the random approach. Results in Fig. 7b illustrate that JQ-LMA and the random approach with any kind of UTs density both save more than the Zigzag approach. More importantly, Fig. 7c demonstrates the JQ-LMA algorithm energy efficiency superiority compared to the cited approaches, with the optimal point at $\lambda_{\mathtt{G}} = 50/\mathrm{km}^2$. Finally, Fig. 8 compares the trajectories. In all three cases, the UAVs totally cover the UTs distributed on the ground. It can be easily noticed that the more confusing trajectory in Fig. 8c does not allow to reach good levels of achievable rate and the energy-intensive trajectory, due to square direction changes, of the Zigzag UAVs in Fig. 8b.

VII. Conclusion

This paper proposes a multiple UAVs-enabled solution to collaboratively optimize the whole system's energy efficiency.



(a) JQ-LMA UAVs trajectories.

(b) Zigzag UAVs trajectories.

(c) Random UAVs trajectories.

Fig. 8: JQ-LMA, Zigzag and Random output trajectories, with $r_{\mathrm{A}} = 90$ m and $h_{\mathrm{A}} = 70$, $r_{\mathrm{B}} = 90$ m and $h_{\mathrm{B}} = 80$ m, $\lambda_{\mathtt{G}} = 250/\mathrm{km}^2$ and $p_{\mathtt{G}} = 0.8$.

The service time has been determined using queue theory and several major energy consumptions, including the UAVs' energy consumption, have been considered. We have proposed JQ-LMA to improve the energy efficiency of the whole system. The parameters for RL have been investigated to achieve a quick training process. Our analysis has shown that the proposed JQ-LMA approach achieves much better performance than the zigzag and random approaches.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper," Feb. 2019. [Online]. Available: https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf

[2] M. Asadpour, B. Van den Bergh, D. Giustiniano, K. A. Hummel, S. Pollin, and B. Plattner, "Micro Aerial Vehicle Networks: an Experimental Analysis of Challenges and Opportunities," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 141–149, July 2014.

[3] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-Efficient UAV Control for Effective and Fair Communication Coverage: A Deep Reinforcement Learning Approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sept. 2018.

[4] Yang, Zhaohui and Xu, Wei and Shikh-Bahaei, Mohammad, "Energy Efficient UAV Communication With Energy Harvesting," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1913–1927, Feb. 2020.

[5] S. Fu, Y. Tang, Y. Wu, N. Zhang, H. Gu, C. Chen, and M. Liu, "Energy-Efficient UAV Enabled Data Collection via Wireless Charging: A Reinforcement Learning Approach," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 10 209–10 219, June 2021.

[6] N. I. Mowla, N. H. Tran, I. Doh, and K. Chae, "AFRL: Adaptive Federated Reinforcement Learning for Intelligent Jamming Defense in FANET," vol. 22, no. 3, pp. 244–258, June 2020.

[7] S. Zhu, L. Gui, N. Cheng, F. Sun, and Q. Zhang, "Joint Design of Access Point Selection and Path Planning for UAV-Assisted Cellular Networks," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 220–233, Jan. 2020.

[8] L. Li, Q. Cheng, K. Xue, C. Yang, and Z. Han, "Downlink Transmit Power Control in Ultra-Dense UAV Network Based on Mean Field Game and Deep Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15 594–15 605, Dec. 2020.

[9] C. K. Armeniakos, P. S. Bithas, and A. G. Kanatas, "SIR Analysis in 3D UAV Networks: A Stochastic Geometry Approach," *IEEE Access*, vol. 8, pp. 204 963–204 973, Nov. 2020.

[10] K. Yoshikawa, K. Yamamoto, T. Nishio, and M. Morikura, "Grid-Based Exclusive Region Design for 3D UAV Networks: A Stochastic Geometry Approach," *IEEE Access*, vol. 7, pp. 103 806–103 814, July 2019.

[11] F. Jiang and A. L. Swindlehurst, "Optimization of UAV Heading for the Ground-to-Air Uplink," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 5, pp. 993–1005, June 2012.

[12] J. Cui, Y. Liu, and A. Nallanathan, "Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[13] Y. Zhu, L. Wang, K.-K. Wong, S. Jin, and Z. Zheng, "Wireless Power Transfer in Massive MIMO-Aided HetNets With User Association," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4181–4195, Oct. 2016.

[14] U. Bhat, *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhäuser Boston, 2015.

[15] J. J.Craig, *Introduction to Robotics, Mechanics and Control*. Upper Saddle River, NJ 07458: Pearson Education International, 2005.

[16] Y. Zeng, J. Xu, and R. Zhang, "Energy Minimization for Wireless Communication With Rotary-Wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[17] C. Zhan and Y. Zeng, "Energy-Efficient Data Uploading for Cellular-Connected UAV Systems," *IEEE Trans. Wireless Commun.*, pp. 7279–7292, Nov. 2020.

[18] C. J. C. H. Watkins and P. Dayan, *Q-learning*, 1992.

[19] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, May 2019.

[20] Y. Zhu, G. Zheng, and M. Fitch, "Secrecy Rate Analysis of UAV-Enabled mmWave Networks Using Matérn Hardcore Point Processes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1397–1409, July 2018.