IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS Connecting the Mobile World | IEEE Transactions on Machine Learning in Communications and Networking

Received 7 October 2023; revised 6 February 2024 and 29 April 2024; accepted 29 April 2024.
Date of publication 2 May 2024; date of current version 15 May 2024.

The associate editor coordinating the review of this article and approving it for publication was M. Mozaffari.

Digital Object Identifier 10.1109/TMLCN.2024.3396438

Energy-Efficient Trajectory Planning With Joint Device Selection and Power Splitting for mmWaves-Enabled UAV-NOMA Networks

AHMAD GENDIA[1,2] **(Member, IEEE), OSAMU MUTA**[3] **(Member, IEEE),**
SHERIEF HASHIMA[4,5] **(Senior Member, IEEE),**
AND KOHEI HATANO[4,6]

[1]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan
[2]Electrical Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo 11884, Egypt
[3]Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan
[4]Computational Learning Theory Team, RIKEN-AIP, Fukuoka 819-0395, Japan
[5]Engineering Department, Egyptian Atomic Energy Authority, Cairo 13759, Egypt
[6]Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan

Corresponding author: Ahmad Gendia (ahmad.gendia@mobcom.ait.kyushu-u.ac.jp)

This work was supported in part by the Ministry of Higher Education (MoHE) of the Arab Republic of Egypt, in part by the Support Center for Advanced Telecommunications (SCAT) Technology Research Foundation, and in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research under Grant 24K07490.

ABSTRACT This paper proposes two energy-efficient reinforcement learning (RL)-based algorithms for millimeter wave (mmWave)-enabled unmanned aerial vehicle (UAV) communications toward beyond-5G (B5G). This can be especially useful in ad-hoc communication scenarios within a neighborhood with main-network connectivity problems such as in areas affected by natural disasters. To improve the system's overall sum-rate performance, the UAV-operated mobile base station (UAV-MBS) can harness non-orthogonal multiple access (NOMA) as an efficient protocol to grant ground devices access to fast downlink connections. Dynamic selection of suitable hovering spots within the target zone where the battery-constrained UAV needs to be positioned as well as calibrated NOMA power control with proper device pairing are critical for optimized performance. We propose cost-subsidized multiarmed bandit (CS-MAB) and double deep Q-network (DDQN)-based solutions to jointly address the problems of dynamic UAV path design, device pairing, and power splitting for downlink data transmission in NOMA-based systems. To verify that the proposed RL-based solutions support high sum-rates, numerical simulations are presented. In addition, exhaustive and random search benchmarks are provided as baselines for the achievable upper and lower sum-rate levels, respectively. The proposed DDQN agent achieves 96% of the sum-rate provided by the optimal exhaustive scanning whereas CS-MAB reaches 91.5%. By contrast, a conventional channel state sorting pairing (CSSP) solver achieves about 89.3%.

INDEX TERMS NOMA resource control, reinforcement learning, UAV emergency communications.

I. INTRODUCTION

BEYOND 5G (B5G) and 6G cellular networks face design challenges due to their increased requirements on massive connectivity and communication speeds [1], [2], [3], [4], [5], [6], [7]. This can be particularly pressing in zones stricken by disasters where the primary base stations (BSs) infrastructure is momentarily out of commission owing to sustaining severe or mild impairment. In such scenarios, ad-hoc intervention based on dispatched unmanned aerial vehicles (UAVs) can allow for a quick and suitable remedy to maintain adequate coverage and provide high-speed, reliable wireless connections to offload downlink data to appropriately activated receivers within the afflicted region [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. The mobile UAV base stations (UAV-MBSs) can therefore play the pivotal role of rapidly dispatched mobile BSs within certain target regions.

Incorporating mmWaves into UAV-based transmissions can provide vital advantages to the communication system. The huge bandwidth resources provided by mmWaves can help the UAV-mounted BS to support high-speed communications as well as flexible coverage [19], [20]. For example, the authors in [19] have studied UAV-mounted BSs to support dynamic rerouting for reconfigurable backhauls operating over mmWaves bands. In [20], the authors proposed a beamforming technique to support flexible coverage within target zones by exploiting mmWave-enabled UAV transmissions. Moreover, the availability of the line-of-sight (LOS) component in UAV-based systems is suitable for mmWaves-based communications aiming to reach high gains [21], [22].

On the other hand, non-orthogonal multiple access (NOMA) downlink protocol is an efficient multiplexing candidate approach for the UAV-MBS to utilize to satisfy the connectivity and transfer speeds requirements set forth for B5G and 6G wireless systems. Contrary to conventional orthogonal multiple access, NOMA-based transmissions are preferred because they have been demonstrated to offer better overall performance through stacking the data of multiple receiving devices (RDs) using a unified resource block (RB) design, wherein jointly-multiplexed devices would enjoy a larger transmission bandwidth as well as more frequent scheduling [23], [24], [25], [26], [27], [28]. Data frames conveying information of multiplexed receivers are sent over the unified RB at varying levels of transmission power to enable each device to successfully recover its own intended data by applying successive interference cancellation (SIC) to sequentially retrieve then remove the messages within the received NOMA stack until it extracts its intended message signal [29]. Energy-efficient planning of the dispatched UAV-MBS flying course throughout the entire communication period is imperative so that the UAV's battery use is optimized. In addition, it is of critical importance to optimize the continuous adaptation of various allocated power portions within a maximum allowable budget of available transmission power as well as the dynamic activation of the receiving devices to reap as much of the promised performance of NOMA-based operation as possible [30], [31]. Moreover, making proper choices regarding the selection of appropriate receiving devices to add to a certain NOMA message stack is important to attain boosted sum-rate levels [32]. In addition, each time the UAV moves position, power allocation and device pairing need to be re-optimized, resulting in a surge in complexity and energy consumption. This is not acceptable especially when coupled with the UAV's limited battery life.

The aforementioned challenges have not been sufficiently addressed to the best of our knowledge. Recently, reinforcement learning (RL)-based methods have been attracting the attention of the research community due to their effectiveness and inherent flexibility in dealing with highly dynamic sequential decision problems. In this paper, we present two proposed algorithms based on the powerful RL framework to address the joint issue of energy-efficient dynamic UAV-MBS path design, receiving device activation, and transmit power distribution for high-speed NOMA-UAV-based downlink wireless communications. In particular, multiarmed bandit (MAB) and double deep Q-network (DDQN) RL agents are employed to leverage their highly adaptable nature to handle various dynamic and complex models. For the proposed MAB approach, we consider the two variants: minimax optimal stochastic strategy (MOSS) and upper confidence bound (UCB), for their simple yet effective deployment. For the DDQN-based algorithm, the RL agent training can be carried out offline where the agent engages in multiple interactions with the UAV-NOMA environment model before it is dispatched for operational deployment. The DDQN RL agent learns an effective, deep neural network (DNN)-based strategy and uses it to determine the appropriate projections of the environment's subsequent states onto a series of decisions yielding high returns in the long term. On the other hand, no DNNs are incorporated in the MAB-based approach, which is deployed directly to make on-the-fly online decisions while aiming to attain adequate performance in terms of the achievable total data rate level through the dynamic selection of various allowable actions according to their varying levels of some appropriate fitness criteria to determine their effective utilities. The utility of making various decisions are updated continuously over the communication time horizon.

The main contributions of this work are:

- We propose two RL-based schemes for energy-efficient UAV trajectory course planning and joint downlink NOMA power allocation and receiver selection. Both schemes are operated within battery-constrained UAV-NOMA environments with dynamic wireless channels.
- The DDQN RL agent is trained to absorb the underlying characteristics of the UAV-NOMA environment within the DNN it uses to implement its action-selection policy. We define the appropriate UAV-NOMA states, actions, and rewards so that the trained agent can achieve energy-efficient, near-optimal sum-rate performance when deployed for operation.
- The MAB-based agent is configurable with either CS-UCB or CS-MOSS operation modes and can learn to quickly converge to a highly-rewarding long-term operation policy by scanning the search space while balancing the exploration-vs-exploitation issue through the dynamic evaluation of various arms' utilities. The agent makes on-the-fly decisions and update them as needed.
- We operate and test the proposed solutions within mmWave-enabled UAV-NOMA environments having LOS signals of variable strength.
- The proposed DDQN and CS-MOSS RL agents reached 96% and 91.5% of the ergodic sum-rate provided by

the exhaustive solver whereas the conventional channel state sorting pairing (CSSP) solver reached 89.3%. Moreover, considerable energy savings of at least 91% were achieved by the proposed agents while supporting the same transfer speed as CSSP.

II. RELATED WORK

Pairing NOMA receiving devices optimally for downlink transmissions generally requires a complete scan exhausting all possible groupings. However the computational burden of such a brute force approach is huge which deems the application of the optimal solver unrealistic from a practical standpoint [32]. Numerous strategies were devised for the appropriate grouping of receiving devices over unified NOMA RBs. Famous benchmarks include the random grouping algorithm (RGA), which samples the action space using a uniform random decisioning strategy, and channel-state grouping (CSG) [33], wherein strong nodes (i.e., devices with boosted channel conditions) are grouped with weaker nodes (i.e., devices with attenuated channel conditions). The authors in [34] developed a technique leveraging unsupervised learning for node clustering wherein they developed an algorithm based on expectation maximization by harnessing spatial correlative patterns among different nodes to solve the device grouping problem.

Although optimized distribution of the power available for signal transmission over receivers multiplexed on a given RB can be attained for power-domain NOMA systems using the technique outlined in [35], optimized operation results can be accomplished only by considering the joint problem of device grouping and power distribution, a taxing NP-hard problem [31]. Upon handling the problem in [31], the authors devised an allocation scheme that operates suboptimally by constructing a correlation structure for the downlink channels, and then deploying difference of convex optimization to distribute the available power. With a focus on reducing the amount of power consumption, the work in [36] leveraged the sparse nature associated with NOMA power distribution to formulate a convex-relaxed version of the power distribution and user grouping problem. However, the presented technique incurs high computational demand to solve the formulated string of problems. A relaxed version based on l_1-norm characterization is formed in [30] for the joint problem of power distribution and device grouping, wherein the authors applied a solution method based on compressive sensing.

Although the above body of research work handles the problem of NOMA user grouping and power distribution for a variety of conventional wireless communication scenarios, it fails to accommodate less common yet important mission-critical scenarios such as emergency-oriented communications wherein an effective and reliable transmission system which can be quickly deployed is needed. For example, in [37], the authors considered a cloud computing environment where computation tasks of mobile users are offloaded to a moving UAV, with an energy minimization objective with QoS constraints. However, NOMA user pairing and power splitting optimization was not considered. In [38], a joint optimization of course planning, device scheduling, and ground users' transmission power is accomplished for an uplink NOMA data collection system to minimize the total flight time of the UAV. However, downlink NOMA sum-rate maximization and UAV energy consumption optimization were not considered. In [39], the authors presented a joint UAV trajectory planning and user scheduling algorithm for downlink NOMA sum-rate maximization. However, dynamic NOMA transmit power splitting and UAV energy optimization were not considered. Other possible use cases include dispatching UAV-mounted mobile BSs to blackout sites where receiving devices are disconnected from the main servicing infrastructure. To incorporate this need, an emergency network UAV-aided framework is developed in [8] for operation in disaster zones. The scheduling and trajectory of UAV-MBSs are firstly designed to support wireless coverage to the receiving devices on the ground. Afterwards, to expand the UAV-MBS wireless service domain, the authors formed a ground-based multi-hop D2D system and studied the UAV-MBS transceiver design. However, the generic system presented in [8] does not account for NOMA resource management optimization. To address this point, the authors in [9] established a UAV-assisted framework for NOMA-based emergency communications. The proposed scheme started by establishing a UAV-active uplink line to collect information relevant to the IoT devices within the areas under emergency operation. Subsequently, to support coverage for IoT users, a joint power management and UAV dispatching scheme is proposed. However, downlink NOMA user grouping is not considered. To handle this issue, the work presented in [40] combines both NOMA power distribution and user grouping with UAV-MBS path design and optimizes the operation jointly with a sumrate-maximization objective in mind. However, energy-efficient operation is not guaranteed since battery-aware design is not considered.

In this paper, we consider energy-driven design of the joint problem of dynamic UAV-MBS path planning and downlink NOMA user grouping and power association where battery-constrained operation is taken into account and optimized routing is accomplished through the deployment of the proposed RL-based frameworks where the RL agents are aiming to maximize the total rate while operating in a battery-constrained mode for energy-efficient UAV-MBS deployment.

III. SYSTEM MODEL & PROBLEM FORMULATION

Consider a downlink UAV-MBS-based system operating via NOMA protocol to offload information data to a group of wireless receiving devices as illustrated in Fig. 1. Application scenarios for such a UAV-based communication system include deployment in emergency cases (e.g., environments affected by a natural disaster [8], [9]) where ground users are experiencing connection issues due to temporary damage of the nearby BS as depicted in Fig. 2. In this case, the

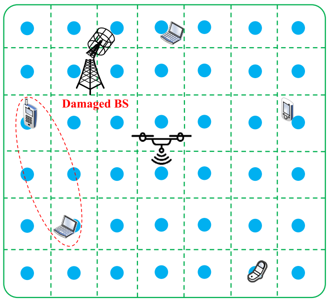


FIGURE 1. NOMA-UAV-MBS system model.

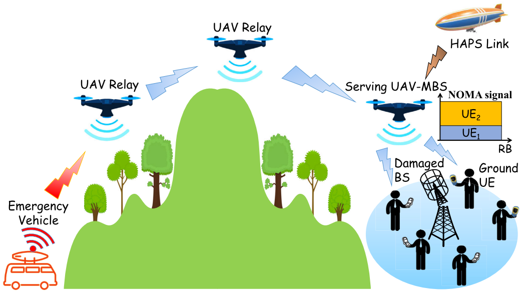


FIGURE 2. NOMA-UAV-MBS emergency environment.

serving UAV-MBS can offload emergency communications data to ground users via a multi-hop UAV relay chain or by utilizing HAPS link communications. Another practical use case is to establish temporary hotspot communication links for suburban or rural environments [41]. These non-terrestrial based use cases are an integral part of B5G and 6G networks [2], [42]. Initially, the transmitting UAV-MBS begins at some arbitrary position (e.g., center of the flying zone). The coverage area is split into multiple small regions of potential spots for hovering, with the blue circles marking the hovering positions' centers. The receiving devices are scattered arbitrarily throughout the flying zone covered by the UAV-MBS. To optimize the rate of transferring data, the UAV-MBS transmitter must carefully select the receiving devices to activate and offload data to, as well as allocate the available transmit power efficiently. In addition, energy-efficient dynamic course planning is important to position the UAV-MBS in an optimized manner across all the allowable hovering spots so as to accommodate and account for the evolving nature of the various wireless links while avoiding rapid depletion of the UAV's battery. In this work, we assume that the UAV-MBS is capable of predetermining its own location within the flying zone.[1] We also assume that the channel-state information (CSI) is available at the UAV-MBS side, which can be accomplished via CSI broadcasting by the ground users over control channels. It is worth mentioning that the operation of the proposed agents does not require the UAV mobile BS to have explicit knowledge of the ground users' locations. CSI availability is sufficient for the proposed RL agents to operate. Alternatively, if the reward feedback information is readily available, the MAB agent can estimate the proper action and navigate the zone accordingly. Since downlink NOMA protocol [23] is utilized, the information offloaded to the activated receiving devices is sent over unified RBs where the corresponding total rate attained by the selected devices may be expressed as

$$R = \sum_{d=1}^{D} W \log_2\left(1 + \frac{p_d g_d}{\sum_{i=1}^{d-1} p_i g_d + \sigma^2}\right), \tag{1}$$

where, without loss of generality, we assume a descending-order CSI (gain information) between the UAV-MBS and the receiving devices: $g_1 > g_2 > \cdots > g_D$. The channel gain, g_d, of the d-th activated device is based on the small-scale Rician fading in Eq. (3) and the distance-dependent large-scale effects in Eq. (18). W represents the available transmission bandwidth, and σ^2 represents the variance of the zero-mean additive white Gaussian noise. p_d is the portion of the transmission power that the UAV-MBS allocates to the signal of the d-th selected receiving device. Under the assumption of a total number of K candidate receiving devices, we have $d \in \{1, 2, \ldots, D\}$, where $D < K$ represents the number of information streams offloaded via the downlink NOMA multiplexing protocol.

We assume that the UAV's battery has a finite capacity of χ energy units (EUs) and that the battery energy level decreases linearly as a function of the traveled distance according to

$$L(i) = \chi - \sum_{j=1}^{i} \eta Z(j) - E_h, \tag{2}$$

where $L(i)$ denotes the UAV's battery level at time step i and $Z(j)$ is the distance traveled at step j, $j = 1, 2, \ldots, i$. η is the energy expense per unit distance. The second term represents the energy consumed to move the UAV around the flying zone. E_h is the hovering energy consumed to keep the UAV in air. Although the UAV will consume energy to remain in air during the entire communication period, this consumed energy will be common among competing solvers for a given service duration. Therefore, the surplus energy consumed to move the UAV around within the coverage zone will be the main differentiating factor for optimizing the total energy consumption. The hovering energy will then represent a minimum common threshold for the energy consumed, and thus we focus on the extra energy required to move the UAV throughout the flying zone since this can differ from one solver to another based on the traversed flight trajectories. Moreover, the energy consumed due to wireless transmission during the service period will be the same for all solvers and was consequently not included in the formulation of the energy consumption that we aim to conserve. A fixed transmission power budget, P_t, is customarily considered, and the maximum achievable sum-rate is pursued. Although such linear energy consumption models are simple, they can

[1] This is typically accomplished through a global navigation satellite system (GNSS). However, if the GNSS is unavailable or more precise location information is needed, then other techniques may be used (e.g., integrating ultra-wideband (UWB) technology with LiDAR-based range finders) [43].

provide a fairly accurate representation of the energy consumed to move the UAV within the flying zone and have been verified empirically in the literature [44], [45], [46], [47].

We assume that the communication channel's wireless link between the UAV-MBS transmitter and the d-th candidate receiving device is given by a Rician channel representation to model the presence of the LOS signal. Therefore, the channel link connecting receiving device d to the UAV-MBS may be written as

$$\hat{g}_d = \sqrt{\frac{F_r}{F_r+1}}\bar{g}_d + \sqrt{\frac{1}{F_r+1}}\tilde{g}_d\,, \tag{3}$$

where $\bar{g}_d$ denotes the LOS deterministic component which is set to a typical value of 1 [41]. $\tilde{g}_d$ represents an NLOS random component that follows the Rayleigh distribution. F_r denotes the Rician channel parameter.

Dynamic spot selection for UAV-MBS adaptive hovering throughout the flying zone is of paramount importance to properly tune the effective wireless links connecting the receiving devices to the UAV-MBS in a way that maximizes the collective acquired rate without expending the UAV's battery inefficiently. By changing the UAV-MBS location, the wireless channel can then be controlled to combat the interference and thus boost the effective SINR level at the receiving devices, and consequently improve the sum-rate [21], [48], [49]. Adjusting the UAV hovering location within the service area is therefore important for the proposed scheme to mitigate the effects of interference among multiplexed NOMA users to maximize the achievable sum-rate level. Consider for example a two-user NOMA downlink transmission with SIC detection: the normalized sum-rate is $log_2(1+\frac{p_1g_1}{\sigma^2}) + log_2(1+\frac{p_2g_2}{p_1g_2+\sigma^2})$. Here, the channel gains, $g_d\,, d = 1, 2$, are time-varying and depend on the Tx-Rx separation distance. Therefore, by changing the UAV Tx location, the channel gains can be adjusted to maximize the sum-rate level. Moreover, changing the UAV's location can be useful for situations where blockers are present in the environment around stationary users. In this case, the UAV can change its location to achieve a LOS link to the ground user and provide better connectivity. In addition, careful activation of receiving devices as well as proper splitting of the UAV-MBS transmission power among the selected devices must be adjusted dynamically to provide and sustain high operational performance. We can therefore formulate the objective problem as,

$$\max_{X,Y,\mathfrak{D},p_d} \frac{\mathbb{E}\left[\sum_{d=1}^{D} W \log_2\left(1+\frac{p_d g_d}{\sum_{i=1}^{d-1} p_i g_d+\sigma^2}\right)\right]}{\sum_{j=1}^{\mathcal{H}} \eta Z(j)} \tag{4}$$

$$\text{s.t.} \quad \left(p_d - \sum_{u=1}^{d-1} p_u\right)\frac{g_{d-1}}{\sigma^2} \geq \mu \quad \forall d \in \mathfrak{D}, \tag{c1}$$

$$\sum_{d=1}^{D} p_d = P_t, \tag{c2}$$

$$E_h\mathcal{H} + \sum_{j=1}^{\mathcal{H}} \eta Z(j) < \chi, \tag{c3}$$

$$X_{min} \leq X \leq X_{max},\ \ Y_{min} \leq Y \leq Y_{max}, \tag{c4}$$

$$\mathfrak{D} \subset \mathcal{K}, \tag{c5}$$

where the ergodic sum-rate in Eq. (4) is maximized with an optimized UAV energy consumption over the service period horizon, $\mathcal{H}$. X and Y are the bounded coordinates of the serving UAV-MBS. $\mathfrak{D}$ is the ordered subset of activated users from the candidate set, $\mathcal{K}$. Constraint (c1) ensures reliable SIC detection of NOMA signals and is described in detail later in Eq. (13). Constraint (c2) ensures the allocated NOMA powers conform to the available transmit power budget of the serving UAV-MBS. Constraint (c3) restricts the consumed energy within the UAV's battery capacity. Constraint (c4) ensures the serving UAV-MBS will remain within the boundaries of the coverage zone. Constraint (c5) ensures the activated NOMA users are selected from the candidate user set.

IV. PROPOSED ALGORITHMS

This section presents RL solution methods based on CS-MAB and DDQN to address the joint problem outlined earlier in section III. Although both methods share the same underlying agent-environment interaction principle, they have distinctive features and operate on different basic concepts: DDQN agents rely on DNNs to represent their decision-making policy and can provide near-optimal performance if environment-related information is collected adequately during offline training sessions. MAB agents, on the other hand, do not employ DNNs, and make on-the-fly decisions through direct online deployment where they adjust their decisions dynamically according to the rewards received during a series of successive interactions with the environment. Since no DNNs are present, MAB-based operation is generally less complex and simpler to implement than its DDQN-based counterpart. It may, however, provide less optimized performance than DDQN. It is worth mentioning that these solving agents will reside at the UAV-MBS side. Such AI-native operation of the communication system is in line with the design principles of future B5G and 6G networks [42].

A. DEEP RL-BASED OPERATION: Q NETWORK METHOD

Algorithm 1 outlines the proposed DDQN solution for the joint UAV-MBS path design and NOMA device activation and power allocation scheme. The fundamental operating premise is to develop a successful strategy to transfer sequential input environmental variables to highly rewarding decisions over the long run, which are reflected in the cumulative rates acquired by active receiving devices. The algorithm leverages the main Q-learning concept that maps a given action $\mathcal{A}$ to a fitness Q-value when taken in a state $\mathcal{L}$ by following some policy π according to

$$V_q^{\pi}(\mathcal{L},\mathcal{A}) = E[r_1 + \beta r_2 + \beta^2 r_3 + \ldots \,|\, \mathcal{A}_0 = \mathcal{A},\ \mathcal{L}_0 = \mathcal{L}]\,, \tag{5}$$

where $V_q^{\pi}(\mathcal{L}, \mathcal{A})$ denotes the mean discounted summation of rewards acquired in a long-term sense. These rewards are earned when, starting at some arbitrary state $\mathcal{L}_0 = \mathcal{L}$, action $\mathcal{A}_0 = \mathcal{A}$ is applied, then the subsequent state-action path is dictated in accordance with the policy π. r_1 denotes the immediately-acquired reward and $r_l \, \forall \, l > 1$ represents the rewards acquired subsequently during future states. $\beta \in [0, 1]$ is a parameter to adjust the amount of discounting to apply to balance r_1 with future rewards.

A DDQN is used to provide $V_q(\mathcal{L}, \mathcal{A}; \Psi_i)$, a configurable parametric implementation of the value function which can be tuned to generate a good approximation of the optimal function $V_q^{*}(\mathcal{L}, \mathcal{A})$. A DNN is employed to store Ψ_i, the parameter set used to generate the optimal strategy. DDQN RL agents employ a separate set of parameters $\bar{\Psi}_i$ to generate Y_i which provides a target for the training of Ψ_i, the main strategy set for the i-th agent-environment interactive exchange. The learnable target is estimated as

$$Y_i = r_i + \beta \hat{V}_q(\mathcal{L}_{i+1}, \arg\max_{\mathcal{A}} V_q(\mathcal{L}_{i+1}, \mathcal{A}; \Psi_i); \bar{\Psi}_i) \,, \tag{6}$$

where the main DNN parameter set Ψ_i is reserved for decision making whereas the assessment of the corresponding fitness value is estimated through the target DNN, $\bar{\Psi}_i$. $\hat{V}_q(.)$ represents the response generated by the target DNN.

1) NOMA-UAV-MBS: STATE & ACTION SPACES, AND REWARDS

The state defining the environment of the NOMA-UAV-MBS system is formed for the i-th interactive step as

$$\mathcal{L}_i = \{g_1(i), g_2(i), \cdots, g_K(i), L(i)\} \,, \tag{7}$$

where $g_k(i)$ and $L(i)$ represent the channel gain of the k-th candidate and the UAV's battery level at time step i, respectively.

The action $\mathcal{A}_i$ applied by the proposed DDQN agent to state $\mathcal{L}_i$ is formed as

$$\begin{aligned} \mathcal{A}_i = \{ & X(i), Y(i), s_1(i), s_2(i), \cdots, s_K(i), \\ & p_1(i), p_2(i), \cdots, p_K(i) \} \,, \end{aligned} \tag{8}$$

where $X(i)$ and $Y(i)$ are the coordinates of the chosen UAV hovering spot at the i-th step, whereas $s_k(i) = 1$ is a binary indicator denoting the selection of the k-th device and $p_k(i)$ represents its corresponding transmission power portion allocated by the UAV-MBS.

The immediate reward resulting from applying $\mathcal{A}_i$ in the i-th interaction step to the environment can be expressed as

$$r_i = \sum_{k=1}^{K} r_k(i) \,, \tag{9}$$

where the reward contribution of the k-th activated device is evaluated as

$$r_k(i) = \log_2\left(1 + \Gamma_k(i)\right) \,, \tag{10}$$

Algorithm 1 Proposed DDQN Agent for Joint UAV-MBS Path Design and NOMA Device Activation and Power Allocation.

```
– Set $P_t$, $\chi$, $g$, $\mu$, $E$, $\eta$, $I$, $\delta$, $\alpha_s$, $\alpha_f$, $\gamma$, $\alpha_d$
– Initialize $\Psi$, $\bar{\Psi} = \Psi$, $d = 1$, $\alpha_1 = \alpha_s$, $\mathcal{M} = \text{NULL}$
– While $e \leq E$  run the following episode:
    • Set  $i = 1$, $L(i) = \chi$
    • Initialize NOMA-UAV-MBS state $\mathcal{L}_i$
    • While $i \leq I$
        1) Draw a random sample $s$ from a uniform distribution $U(0, 1)$:
           If $s \leq \alpha_i \Rightarrow$ Pick decision $\mathcal{A}_i$ randomly. Else, $\mathcal{A}_i = \arg\max_{\mathcal{A}} V_q(\mathcal{L}_i, \mathcal{A}; \Psi)$.
        2) If $\mathcal{A}_i$ does not conform to the QoS requirement (13):
            ∗ For $k \in \mathcal{V}_u$, $\mathcal{V}_u \equiv$ set of violating devices, enforce (13) on $\mathcal{A}_i$ by
              rectifying the allocated power portion for each active device in $\mathcal{V}_u$:
                  $p_k = \frac{\mu\sigma^2}{g_{k-1}} + \sum_{u=1}^{k-1} p_u$
            ∗ Normalize Tx power level of all active devices:
                  $p_d \leftarrow \frac{p_d}{\sum_{u=1}^{D} p_u} P_t$
        3) Execute $\mathcal{A}_i$ in the environment, then monitor its resulting state $\mathcal{L}_{i+1}$
           and the UAV's battery level $L(i)$, and acquire the generated reward $r_i$
        4) Append $(\mathcal{L}_i, \mathcal{A}_i, r_i, \mathcal{L}_{i+1})$, the experience gathered through interaction, to the memory unit $\mathcal{M}$
        5) Randomly pick an experience mini-batch $(\mathcal{L}_j, \mathcal{A}_j, r_j, \mathcal{L}_{j+1})$ from the memory $\mathcal{M}$:
            ∗ Set $A^* = \arg\max_{\mathcal{A}} V_q(\mathcal{L}_{j+1}, \mathcal{A}; \Psi)$
            ∗ Form the agent's training target as
                  $Y_j = r_j + 1[\mathcal{L}_j \neq \textit{Terminal}]\beta\hat{V}_q(\mathcal{L}_{j+1}, A^*; \bar{\Psi})$
            ∗ Take a single gradient descent step on
                  $\sum_j (Y_j - V_q(\mathcal{L}_j, \mathcal{A}_j; \Psi))^2$ w.r.t $\Psi$
        6) If $i Mod \delta = 0$:Adjust $\bar{\Psi}$ softly using a smoothing parameter $\gamma$:
                  $\bar{\Psi} \leftarrow \gamma\Psi + (1 - \gamma)\bar{\Psi}, \; 0 < \gamma < 1$
        7) If $\alpha_f < \alpha_i$:Reduce $\alpha_i$ further to approach the final level $\alpha_f$ through the
           annealing factor $\alpha_d$:
                  $\alpha_{i+1} = \alpha_i(1 - \alpha_d), 0 < \alpha_d < 1$
        8) If $L(i) \leq (1 - g)\chi$ (i.e., UAV's battery is depleted):
            ∗ Penalize the reward for battery draining:
                  $r_i = r_i - \rho$
            ∗ Mark state $\mathcal{L}_i$ as Terminal, and set $\alpha_1 = \alpha_i$.
            ∗ End Episode
        9) Increment iteration index: $i \leftarrow i + 1$
    • end While
    • Configure probability of randomized decisioning for the upcoming episode:
                  $\alpha_1 = \alpha_I$
    • Advance episode: $e \leftarrow e + 1$
– end While
where
    • 1[.] is the indicator function. $E$ is the number of episodes and $I$ is the maximum number of iterations within an episode.
    • $\alpha_s$ is the initial probability of randomized decisioning, whereas $\alpha_f$ represents the final probability of picking an action in a random fashion during advanced interactions.
    • $\rho$ is the penalty for battery draining and $\delta$ is a configurable controller to adjust the target network's update interval.
```

with the ratio of desired signal power to the collective power of interference and noise (SINR) evaluated as

$$\Gamma_k(i) = \frac{s_k(i)p_k(i)g_k(i)}{\sum_{u=1}^{k-1} s_u(i)p_u(i)g_k(i) + \sigma^2}. \tag{11}$$

The β-discounted total reward accumulated throughout an interaction horizon I may therefore be expressed as

$$R(I) = E[r_1 + \beta r_2 + \beta^2 r_3 + \cdots + \beta^{I-1} r_I]. \tag{12}$$

The proposed RL operation in Algorithm 1 is designed so that the DDQN agent interacts continuously for E episodes, each comprising up to I interactions with the NOMA-UAV-MBS environment. The proposed agent aims to learn a successful strategy that sequentially projects the environment's states to a series of actions that maximize the total long-run reward in (12). At the start, two identical random instantiations of the parameter sets Ψ and $\bar{\Psi}$ are generated for the main and target DNN-based policies. The NOMA-UAV-MBS environment is then subjected to a uniformly distributed random action with probability α_s. The decision is made with probability $1 - \alpha_s$ depending on the existing DNN strategy, Ψ. Subsequently, the SIC-related QoS requirement for proper detection,

$$\left(p_d - \sum_{u=1}^{d-1} p_u\right) \frac{g_{d-1}}{\sigma^2} \geq \mu\,, \tag{13}$$

can be checked and, if necessary, enforced whenever the action is not conforming to the condition. p_d and μ respectively denote the power portion allocated for active receiving device d, $d = 2, 3, \ldots, D$, and a reliable detection threshold for the SIC operation. The sum-rate immediate reward as well as the next resultant NOMA-UAV-MBS system state can then be buffered into an experience-gathering memory unit $\mathcal{M}$ which collects important information relevant to the agent's ongoing interaction with the environment. The collected information helps the agent to form a concrete set of experiences which can be progressively fused and harnessed to update the agent's acting policy Ψ during each round of interaction. The buffering memory unit, $\mathcal{M}$, is propagated with training data objects taking the form of 4-tuple items each consisting of a possible state of the environment, an associated action performed during that state, a resultant subsequent state, and the reward collected from that interaction. A mini-batch of random data items is fetched from the experience memory unit for tuning and adjustment of the main policy DNN toward the agent's learnable target, Y_j. In *double* DQN operation, each training data item within the mini-batch is used to compute a corresponding learnable value over a two-step process: firstly, the training item's next state information is passed through the main DNN, Ψ, to find the action with the associated highest Q-value. Secondly, the training item's immediate reward component is combined with the output of the secondary DNN, $\bar{\Psi}$, which corresponds to the action selected in the first step. Although the main DNN parameter set is updated at each interaction, the critic's network, $\bar{\Psi}$, is smoothly updated in a periodic fashion every δ iterations. To accomplish this soft update, a fractional smoothing factor, $\gamma \in (0, 1)$, is used to fuse the updated parameter set of the main DNN with the current parameter set of the target DNN.

To optimize the operation for battery-aware decisioning, at each interactive iteration, the agent's inspect the status of the UAV's battery to determine whether to prematurely halt the ongoing episode. If the UAV's battery is drained (i.e., $L(i) \leq (1-g)\chi$, g is the battery drain percentage), the agent's reward is discounted by a configurable penalty parameter, ρ. The agent then saves the current value of the probability of making its decision on a random basis, α_i, to be used as the starting value, α_1, for the next episode, then the current episode is abruptly terminated, thus giving the agent an incentive towards deciding in favor of actions that, in the long run, do not drain the UAV's battery rapidly. Otherwise, towards the end of the interaction iteration, if the current probability of executing a random decision is greater than a preconfigured minimum end value, α_f, then the probability is reduced through a controllable decay parameter, α_d. This probability reduction mechanism is gradual and lasts for an amount of iterations controlled through α_d to allow the DDQN agent to build more confidence to follow its developing internal strategy, Ψ, more frequently while it absorbs more knowledge of the underlying characteristics of the environment it is interacting with. Once the probability reduction process halts, the agent subsequently maintains a fixed level of α_f for sampling decisions randomly throughout all remaining interactions. It is worth mentioning that, from a practical perspective, while the inference of the deployed agent's policy consumes additional energy, it is overwhelmingly eclipsed by the energy consumed to keep flying the UAV around within the service zone.

B. MULTIARMED BANDIT-BASED SOLUTION

Algorithm 2 outlines the proposed CS-MAB solution for the joint UAV-MBS path design (deciding which coordinates to choose for subsequent hovering positions within the defined grid for UAV-MBS operation) and NOMA device activation and power allocation scheme. The algorithm's deployment can be implemented via either CS-MOSS or CS-UCB options. The UAV-MBS player starts initially at some arbitrary state of the environment and pulls a decision arm then loops successively over all the allowable decision arms in a number of interactive steps in order to form initial crude estimates for the various utilities of playing different decision arms in accordance with Eq. (14). During this initialization phase, a separate reward jar is dedicated to collect the achievements of each decision arm. Each jar's initial value is set to the immediate reward acquired by executing the corresponding decision arm in the environment. As for the DDQN case, Eq. (9) reflects how much reward is generated for a given arm play. Each decision arm $\mathcal{A}_i$ is a three-tuple action object containing the hovering position coordinates (X,Y) of the UAV-MBS, along with the UAV-MBS transmit power portions $(p_1, p_2, \ldots p_K)$, and a set of active receiving devices

Algorithm 2 Proposed Cost-Subsidized MAB Operation for Joint UAV-MBS Path Design and NOMA Device Activation and Power Allocation.

```
– Set $P_t$, $\lambda$, $\mu$, $\chi$, $\eta$, $\mathcal{H}$
– Initialize $i = 1$, NOMA-UAV-MBS state $\mathcal{L}_i$
– While $i \leq \mathcal{H}$
  • If $i \leq \mathcal{A}_{max}$
    1) Pull decision arm $\mathcal{A}_i$
    2) If $\mathcal{A}_i$ does not conform to the QoS requirement (13):
       ∗ For $k \in \mathcal{V}_u$, $\mathcal{V}_u \equiv$ set of violating devices, enforce (13)
         on $\mathcal{A}_i$ by rectifying the allocated power portion for each
         active device in $\mathcal{V}_u$:
           $$p_k = \frac{\mu\sigma^2}{g_{k-1}} + \sum_{u=1}^{k-1} p_u$$
       ∗ Normalize power level of all active devices:
           $$p_d \leftarrow \frac{p_d}{\sum_{u=1}^{D} p_u} P_t$$
    3) Set up a reward jar for current arm: $m_i = r_i$
    4) Initialize a pull counter for current arm: $n_i = 1$
    5) Evaluate the utility of pulling current arm:
           $$f_i = \begin{cases} m_i + \sqrt{2\log(i)}, & \textbf{CS}-\textbf{UCB} \\ m_i + \sqrt{\max(\log(i), 0)}, & \textbf{CS}-\textbf{MOSS} \end{cases} \qquad (14)$$
  • Otherwise
    1) Set $k = \arg\max_j f_j$
    2) Form a candidate subset of decision arms:
           $$\Omega(i) = \left\{ j : f_j \geq (1-\lambda) f_k \right\} \qquad (15)$$
    3) Pull battery-aware arm $\mathcal{A}_{i^*}$:
           $$i^* = \arg\max_{j \in \Omega(i)} L(j) \qquad (16)$$
    4) If $\mathcal{A}_{i^*}$ does not conform to (13):
       ∗ For $k \in \mathcal{V}_u$, enforce (13) on $\mathcal{A}_{i^*}$ by rectifying the allocated
         power portion for each active device in $\mathcal{V}_u$:
           $$p_k = \frac{\mu\sigma^2}{g_{k-1}} + \sum_{u=1}^{k-1} p_u$$
       ∗ Normalize power level of all active devices:
           $$p_d \leftarrow \frac{p_d}{\sum_{u=1}^{D} p_u} P_t$$
    5) Update reward jar of $\mathcal{A}_{i^*}$: $m_{i^*} \leftarrow m_{i^*} + r_i$
    6) Increment corresponding counter: $n_{i^*} \leftarrow n_{i^*} + 1$
    7) Update the utility corresponding to playing $\mathcal{A}_{i^*}$:
           $$f_{i^*} = \begin{cases} \frac{m_{i^*}}{n_{i^*}} + \sqrt{\frac{2\log(i)}{n_{i^*}}}, & \textbf{CS}-\textbf{UCB} \\ \frac{m_{i^*}}{n_{i^*}} + \sqrt{\frac{\max\left(\log\left(\frac{i}{n_{i^*}}\right), 0\right)}{n_{i^*}}}, & \textbf{CS}-\textbf{MOSS} \end{cases} \qquad (17)$$
  • Move to new NOMA-UAV-MBS state, $\mathcal{L}_{i+1}$.
  • Increment iteration index: $i \leftarrow i + 1$
– end While
where
  • The horizon, $\mathcal{H}$, is the total number of played arms.
  • $\mathcal{A}_{max}$ denotes the number of arms available to the agent.
  • $\lambda$ is the cost-subsidizing control parameter.
```

as defined in Eq. (8). Moreover, a set of counters is initialized to record the frequencies of pulling various decision arms in later stages. During each interaction, and before executing the chosen decision arm on the NOMA-UAV-MBS system, if the chosen decision arm is not conforming to the data detection fidelity requirement in (13), the agent enforces the condition by boosting the power portions associated with the violating receiving devices and applying power normalization to maintain the feasibility of transmission power budget.

As soon as the initialization phase terminates, dynamic selection of decision arms is accomplished by harnessing the available information of various arm-pulling utilities as well as the UAV-MBS battery level, which are updated dynamically in every interactive iteration in accordance with Eqs. (17), and (16) and (2), respectively. In particular, a feasible subset of candidate receiving devices exceeding a configurable QoS threshold on utility is formed according to Eq. (15) where a recommended cost-subsidizing factor of $\lambda = 0.1$ is used [50]. Next, to control the battery energy consumption and facilitate for a battery-aware operation, the feasible decision arm resulting in the current highest battery level is played in the environment as dictated by Eq. (16). The corresponding generated reward is then added to the associated reward jar of the played arm. In addition, the corresponding counter of the played arm is incremented by one, and the associated utility value is updated. The state of the environment advances subsequently to a new state and another iteration of interaction begins. The CS-MAB algorithm keeps on interacting with the environment for a predefined horizon, $\mathcal{H}$.

Numerous hyper-parameters govern the process of training a DDQN RL agent and, generally speaking, high complexity is associated with the required computations. This concern may nonetheless be bypassed if computations are offloaded to a prior stage of offline training wherein the DDQN RL agent's skills are honed through interactive training on the possibly-simulated environment. Afterwards, online operation is commenced and the experienced agent is deployed. The proposed DDQN agent (with A layers) has a deployment complexity of $\mathcal{O}(vAK)$ for K candidate devices within an area of v hovering positions. On the other hand, the deployment complexity of the CS-MAB agent is $\mathcal{O}(vK)$. By contrast, in optimal operation, UAV-MBS placement and the grouping of D devices for active operation over an allocated RB would incur $\mathcal{O}\left(v\binom{K}{D}\right)$, thereby demanding a far greater implementation cost.

V. NUMERICAL ANALYSIS

A. SIMULATION ENVIRONMENT

The settings of the simulated environment are given in Table 1. The table presents the default simulation values of the used settings for the evaluated scenarios. We assume a data-offloading NOMA-UAV-MBS downlink system where 5 candidate receiving devices are scattered arbitrarily within a 100 × 100 m^2 2-D zone. Various grid sizes ranging from less than 100-by-100 m^2 and going well beyond 100-by-100 m^2 have been considered in the literature. For instance,

TABLE 1. Simulation settings.

Scheme	Setting	Value
Common settings	mmWave Carrier frequency	60 GHz
	System bandwidth, W	100 MHz
	Channel exponent (pathloss), υ	2.1
	Standard deviation for Shadowing, ζ	4.4 dB
	Type of wireless channel	Rician
	Rician channel parameter, F_r	10 dB
	Noise spectral density	-174 dBm/Hz
	Grid zone size	100×100 m^2
	Grid zone spacing	10 m
	UAV-MBS battery capacity, χ	1 EU
	UAV-MBS energy parameter, η	10^{-4} EU
	UAV-MBS transmit power, P_t	20 dBm
Proposed DDQN solution	Episodes	100
	Episode interactive iterations	100
	DNN optimizer method	SGDM
	Learning rate	0.001
	Secondary DNN update period, δ	4
	Gradual-update softening parameter, γ	0.001
	Capacity of experience memory unit	5000
	Mini-batch training items	8
	Randomized initial action probability, α_s	1
	Randomized action probability end-level, α_f	0.01
	Probability decaying parameter, α_d	0.005
	Battery draining penalty, ρ	100

the authors in [51] considered a grid size of 30-by-30 m^2 whereas the authors in [21] considered a 200-by-200 grid. Similarly, the authors in [9] considered a 200-by-200 emergency communications area with four deployed UAVs. In our work, we consider an in-between 100-by-100 area which can be suitable for spots not covered yet by the main network's infrastructure (e.g., in remote areas) or for zones suffering network instability such as in areas hit by disasters thus temporarily rendering the main network out-of-service. It is worth mentioning that other grid sizes can be used depending on the application. All receiving devices employ antenna elements at a fixed 1 m height. Initially, and at a fixed 10 m height, the UAV-MBS begins at the (0,0) hovering point within the flying region. The carrier setting for the mmWave is configured to 60 GHz. The system bandwidth setting is 100 MHz. We also consider the close-in model (CI),

$$PL_{CI}^{dB}(f,Z) = PL_{FS}^{dB}(f,Z_0) + 10\upsilon\log_{10}\left(\frac{Z}{Z_0}\right) + \Upsilon_{CI}^{\zeta}, \tag{18}$$

for large-scale channel disturbances. The used channel exponent setting for pathloss is $\upsilon = 2.1$ (typical for LOS propagation of mmWaves in urban environments) [52]. Taking such mmWave signal propagation models into account is important since B5G and 6G will heavily leverage the broad spectrum offered by mmWaves. The model simulates large-scale fluctuations which is combined with small-scale fading of Eq. (3) to account for overall variation of the wireless channel. The channel gain of the d-th candidate receiving device is thus $PL_{CI}^{dB}+20\log_{10}\left|\hat{g}_d\right|$ dB. The standard deviation, ζ, for CI model emulated shadowing, Υ_{CI}^{ζ}, is configured as 4.4 dB. PL_{FS}^{dB} is the nominal pathloss for free-space in dB. Since it delivers appropriate model accuracy and maintains parameter stability both for outdoor as well as indoor urban environments (including micro and macro variants) spanning a broad frequency spectrum within microwave and mmWave bands, a $Z_0 = 1$ m reference distance is utilized for typical CI models. A 1 m reference point might be crossing the boundaries of the near-field emitted by massive antennae arrays. However, the inaccuracy introduced by such small distance is mostly trivial from the perspective of practical wireless communication systems [52], [53].

The UAV-MBS operates at a 20-dBm default transmission power level to offload the NOMA message relaying the data of active receiving devices. The simulated AWGN noise is generated using the typical -174 dBm/Hz for its PSD level.

The primary and secondary DNNs employed in the RL-based DDQN operation are built with a total of 5 fully-connected layers. Layers 1 through 4 have rectifying linear units (ReLUs) as non-linear activation functions: $A(x) = \max(0, x)$. The fifth layer (output layer) has the linear characteristic $A(x) = x$. The structure of the two sets of neurons configured for both DNNs is identical: 100 neurons within each of layer 1 and 5, whereas the remaining three layers in the middle consist of 95,90, and 85 neurons respectively. Moreover, each neuron in both DDNs is configured with an adjustable bias term.

DNN training for updating the main parameter set Ψ at each interactive iteration is accomplished by performing a single optimization step using stochastic gradient descent with momentum operation (SGDM optimizer) [54] towards an energy-efficient and highly rewarding mapping strategy from input environmental states to appropriate decisions that the agent can apply. The associated learning step size is configured as 0.001. The critic's secondary DNN set, $\bar{\Psi}$, is only periodically updated every fourth interactive step. The gradual update of $\bar{\Psi}$ is performed in a smooth fashion using a soft mixture of $\bar{\Psi}$ and Ψ with a 10^{-3} softening parameter. A limit of five thousand items is set on the size of the memory tank buffering experience data items collected by the DDQN agent over successive interactive steps.

The end policy for the trained DDQN agent is acquired by progressively updating the main DNN set Ψ, using an 8-item mini-batch (which is composed through random sampling from the previously stored experience items) during each update step. The DDQN agent's training starts with a 100% chance of choosing a random decision (with uniform selection distribution among the decision set). The agent gradually shifts away from this randomized decisioning by annealing its probability over successive interactive iterations until the chance of operating in a random fashion reaches a preset terminal level of 0.01. By setting the decaying parameter $\alpha_d = 5 \times 10^{-3}$, this probability decay process runs for slightly over 900 interactive iterations before it halts when the terminal level α_f is attained. The training phase comprises a 100 episodes in total where each one runs for up to a 100 interactive iterations.

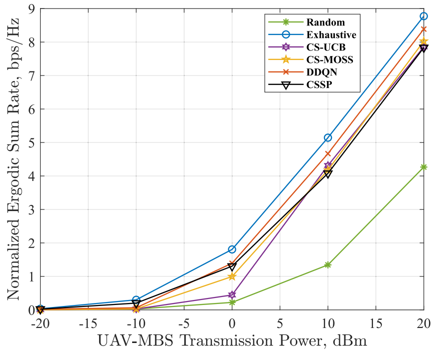
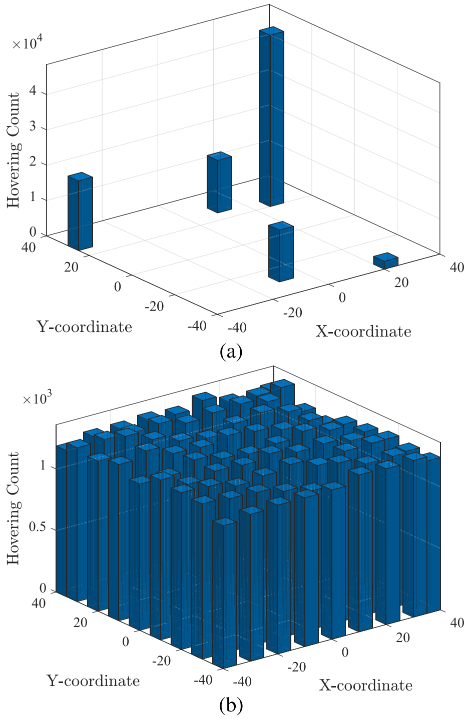


FIGURE 3. Evaluation of the performance in terms of the acquired ergodic sum-rate over 100, 000 simulated interactions.

B. RESULTS AND DISCUSSION

In Fig. 3 we evaluate the performance of the proposed algorithms in terms of the normalized ergodic total rate with the total transmission power of the UAV-MBS ranging from −20 dBm to 20 dBm. Upper and lower performance benchmark references are respectively represented by the exhaustive and random solution strategies. In addition, to show the performance gain compared with existing NOMA user paring approaches that maximize the achievable sum-rate, conventional CSSP resource allocation reference baseline [33] is included with optimal UAV-MBS positioning. In CSSP, candidate receiving devices are arranged based on their channel state quality conditions wherein the devices experiencing larger gaps in their channel gains are paired. Although UAV energy efficiency is an important aspect that we geared the proposed RL-based algorithms to tackle, the main objective of this work is the maximization of the achievable ergodic sum-rate of downlink NOMA. Therefore, we compare with the conventional CSSP benchmark along with the exhaustive strategy for the upper performance bound. Clearly, the optimal exhaustive approach can support the highest total rate level in a consistent manner by virtue of exploring every available point within the action space before applying the most-rewarding alternative. On the other hand, random-based operation runs by choosing to execute some randomly sampled alternative from the available decision set, thereby resulting in a heavily non-optimized and weak performance, which mainly functions as an indicator of achievable levels for the total rate. Coming on top of the proposed algorithms is the DDQN solution where it manages to remain within a close performance gap from the upper level provided by the exhaustive scan. Beginning with extremely low levels of UAV-MBS transmission power, the proposed algorithms produce performance results close to those achieved by a random approach, with CSSP taking the lead until around −4 dBm where the proposed DDQN first overtakes CSSP. As the UAV-MBS transmission power is increased, noticeable distinct performance gaps of the achievable sum-rate then begins to emerge until the DDQN agent rises to slightly over 77% of the exhaustive at 0 dBm as opposed to 72% for CSSP, with the CS-MOSS agent following next at about 55% whereas the CS-UCB agent falls behind around the 25% mark. The proposed CS-MAB agents then quickly rise to surpass CSSP starting around 8 dBm. All schemes then keep on rising until the DDQN agent manages to attain almost 96% of the sum-rate level achievable by an exhaustive scan at 20 dBm, whereas CS-MOSS rises to 91.5% and CS-UCB follows closely after at 89.5% which is immediately followed by CSSP at 89.3%. It is worth mentioning that CS-UCB surpassing CS-MOSS at around 7.5 dBm can be attributed to the fact that these two settings of the CS-MAB agent employ different utilities in Eq. (17) of Algorithm 2 to balance exploitation of previously gained knowledge with the exploration of different available arms in the action space. This can consequently lead to the formation of different candidate subsets of decision arms in Eq. (15), which may in turn be subject to further rectification in step 4. The intricate dynamics at play here will ultimately manifest as a sum-rate performance increase in favor of one setting with respect to the other which can evidently vary depending on the operating power regime.



(a)

(b)

FIGURE 4. UAV hovering distribution for 100, 000 simulated interactions for (a) Exhaustive search, and (b) Random selection strategy.

Figure 4 illustrates the distribution of UAV hovering spot selection based on an exhaustive search for sum-rate maximization in part (a) and a random selection strategy in part (b). Figure 4 (a) shows the critical positions within the flying zone that the UAV-MBS visits for hovering during a

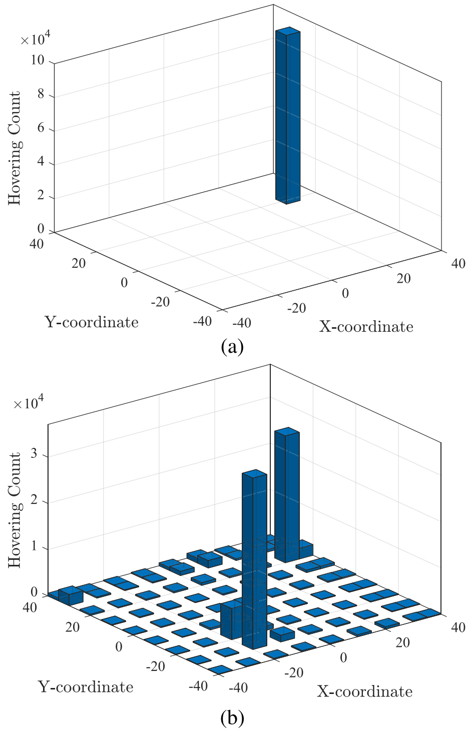


FIGURE 5. **UAV hovering distribution for 100, 000 simulated interactions for the proposed (a) DDQN RL agent, and (b) CS-MAB-UCB RL agent.**

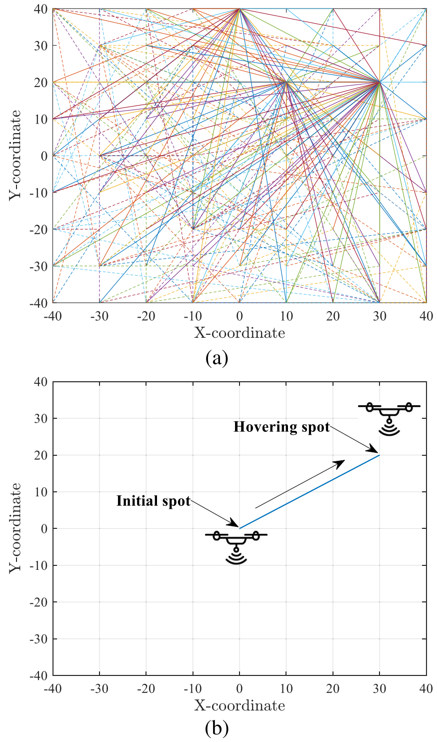


FIGURE 6. **Path traced by the UAV for DDQN-based operation, for (a) Training, and (b) Deployment: DDQN agent decides to fly over to the marked spot where it maintains position during the service period.**

series of 10^5 interactive steps while following an optimal solution policy as given by an exhaustive search covering the entirety of the action space. This intensive scanning of the grid filters the ineffective hovering spots while leaving intact the five positions represented by the blue bars due to their optimized positioning relative to device scatter pattern. These five critical hovering locations are chosen exclusively by the UAV-MBS to attain the transmission rates of Figs. 3 and 10. Despite the fact that all of the 5 locations are preferred and chosen by the optimal strategy, their associated visiting frequencies are not the same as indicated by the difference in their corresponding bar heights, and the UAV-MBS predominantly converges to the position (30,20) for hovering. Furthermore, in part (b) we can observe the non-optimized behavior of the random strategy where, as expected, all spots are selected with almost equal visiting frequencies.

Figure 5 illustrates in part (a) the distribution of UAV hovering spot selection based on a trained DDQN RL agent, whereas in part (b) the CS-UCB agent's selection strategy is presented. For the entire deployment stage, the trained DDQN agent decides to maintain a fixed position at the spot most frequently visited by the optimal scanning as indicated by the presence of a singular blue prism at the (30,20) grid point. On the other hand, the CS-UCB agent switches back and forth between the same spot and the spot at $(-20,-30)$ while occasionally visiting other positions with less relative frequency as indicated by the blue bars of varying heights in part (b).

To get a complementary view of the DDQN operation, the trajectories traced by the UAV during both training and deployment are shown in Fig. 6. In part (a), each line in this figure represents a traced path segment connecting a departure point to a destination point where the UAV traverses the path connecting the two spots as it makes successive decisions while interacting with the environment. The colors represent different traversed path segments for different agent-environment interactions. The agent learns by exploring various hovering positions to absorb the essential features of the environment prior to deployment. Noticeably, the agent goes through the three spots at (30,20), (10,20), and (0,40) more often than it routes through other spots on the grid. Upon deployment in part (b), the trained DDQN agent opts to route directly from the (0,0) starting point to the critical position found at (30,20) where it maintains position as discussed earlier in part (a) of Fig. 5.

The energy efficiency of the UAV movement is presented in Fig. 7 for CSSP and the proposed RL-based schemes where the Y-axis represents the total energy consumption level per normalized ergodic sum-rate. As shown, an energy efficiency level of about 0.44 EU per bps/Hz is achieved by the DDQN solution and is lower than those achieved by both CS-UCB and CS-MOSS solutions. The CS-MOSS solution comes second in line at around 1.31 EU per bps/Hz, thus requiring just a little below 200% higher energy to support the same total transfer speed provided when operating using the DDQN solution. The CS-UCB solution consumes about 1.4 EU per bps/Hz which sets it at around 220% and 7.5% behind the DDQN and CS-MOSS solutions, respectively. Lastly, to maintain adequate total rate levels, CSSP incurs a significant energy loss in comparison where it consumes over 11 times higher energy than CS-UCB. Operating in such energy-efficient modes is essential for meeting the high expectations set forth for 6G networks where a minimum of 10-fold improvement is expected in terms of energy efficiency [55]. This demonstrates the effectiveness of the proposed RL-based approaches.

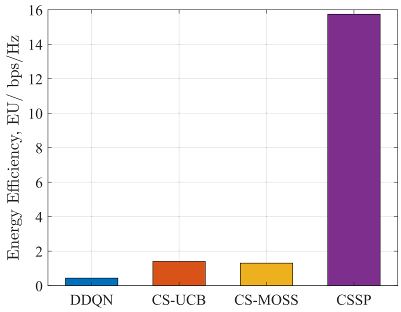


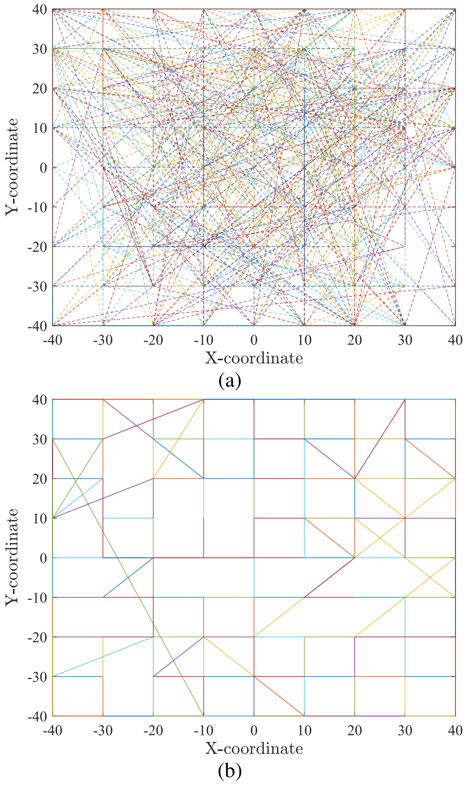
FIGURE 7. UAV movement energy efficiency of the proposed RL-based algorithms evaluated using 100, 000 simulated interactions.

The trajectory traced by the UAV is shown in Fig. 8 for CS-MAB-based operation with and without battery optimization. In part (a), the subsidizing factor of the deployed CS-UCB agent is configured to $\lambda = 0$ whereas in part (b) a recommended typical value of $\lambda = 0.1$ is used [50]. When the operation is not optimized for energy efficiency, the CS-UCB agent keeps rerouting through all the spots on the grid extremely intensively without any signs of filtering down the traced path. On the other hand, when battery optimization is turned on in part (b), the CS-UCB agent prunes the traversed path significantly and flies along a much smaller subset of path segments compared to the operation illustrated in part (a) when battery usage optimization is turned off. This clearly shows the effectiveness of cost-subsidizing operation when a MAB solution is deployed.

Figure 9 illustrates the effect of more noise accommodation as the system bandwidth increases (for a fixed transmit power level of 20 dBm) and the resulting impact on the training performance of the proposed DDGN agent. As expected, since the reward is based on the normalized sum-rate, increasing the system bandwidth allows in more noise and leads consequently to less accumulated episode reward for the interacting agent. Nonetheless, the proposed agent learns a stable, high-reward policy in all cases where the bandwidth is varied from 100 to 500 MHz. Specifically, when operating over a 100 MHz link, training rewards as high as 8.53 bps/Hz are achieved. The training reward then dips down to near 7.5 bps/Hz over 200-MHz links. Switching the operation to the 500-MHz bandwidth results in episode rewards below the 6 bps/Hz level. The figure also tracks the five-episode reward moving average level as the agent's training progresses towards a mature high-reward policy. This demonstrates the successful training operation of the proposed DDGN agent for various noise levels.

To verify the ability of our proposed solutions to support proportionally increasing total rate versus a variable transmit bandwidth range, Fig. 10 demonstrates the total achievable rate corresponding to 100 ∼ 500 MHz Tx bandwidths at 60-GHz NOMA mmWave carrier. The main purpose of Fig. 10 is to validate that the RL algorithm's performance will not deviate significantly from the optimal solution as the system bandwidth is increased and more pronounced noise effect is allowed. This is particularly important for communication systems operating over the mmWave frequencies where large bandwidths are commonly used. The linear scaling by the bandwidth serves to emphasis the fact that larger bandwidths can accommodate higher speeds holds for downlink NOMA. When operating at a hundred MHz, all solutions result in sub-Giga data transfer speeds. All approaches then rapidly exceed 1 Gbps except for the random strategy which grows very slowly towards the 1-Gpbs level where it breaks it around the 500 MHz mark at which point the proposed methods manage to support rates beyond 2.5 Gbps. The DDQN agent comes on top of the proposed RL solutions where it achieves speeds as high as 2.92 Gbps while enjoying the full 500 MHz of system bandwidth. By contrast, CS-MOSS attains 2.81 Gbps whereas CS-UCB reaches 2.68 Gbps when utilizing the full bandwidth. CSSP trails behind and achieves 2.57 Gbps at the same 500-MHz point. Nonetheless, the optimal exhaustive scan can evidently support even faster transmission rates going well beyond 3 Gbps by running through all permissible decisions.

The capability of the proposed algorithms to learn successful adaptations to changing LOS circumstances within the wireless links of the used communication channels is validated in Fig. 11 where a comparative analysis of the DDQN agent versus cost-subsidized MAB methods is presented. The main purpose of Fig. 11 is to ensure that the proposed schemes will maintain stably high performance around the default quiescent point and will remain resilient to perturbations in the wireless environment. Therefore, only the proposed schemes are presented in this figure (simulation results at the default point defined in Table 1 are already presented previously for all schemes; in Fig. 3 for example). The strength of the simulated LOS channel component is swept by adjusting the Rician parameter LOS control over the range $-40 \sim 40$ dB. Both DDQN and CS-MAB solutions show similar trends. When operating over non-LOS-dominated links (corresponding to $-40 \sim -10$ dB), the response exhibited by the DDQN agent is flat at around 7.7 bps/Hz. Similarly, CS-MOSS operates around 7.34 bps/Hz whereas CS-UCB provides 7.17 bps/Hz over the same region. Next, the total rates of both DDQN and CS-MAB solutions begin to accelerate as operation is shifted towards channels with more inherent presence of the LOS component as illustrated in Fig. 11 for the $-10 \sim 10$ dB



(a)

(b)

FIGURE 8. Path traced by the UAV for MAB-based operation, for (a) No battery optimization: λ = 0 CS-UCB, and (b) With battery optimization: λ = 0.1 CS-UCB.

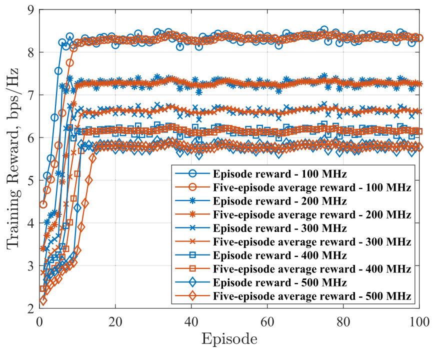


FIGURE 9. Tracking the training performance of the proposed DDQN agent over a 100 episodes in terms of the acquired episode reward and the five-episode reward moving average for varying system bandwidth levels.

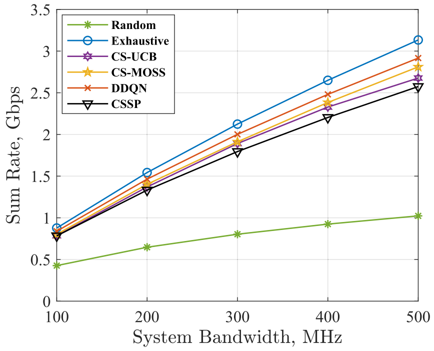


FIGURE 10. Evaluation of the deployment performance in terms of the ergodic sum-rate acquired at variable system bandwidth levels with 100, 000 simulations.

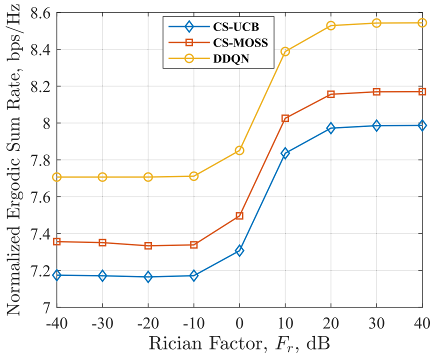


FIGURE 11. Tracking the maximum deployment performance of the DDQN agent vs cost-subsidized MAB for varying levels of LOS component presence.

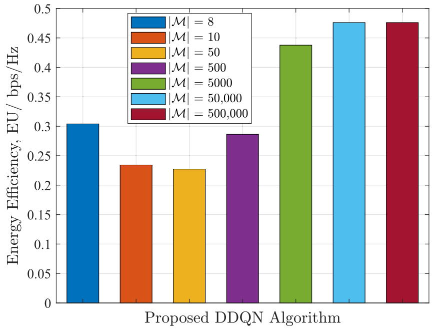


FIGURE 12. Tracking the performance of the proposed DDQN agent for various memory unit sizes.

region. As shown for the region 10 $\sim$ 40 dB, the achieved levels of the total ergodic rate that can be supported by the proposed algorithms saturate eventually when operation is heavily geared toward LOS-dominated links where, at the end point of 40 dB, the DDQN solution provides 8.54 bps/Hz, whereas CS-MOSS supports 8.17 and CS-UCB comes next at 8 bps/Hz.

To investigate the sensitivity of the proposed DDQN algorithm to varying buffering unit sizes, we study its performance in Fig. 12. Starting at the lowest possible capacity of $|\mathcal{M}| = 8$ (i.e., same as training batch size), the achieved normalized energy consumption is about 0.304 EU per bps/Hz. As expected, when the unit capacity is increased beyond the batch size, the performance is improved where 0.234 and 0.227 EU per bps/Hz are obtained at $|\mathcal{M}| = 10$ and $|\mathcal{M}| = 50$, respectively. As the unit capacity is increased further, more datapoints from the action space are gathered and retained while training the agent. This leads to the exploration of more paths and consequently drives the total energy consumption up, where 0.286 and 0.438 EU per bps/Hz are reached at $|\mathcal{M}| = 500$ and $|\mathcal{M}| = 5000$, respectively. As the size of the unit is increased even further, the number of new potentially improving exploratory points dwindles, which drives the consumption up only slightly where it saturates at around 0.476 EU per bps/Hz when $|\mathcal{M}| = 500,000$. In all cases, the proposed DDQN algorithm manages to achieve highly efficient normalized energy consumption performance which demonstrates its effectiveness.

VI. CONCLUSION

In this article, we have conducted an investigative study on the utilization of CS-MAB as well as DDQN agents as viable RL-based data-offloading solutions for emergency use cases deploying ready-to-dispatch UAV-MBSs for NOMA-based downlink transmissions. The DDQN agent's training was accomplished in an offline stage wherein the agent engages with the UAV-MBS-NOMA environment in a multi-iteration interactive mode prior to operational deployment. CS-MAB agents on the other hand were directly operated online as they do not utilize DNNs to require a training stage. Due to its tailored ability to resolve highly complex dynamic sequential decision problems, the proposed RL DDQN approach succeeded in supporting an energy-efficient near-optimal total rate level consistently in various battery-constrained transmission scenarios, whereas the proposed cost-subsidized MAB-based approaches followed closely after. Both proposed approaches have been tested via operation in mmWave-enabled propagation modes with varying dominance levels of the LOS Rician channel component. This is of particular importance to B5G and 6G networks where higher spectral and energy efficiencies are targeted. Both CS-MAB and DDQN solutions exhibited accelerated performance over links with strong LOS presence where they respectively supported as high an ergodic total rate level as 8.17 bps/Hz and 8.54 bps/Hz. We tackled the joint dynamic UAV-MBS trajectory design and NOMA transmit power splitting and receiving device activation problem to adequately support ready-to-deploy energy-efficient solutions accommodating increased transfer speeds breaking beyond 2.5 Gbps which may prove critical importance for deployment in emergency cases with high-speed data-offloading demands as in the regions afflicted by disasters. Future extension to multi-UAV scenarios will be considered where collision avoidance mechanisms will be employed.

REFERENCES

[1] W. Feng et al., "Joint 3D trajectory design and time allocation for UAV-enabled wireless power transfer networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9265–9278, Sep. 2020.

[2] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2021.

[3] K. David and H. Berndt, "6G vision and requirements: Is there any need for beyond 5G?" *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 72–80, Sep. 2018.

[4] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward tactile internet in beyond 5G era: Recent advances, current issues, and future directions," *IEEE Access*, vol. 8, pp. 56948–56991, 2020.

[5] M. Asad, S. Qaisar, and A. Basit, "Client-centric access device selection for heterogeneous QoS requirements in beyond 5G IoT networks," *IEEE Access*, vol. 8, pp. 219820–219836, 2020.

[6] R. Ali, Y. B. Zikria, A. K. Bashir, S. Garg, and H. S. Kim, "URLLC for 5G and beyond: Requirements, enabling incumbent technologies and network intelligence," *IEEE Access*, vol. 9, pp. 67064–67095, 2021.

[7] M. M. D. Silva and J. Guerreiro, "On the 5G and beyond," *Appl. Sci.*, vol. 10, no. 20, p. 7091, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/20/7091

[8] N. Zhao et al., "UAV-assisted emergency networks in disasters," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 45–51, Feb. 2019.

[9] W. Feng et al., "NOMA-based UAV-aided networks for emergency communications," *China Commun.*, vol. 17, no. 11, pp. 54–66, Nov. 2020.

[10] K. G. Panda, S. Das, D. Sen, and W. Arif, "Design and deployment of UAV-aided post-disaster emergency network," *IEEE Access*, vol. 7, pp. 102985–102999, 2019.

[11] M. Liu, J. Yang, and G. Gui, "DSF-NOMA: UAV-assisted emergency communication technology in a heterogeneous Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5508–5519, Jun. 2019.

[12] Z. Huang, C. Chen, and M. Pan, "Multiobjective UAV path planning for emergency information collection and transmission," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6993–7009, Aug. 2020.

[13] M. Y. Arafat and S. Moh, "Localization and clustering based on swarm intelligence in UAV networks for emergency communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8958–8976, Oct. 2019.

[14] G. Peng, Y. Xia, X. Zhang, and L. Bai, "UAV-aided networks for emergency communications in areas with unevenly distributed users," *J. Commun. Inf. Netw.*, vol. 3, no. 4, pp. 23–32, Dec. 2018.

[15] T. Zhang, J. Lei, Y. Liu, C. Feng, and A. Nallanathan, "Trajectory optimization for UAV emergency communication with limited user equipment energy: A safe-DQN approach," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1236–1247, Sep. 2021.

[16] W. Feng et al., "UAV-enabled SWIPT in IoT networks for emergency communications," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 140–147, Oct. 2020.

[17] S. K. Datta, J.-L. Dugelay, and C. Bonnet, "IoT based UAV platform for emergency services," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2018, pp. 144–147.

[18] P. Boccardo, F. Chiabrando, F. Dutto, F. Tonolo, and A. Lingua, "UAV deployment exercise for mapping purposes: Evaluation of emergency response applications," *Sensors*, vol. 15, no. 7, pp. 15717–15737, Jul. 2015. [Online]. Available: https://www.mdpi.com/1424-8220/15/7/15717

[19] M. Gapeyenko, V. Petrov, D. Moltchanov, S. Andreev, N. Himayat, and Y. Koucheryavy, "Flexible and reliable UAV-assisted backhaul operation in 5G mmWave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2486–2496, Nov. 2018.

[20] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "3-D beamforming for flexible coverage in millimeter-wave UAV communications," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 837–840, Jun. 2019.

[21] Z. Xiao, H. Dong, L. Bai, D. O. Wu, and X.-G. Xia, "Unmanned aerial vehicle base station (UAV-BS) deployment with millimeter-wave beamforming," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1336–1349, Feb. 2020.

[22] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 66–73, May 2016.

[23] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 611–615.

[24] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance of downlink non-orthogonal multiple access (NOMA) under various environments," in *Proc. VTC (Spring)*, May 2015, pp. 1–5.

[25] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, Oct. 2017.

[26] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[27] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[28] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, Jan. 2020.

[29] A. H. Gendia, M. Elsabrouty, and A. A. Emran, "Cooperative multi-relay non-orthogonal multiple access for downlink transmission in 5G communication systems," in *Proc. Wireless Days*, Mar. 2017, pp. 89–94.

[30] A. Nasser, O. Muta, H. Gacanin, and M. Elsabrouty, "Joint user pairing and power allocation with compressive sensing in NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 151–155, Jan. 2021.

[31] S. Zhang, N. Zhang, G. Kang, and Z. Liu, "Energy and spectrum efficient power allocation with NOMA in downlink HetNets," *Phys. Commun.*, vol. 31, pp. 121–132, Dec. 2018.

[32] A. Gendia, O. Muta, and A. Nasser, "Cache-enabled reinforcement learning scheme for power allocation and user selection in opportunistic downlink NOMA transmissions," *IEEJ Trans. Electr. Electron. Eng.*, vol. 17, no. 5, pp. 722–731, May 2022.

[33] H. Zhang, D.-K. Zhang, W.-X. Meng, and C. Li, "User pairing algorithm with SIC in non-orthogonal multiple access system," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[34] J. Ren, Z. Wang, M. Xu, F. Fang, and Z. Ding, "An EM-based user clustering method in non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8422–8434, Dec. 2019.

[35] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, Jul. 2017.

[36] Z. Yang, C. Pan, W. Xu, and M. Chen, "Compressive sensing-based user clustering for downlink NOMA systems with decoding power," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 660–664, May 2018.

[37] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

[38] W. Wang, N. Zhao, L. Chen, X. Liu, Y. Chen, and D. Niyato, "UAV-assisted time-efficient data collection via uplink NOMA," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7851–7863, Nov. 2021.

[39] X. Wu, Z. Wei, Z. Cheng, and X. Zhang, "Joint optimization of UAV trajectory and user scheduling based on NOMA technology," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[40] A. Gendia, O. Muta, S. Hashima, and K. Hatano, "UAV positioning with joint NOMA power allocation and receiver node activation," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2022, pp. 240–245.

[41] X. Mu, Y. Liu, L. Guo, J. Lin, and H. V. Poor, "Intelligent reflecting surface enhanced multi-UAV NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3051–3066, Oct. 2021.

[42] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[43] F. Orjales, J. Losada-Pita, A. Paz-Lopez, and Á. Deibe, "Towards precise positioning and movement of UAVs for near-wall tasks in GNSS-denied environments," *Sensors*, vol. 21, no. 6, p. 2194, Mar. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2194

[44] H. V. Abeywickrama, B. A. Jayawickrama, Y. He, and E. Dutkiewicz, "Comprehensive energy consumption model for unmanned aerial vehicles, based on empirical studies of battery performance," *IEEE Access*, vol. 6, pp. 58383–58394, 2018.

[45] S. Ahmed, A. Mohamed, K. Harras, M. Kholief, and S. Mesbah, "Energy efficient path planning techniques for UAV-based systems with space discretization," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2016, pp. 1–6.

[46] H. V. Abeywickrama, B. A. Jayawickrama, Y. He, and E. Dutkiewicz, "Empirical power consumption model for UAVs," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–5.

[47] H. Huang, A. V. Savkin, and C. Huang, "Reliable path planning for drone delivery using a stochastic time-dependent public transportation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4941–4950, Aug. 2021.

[48] L. Zhu, J. Zhang, Z. Xiao, and R. Schober, "Optimization of multi-UAV-BS aided millimeter-wave massive MIMO networks," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[49] Z. Xiao et al., "A survey on millimeter-wave beamforming enabled UAV communications and networking," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 557–610, 1st Quart., 2021.

[50] D. Sinha, K. A. Sankararaman, A. Kazerouni, and V. Avadhanula, "Multi-armed bandits with cost subsidy," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3016–3024.

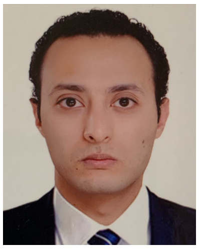
[51] E. Koyuncu, M. Shabanighazikelayeh, and H. Seferoglu, "Deployment and trajectory optimization of UAVs: A quantization theory approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8531–8546, Dec. 2018.

[52] S. Sun et al., "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2843–2860, May 2016.

[53] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.

[54] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2022.

[55] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.



AHMAD GENDIA (Member, IEEE) received the B.Sc. degree in electronics and communications engineering from Al-Azhar University, Cairo, Egypt, in 2011, the M.Sc. degree in electronics and communications engineering from Egypt-Japan University for Science and Technology (E-JUST), Alexandria, Egypt, in 2017, and the Ph.D. degree in electronics and communications engineering from Kyushu University, Fukuoka, Japan, in 2023. He worked in the automotive industry as an Embedded Software Developer at Valeo, where he participated in developing a parking-assistance system for certain VW, Audi, and Porsche vehicles. He is currently with the Department of Electrical Engineering, Al-Azhar University. His research interests include the application of machine learning to future wireless networks, OFDM systems, non-orthogonal multiple access (NOMA), UAV-based communications, RIS-assisted transmissions, cache and compute resource allocation for fog and cloud-based systems, heterogeneous networks, massive MIMO, and emerging technologies for 6G wireless networks. He was the recipient of Kyushu University 3MT Ph.D. Thesis Award in 2021. He also received the Institute of Electronics, Information, and Communication Engineering (IEICE) Radio Communication Systems (RCS) Active Research Award in 2023.

OSAMU MUTA (Member, IEEE) received an Associate B.E. degree from the Sasebo Institute of Technology in 1994, the B.E. degree from Ehime University, in 1996, the M.E. degree from the Kyushu Institute of Technology in 1998, and the Ph.D. degree from Kyushu University in 2001. In 2001, he joined the Graduate School of Information Science and Electrical Engineering, Kyushu University, as an Assistant Professor. During 2010 to 2023, he was an Associate Professor at the Center for Japan-Egypt Cooperation in Science and Technology, Kyushu University. Since 2023, he has been a Professor at the Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. His research interests include signal processing techniques for wireless communications and powerline communications, MIMO techniques, interference coordination techniques, low-power wide-area networks, and nonlinear distortion compensation techniques for high-power amplifiers. He is a Senior Member of the Institute of Electronics, Information, and Communication Engineering (IEICE). He was the recipient of the 2005 Active Research Award in IEICE Radio Communication Systems Technical Committee, the Chairperson's Award for Excellent Paper in IEICE Communication Systems Technical Committee (2014, 2015, and 2017), the 2020 IEICE Communications Society Best Paper Award, and the International Symposium on Computing and Networking 2022 (CANDAR'22) Best Paper Award, respectively.

SHERIEF HASHIMA (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in electronics and communication engineering (ECE) from Tanta and Menoufiya University, Egypt, in 2004 and 2010, respectively, and the Ph.D. degree from the Egypt-Japan University of Science & Technology (EJUST), Alexandria, Egypt, in 2014. He is a Post-Doctoral Researcher, computational learning theory team, RIKEN AIP, Japan, since July 2019. He is working as an Associate Professor at the Engineering and Scientific Equipment Department, Nuclear Research Center (NRC), Egyptian Atomic Energy Authority (EAEA), Egypt, since 2014. From January to June 2018, he was a Visiting Researcher at Center for Japan-Egypt Cooperation in Science and Technology, Kyushu University. He is a technical committee member in many international conferences and a reviewer in many international conferences, journals and transactions. His research interests include wireless communications, machine learning, online learning, 5G, B5G, 6G systems, image processing, millimeter waves, nuclear instrumentation, and the Internet of Things. He is a member of AAAI.

KOHEI HATANO received Ph.D. degree from Tokyo Institute of Technology in 2005. Currently, he is an Associate Professor with the Research and Development Division, Kyushu University Library. He is also the Leader of the Computational Learning Theory team at RIKEN AIP. His research interests include machine learning, computational learning theory, online learning and their applications.