# Q-learning aided intelligent routing with maximum utility in cognitive UAV swarm for emergency communications

Zhang, L., Ma, X., Zhuang, Z., Xu, H., Sharma, V., & Han, Z. (2022). Q-learning aided intelligent routing with maximum utility in cognitive UAV swarm for emergency communications. *IEEE Transactions on Vehicular Technology*. Advance online publication. https://doi.org/10.1109/TVT.2022.3221538

**Published in:**
IEEE Transactions on Vehicular Technology

**Document Version:**
Peer reviewed version

# $Q$-Learning Aided Intelligent Routing with Maximum Utility in Cognitive UAV Swarm for Emergency Communications

Long Zhang, *Member, IEEE,* Xiaozheng Ma, Zirui Zhuang, *Member, IEEE,* Haitao Xu, *Member, IEEE,* Vishal Sharma, *Senior Member, IEEE,* and Zhu Han, *Fellow, IEEE*

***Abstract*—This paper studies the routing problem in a cognitive unmanned aerial vehicle (UAV) swarm (CU-SWARM), which employs the cognitive radio into a swarm of UAVs within a three-layer hierarchical aerial-ground integrated network architecture for emergency communications. In particular, the flexibly converged architecture utilizes a UAV swarm and a high-altitude platform to support aerial sensing and access, respectively, over the disaster-affected areas. We develop a $Q$-learning framework to achieve the intelligent routing to maximize the utility for CU-SWARM. To characterize the reward function, we take into account both the routing metric design and the candidate UAV selection optimization. The routing metric jointly captures the achievable rate and the residual energy of UAV. Besides, under the location, arc, and direction constraints, the circular sector is modeled by properly choosing the central angle and the acceptable signal-to-noise ratio for UAV to optimize the candidate UAV selection. With this setup, we further propose a low-complexity iterative algorithm using the dynamic learning rate to update $Q$-values during the training process for achieving a fast convergence speed. Simulation results are provided to assess the potential of the $Q$-learning framework of intelligent routing as well as to verify our overall iterative algorithm via the dynamic learning rate for training procedure. Our findings reveal that the proposed algorithm converges in a few number of iterations. Furthermore, the proposed algorithm can increase the accumulated rewards, and achieve significant performance gains, as compared to the benchmark schemes.**

***Index Terms*—Emergency communications, UAV swarm, cognitive radio, intelligent routing, maximum utility, $Q$-learning.**

## I. Introduction

In times of emergencies and natural disasters, one of significant impact is the sudden and wide-scale breakdown or interruption of terrestrial communications infrastructure. For instance, cellular base stations (BSs) may be partially or totally dysfunctional or paralyzed in disaster-affected areas due to physical destructions and power outages [1]. The wireless networks might fail to provide the necessary coverage and capacity for the public and disaster responders, hindering emergency response and disaster relief operations. On the one hand, the public needs to connect to important resources or declare their safety to responders via emergency calls, text messages, or even social media updates. On the other hand, the disaster responders are dispatched to the affected region for disaster response and relief. As the first 72 hours following the disaster are the most critical, an alternative post-disaster communication system is needed to be quickly deployed for providing *disaster sensing* (e.g., situational awareness), coordinating emergency and recovery operations, as well as allowing the public to deliver massive amounts of information timely [2]. Consequently, it is highly necessary to set up emergency communications in post-disaster areas for extended, flexible, resilient, and rapidly deployable network coverage.

For reaching this goal, several technical solutions have been recognized as potential candidates for disaster and emergency scenarios, such as device-to-device [3], full-duplex [4], movable BSs [5], satellite communications [6], etc. However, the existing solutions lack efficient situational awareness over disaster areas, and are limited by complex ground conditions and uninterruptible power supplies. Particularly, some of the methods are mostly infrastructure based, depending on ground BSs to schedule the resource allocation [7]. Moreover, constrained wireless backhaul capacity and inefficient deployment of ground devices (GDs) pose new technical challenges on network design and planning of emergency communications [5]. To overcome these challenges, unmanned aerial vehicles (UAVs) can be utilized for quickly restoring communications in emergency and disaster situations [7]–[9]. Due to their high mobility and flexible deployment in the three-dimensional (3D) space, UAVs are often exploited as aerial BSs (ABSs) for temporary coverage and traffic offloading over disaster areas. For air-to-ground communications, UAVs are more likely to set up strong line-of-sight (LoS) links with GDs and normally working BSs in disasters, thus enhancing the wireless coverage and improving the system capacity [10].

Beyond that, UAVs loaded with sensors and cameras are also being used to provide the disaster responders with better disaster sensing timely, enabling them to assess situations and efficiently respond based on the aerial sensing data of the disaster areas [11]. Intuitively, the more UAVs deployed to an affected area for disaster sensing, the better response efforts could be achieved, because additional UAVs add more sensing data of such areas. Compared with a single UAV, UAV swarm has the potential to create an autonomous multi-UAV network by distributing disaster sensing tasks and coordinating operation of a large number of small UAVs [12]. Regarding the UAV swarm, the UAVs communicate with each other while in flight for self-organization and collaboration in an airborne flying ad hoc manner. Without the support of centralized infrastructure,

L. Zhang and X. Ma are with the School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China (e-mail: zhanglong@hebeu.edu.cn, xiaozhengma@hotmail.com).

Z. Zhuang is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhuangzirui@bupt.edu.cn).

H. Xu is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: alex_xuht@hotmail.com).

V. Sharma is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Northern Ireland BT9 5BN, U.K. (e-mail: V.Sharma@qub.ac.uk).

Z. Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, Republic of Korea (e-mail: zhan2@uh.edu).

each UAV connects and transmits the aerial sensing data with neighboring UAVs which are the intermediate UAVs within its communication range. Compared with the UAVs in lower altitude above the ground, high-altitude platforms (HAPs) in the stratosphere are more suitable to provide the *disaster information fusion* for emergency situations, due to larger area coverage, bigger payload capacity, and longer flight endurance [1]. Therefore, the integration of HAPs with UAV swarm motivates the prospect of implementing the hierarchical aerial sensing and access for emergency communications, which has been recognized as a key enabler for 6G systems of 2030s [1].

Despite the appealing potential of HAP-UAV hierarchical architecture for converged aerial sensing and access, such a design poses extra challenges particularly with spectrum usage for practical system design in emergency scenarios[1]. First, relying on the ad hoc mode, it will not be an effective way to use static spectrum access for deploying the UAV swarm. Second, it will not be sufficient to collect and transmit the aerial sensing data with high capacity via the pre-allocated spectrum resource for UAVs. Hence, efforts on dynamic and on-demand spectrum access for HAP-UAV hierarchical networking should be made to tackle this problem. Fortunately, the paradigm of cognitive UAV swarm (CU-SWARM) that employs cognitive radio (CR) to realize the spectrum coexistence of UAVs and cellular primary networks has gained growing interests from the research community [13]–[15]. With the built-in cognitive capabilities, the UAVs exploit dynamic spectrum access for opportunistic utilization of the licensed sub-channels over the sharing resource block (RB) held by primary users (PUs) in disasters to optimize the overall spectrum usage.

### A. Motivation and Contributions

Under the HAP-UAV hierarchical structure, when the UAVs are located outside the coverage area of an HAP serving as the disaster data fusion, it is essential for creating the *multi-hop routing* for them by choosing immediate UAVs as the relays to forward the sensing data timely, until these data finally reach the UAVs located inside the HAP coverage area. However, the inherent features of CU-SWARM, e.g., flying constraints, high mobility, dynamic spectrum access, drastically changing network topology, lack of global information, etc., also bring a number of issues to overcome particularly with the multi-hop routing design. To fully unleash the potentials of such an HAP-UAV hierarchical design, the following challenges must be well solved at the routing design so as to effectively realize the disaster information fusion for emergency situations:

1) *Dynamic mobility*: Due to the high mobility and flying constraints in CU-SWARM, the flying trajectory of UAV must be carefully identified and discussed by taking the practical mobility modelling into account when designing the routing scheme.
2) *Dynamic spectrum access with imperfect sensing*: Joint spectrum sensing (SS) and routing design is a necessary consideration in spectrum sharing CU-SWARM. Imperfect SS needs to be well captured at the sensing decision for each UAV due to the varying ground-to-air channel conditions and physical hardware impairments.
3) *Distributed implementation*: The problem of local information and limited interaction for UAVs arises due to the lack of global information, and thereby, the routing decision should be performed in a distributed manner.

Aiming to tackle these challenges for better adapting to the dynamically changing environment and distributed implementation, flexible and efficient routing is required to be carefully designed in a smarter and more agile manner. Reinforcement learning (RL), more specifically $Q$-learning, shows powerful capabilities to solve complicated decision-making problems in dynamically varying environment. Since it does not need a model of the environment, $Q$-learning is more suitable for the dynamic and partially observed external environment with low computational complexity, ease of implementation, and guaranteed convergence [16], [17]. Besides, $Q$-learning process can be executed in a distributed manner without global information, which is more in line with the inherent properties of UAV's local information and limited interaction in CU-SWARM.

Motivated by the aforementioned discussions, this paper is an attempt to address the routing problem in CU-SWARM under the HAP-UAV hierarchical structure by resorting to the $Q$-learning. We aim to design an intelligent routing framework with maximum utility through adaptive learning and intelligent decision making, while complying to the requirements of dynamically changing environment and distributed implementation. Notably, to the best knowledge of authors, this is the first trial establishing an intelligent routing framework with maximum utility in HAP-UAV hierarchical aerial-ground integrated network architecture for emergency communications. Our contributions can be summarized as follows:

- We present a three-layer hierarchical aerial-ground integrated network architecture for emergency communications over the large-scale disaster-affected areas. This is a new architectural design approach to beyond-terrestrial domain view of network design via flexibly converged aerial and terrestrial networks, including a UAV swarm and an HAP to support aerial sensing and access, respectively.
- We develop a $Q$-learning framework to achieve the intelligent routing with maximum utility for CU-SWARM, which considers the integration of CR with a swarm of UAVs via an ad hoc manner across the aerial sensing layer of the hierarchical architecture. A Gauss-Markov (GM) mobility model is employed to model the random flying trajectory for the UAV in CU-SWARM. The convergence of aerial and terrestrial networks is carefully considered by applying the underlay paradigm into the implementation of CU-SWARM, which allows the coexistence of CU-SWARM and ground cellular primary network.
- We derive the reward function of the $Q$-learning framework by designing the routing metric and optimizing the selection of candidate UAVs. We determine the routing metric by maximizing the utility, which jointly captures the achievable rate and the residual energy of UAV. The

[1]Unless otherwise stated, we use the terms HAP-UAV hierarchical architecture and three-layer hierarchical aerial-ground integrated network architecture interchangeably throughout this paper.

imperfect SS is incorporated into the sensing decision for each UAV, which affects the achievable rate of UAV pair. With the location, arc, and direction constraints, we particularly devise a circular sector over the two-dimensional (2D) rectangular area to optimize the candidate UAV selection by appropriately setting up the central angle of circular sector as well as the acceptable signal-to-noise ratio (SNR) of UAV. The dynamic learning rate is adopted for updating $Q$-values during the training process to yield a fast convergence speed.

### B. Paper Organization

The rest of the paper is organized as follows. In Section II, we introduce the related works. The network architecture is discussed in Section III. Section IV describes the system model. In Section V, the $Q$-learning framework is proposed for achieving the intelligent routing with maximum utility. The performance evaluation results are presented in Section VI. Finally, concluding remarks are provided in Section VII.

## II. Related Works

### A. Aerial-Ground Integrated Network Architecture

Incorporating aerial access networks into ground networks for realizing the aerial-ground integrated networks has drawn considerable attention from the research community recently. Majority of the existing studies focus on deploying the UAVs, also referred to as the low-altitude platforms (LAPs), to assist wireless communications and networking for GDs, through the double-layer aerial-ground integrated networks, e.g., millimeter wave communications [18], mobile edge networks [19], vehicular networks [20], etc. However, it is challenging to only deploy the UAVs at lower altitudes in sophisticated mobile environments with more performance benefits of longer flying endurance and wider area coverage.

More recently, there has been increasing interest in fusing HAPs, UAVs, and ground networks for setting up the three-layer aerial-ground integrated network architecture with different applications [21]–[25]. Ahmadinejad and Falahati in [21] designed an aerial heterogeneous wireless with an HAP and multiple UAVs serving as the quasi-stationary ABSs to provide radio access services to the GDs via the downlink orthogonal division multiple access. A non-orthogonal multiple access (NOMA)-enabled airborne access vehicular ad hoc networks (VANETs) architecture was proposed in [22] to provide reliable downlink communication services to vehicles by an HAP and several UAV relays via the decode-and-forward protocol. Apart from the downlink access services brought by the hierarchical aerial-ground network architecture in the research efforts [21], [22], some related works [23]–[25] have studied the uplink transmission or offloading as well. In [23], Qin *et al*. adopted an XAPS model with an HAP serving as a macro ABS and several LAPs serving as small ABSs to empower the clustered-NOMA systems in 6G heterogeneous Internet of Things (IoTs). Lakew *et al*. in [24] explored a heterogeneous aerial access IoT network consisting of an HAP, several UAVs, and IoT devices in underserved areas for achieving joint intelligent computation offloading and resource allocation. An HAP-aided aerial edge computing architecture was introduced in [25], where the HAP was used to execute the offloading tasks from the UAVs that were deployed to collect the data from the GDs.

The above solutions [21]–[25] have laid a solid foundation on the three-layer aerial-ground integrated network architecture. To our best knowledge, the use of UAV self-organization and collaboration to create the UAV swarm for distributing sensing tasks has not been addressed in HAP-UAV hierarchical architecture and remains an appealing study. With the flight property of cooperation and self-organization, the UAVs are more likely to send the sensing data with neighboring UAVs in a hop-by-hop manner, which motivates us to explore the multi-hop routing problem in this work. Furthermore, different from previous studies on the HAP-UAV hierarchical structure, we design a novel three-layer aerial-ground integrated network architecture by fusing the sensing in the UAV swarm and the access provided by the HAP.

### B. Routing in UAV Swarms

A large body of related works in the literature [26]–[30] have been proposed to investigate the problem of routing for UAV swarms from different perspectives. In [26], Mukherjee *et al*. proposed an offloading path selection scheme in ad hoc edge UAV swarms, where the UAVs were deployed close to the GDs to offload their computation tasks. The optimal multi-hop path through the UAVs was obtained via Multi-Armed Bandit by jointly optimizing the residual energy, the available processing power, the hop distance, and the task load of each UAV. By minimizing the computation and routing cost of the running workflows, Liu *et al*. in [27] proposed an online algorithm via Markov approximation to jointly optimize the computation offloading and multi-hop routing scheduling for UAV swarms in dynamic edge-cloud computing systems. In [28], Song *et al*. designed an enhanced flooding-based routing protocol by employing the random network coding and clustering for UAV swarms, which realized the efficient routing without any routing path discovery or network topology information. By using the particle swarm optimization, Arafat and Moh in [29] presented an energy-efficient swarm-intelligence-based clustering algorithm in UAV networks for emergency communications. The particle fitness function was defined by capturing multiple factors, such as inter-cluster distance, intra-cluster distance, residual energy, and geographic location. In [30], Li *et al*. proposed a mean field game theoretical approach for cross-layer dynamic source routing protocol in flying ad-hoc networks, by incorporating the link quality into the cross-layer cost function design. Despite the above mentioned works devoted to solving the routing problem in UAV swarms, the underlay spectrum usage for practical system design in emergency situations of interest was not considered in their studies, which may cause low efficiency in the transmission of aerial sensing data. Besides, the 3D UAV movement and simplified collision policy were assumed in [29] for the simulation setup, while the Poisson cluster process was adopted to model the random positions of UAVs in [28]. However, all of the research progress in [26]–[30] did not take the flying trajectories of UAVs into account when designing their routing schemes or algorithms.

### C. Q-Learning Aided Routing for Network Scenarios

Recent progress has been made to explore the potential of $Q$-learning in routing design for various network scenarios, e.g., CR networks [16], transportation systems [31], VANETs [32], [33], optical networks-on-chips [34], etc. Paul and Maity in [16] formulated an outage minimization problem for multi-hop routing in CR networks under multiple constraints, e.g., SS reliability, energy causality, etc. A RL-based $Q$-routing algorithm was designed to find the optimal routing, and the impact of network topologies on the runtime complexity of $Q$-routing was also examined. To solve a stochastic shortest path problem for sustainable transportation systems, Cao *et al*. in [31] developed a $Q$-learning approach, of which the converged $Q$-values was defined the actual probabilities of arriving on time. By employing the dynamic neural networks to learn the $Q$-values, the proposed method can scale well to larger road networks with arbitrary deadlines. In [32], Li *et al*. proposed a position-based hierarchical protocol known as QGrid via the RL to improve the message delivery rates with minimum latency and hop counts in VANETs. The QGrid combined both the macroscopic aspect and microscopic aspect when making its routing decision via the $Q$-value table. In [33], Luo *et al*. designed an intersection-based hierarchical routing protocol in VANETs, and designed a multidimensional $Q$-table to select the optimal road segments for packet forwarding at intersections. Zhang and Ye in [34] introduced a thermal-aware adaptive routing strategy using the tableless approximation $Q$-learning to find the optimal low-loss paths in the presence of on-chip temperature changes for optical networks-on-chips.

For the above research, only the work in [16] was focused on the CR network scenario, where the impact of PU reappearance on the amount of energy harvesting and the outage during secondary transmissions was considered. However, there was no explicit consideration for imperfect SS used for detecting the PU's transmission or non-transmission over a frequency channel, and therefore, the application of this work in practice may be limited. By contrast, in this work, we incorporate the imperfect SS into the sensing decision for each UAV, which is more in line with the practical considerations that both the varying channel conditions and physical hardware impairments may affect the sensing results. In addition, none of the aforementioned works [16], [31]–[34] considered to optimize the action selection taken by the agent for reducing the complexity of $Q$-learning algorithm when designing the reward function in their $Q$-learning frameworks. Therefore, a key motivation of this paper is to explore the candidate UAV selection strategy for optimizing the action selection taken by each agent, which is also the challenge brought by $Q$-learning.

## III. Network Architecture

We consider a hierarchical aerial-ground integrated network architecture for emergency communications over a terrestrial large-scale disaster scenario, as shown in Fig. 1. The hierarchical architecture is composed of three layers with different functionalities and properties, i.e., terrestrial layer, aerial access layer, and aerial sensing layer.

- In the **terrestrial layer**, the disaster-affected area consists of a finite number of cells that constitute a ground cellular primary network. Each cell owns a multi-antenna primary BS (PBS) that communicates with multiple associated PUs distributed in the cell coverage area via licensed orthogonal sub-channels over a spectrum RB. Here, normally working cells and dysfunctional cells coexist in the primary network since part of the PBSs become severely damaged during disasters. Partial cellular coverage will be guaranteed for a certain number of PUs.
- In the **aerial sensing layer**, a large number of *rotary-wing* UAVs equipped with onboard sensors, cameras, GPS devices, and radio transceiver modules are released and dispatched over the disaster-affected area to participate in the missions of disaster sensing. The large-scale deployment of UAVs forms a swarm of UAVs that communicate with each other while in flight for self-organization and collaboration in an airborne flying ad hoc manner. Each UAV is able to transmit the sensing data with its neighboring UAVs located within its communication range.
- In the **aerial access layer**, a quasi-stationary HAP is deployed as a standalone infrastructure for disaster information fusion. The HAP is capable of carrying communication payloads and operating at stratospheric altitude above the disaster area. To enable efficient disaster response, the HAP is connected to an emergency communications response vehicle (ECRV) located in a remote ground area via a feeder link. The HAP assumes the role of a *network gateway*, through which the sensing data are gathered from the UAVs via uplink, and where from the ECRV extracts the sensing data from the UAV swarm.

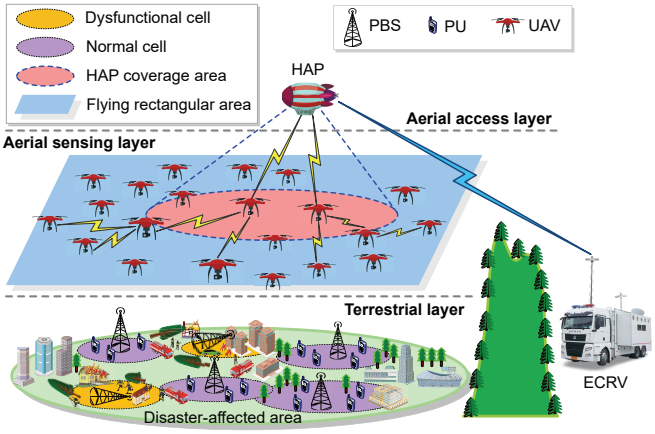



Fig. 1. Illustration of a hierarchical aerial-ground integrated network architecture for emergency communications.

In particular, we consider the integration of CR technology with the UAV swarm to implement a CU-SWARM in the aerial sensing layer for emergency communications. In this way, the overall spectrum usage can be enhanced by allowing each UAV to coexist with the PUs over the same RB. For enabling the cognitive capability, the radio transceiver of each UAV can be tuned to any licensed sub-channel of the RB. To accomplish the disaster information fusion at the aerial access layer, we have to take into account the following two cases:

- **Single-hop case**: From the HAP's point of view, by integrating with multiple antennas, it can directly collect the sensing data transmitted from the UAVs located inside

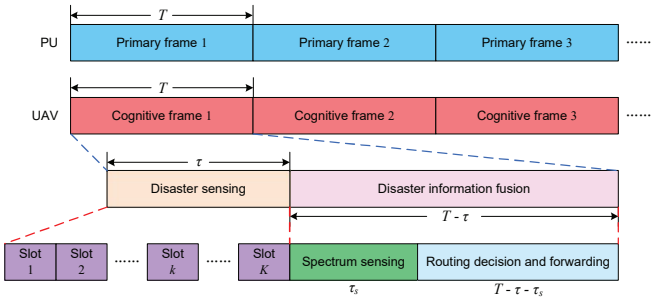



Fig. 2. The synchronized time-slotted frame structure.

the HAP coverage area via single hop.

- **Multi-hop case**: For the UAVs located outside the HAP coverage area, the multi-hop routing is required for them to employ other immediate UAVs as the relays to forward the sensing data in a hop-by-hop manner, until these data reach the UAVs located inside the HAP coverage area.

Having this in mind, we in this paper focus on the problem of multi-hop routing in CU-SWARM for identifying the optimal next-hop UAV as the relay to forward the sensing data. Under such a setup, we concentrate on an *underlay* CU-SWARM coexisting with the ground cellular primary network sharing the authorized orthogonal sub-channels over the same RB simultaneously. With the underlay paradigm, the PUs have the full privilege to access their licensed sub-channels at any time, while the UAVs are allowed to opportunistically utilize idle sub-channels unoccupied by the PUs.

## IV. System Model

For illustration convenience, we use the synchronized time-slotted frame structure to describe the coexistence of the CU-SWARM and cellular primary network, as shown in Fig. 2. The CU-SWARM operates periodically within the cognitive frame of duration $T$. Each cognitive frame consists of the disaster sensing phase with duration $\tau$ and the disaster information fusion phase with duration $T-\tau$. In the disaster sensing phase, the UAVs are dispatched to perform the disaster sensing over the disaster area of interest. For ease of exposition, the duration $\tau$ of disaster sensing phase is further discretized into $K$ equally spaced time slots with length $\delta_l = \frac{\tau}{K}$. Note that $K$ is chosen to be sufficiently large, which makes $\delta_l$ so small that the UAV's location can be considered as approximately unchanged within each time slot. After the disaster sensing phase, each UAV performs the spectrum sensing (SS), and then selects the optimal route and forwards the sensing data to the HAP in the disaster information fusion phase. Thereby, the disaster information fusion phase is divided into two subphases: the SS subphase with duration $\tau_s$, and the routing decision and forwarding subphase with duration $T-\tau-\tau_s$.

In this section, we first introduce the primary network model in Subsection IV-A, and then discuss the CU-SWARM network model in IV-B.

### A. Primary Network Model

For the primary network, the whole RB is divided into $C$ licensed sub-channels, denoted by a set $\mathcal{C} = \{1, 2, \cdots, C\}$, each having an equally-sized bandwidth of $B_P$. During each primary frame, the licensed sub-channel is either occupied by a PU or unoccupied. From the UAV's point of view, the licensed sub-channel is alternatively switching between the ON (*busy*) state and the OFF (*idle*) state. The ON state means the sub-channel is being occupied by the PU, whereas the OFF state indicates the PU is absent and the sub-channel can be freely occupied by UAV. We thus model the PU behavior over sub-channel $c$ as an independent and identically distributed alternating ON-OFF random process within each primary frame. The duration of the ON state and the OFF state on sub-channel $c$ is statistically independent of each other, and is denoted by two random variables $T^c_{\text{ON}}$ and $T^c_{\text{OFF}}$, respectively, for $c \in \mathcal{C}$. As in [35], $T^c_{\text{ON}}$ and $T^c_{\text{OFF}}$ generally follow an exponential distribution with a mean of $\mathbb{E}\left[T^c_{\text{ON}}\right] = \frac{1}{\lambda^c_1}$ and $\mathbb{E}\left[T^c_{\text{OFF}}\right] = \frac{1}{\lambda^c_0}$, respectively, and thereby have the probability density functions expressed as follows

$$T^c_{\text{ON}} \sim f^c_{\text{ON}}(\varsigma) = \lambda^c_1 e^{-\lambda^c_1 \varsigma}, \quad \varsigma \in [0, T], \tag{1}$$

$$T^c_{\text{OFF}} \sim f^c_{\text{OFF}}(\varsigma) = \lambda^c_0 e^{-\lambda^c_0 \varsigma}, \quad \varsigma \in [0, T]. \tag{2}$$

Note that distribution parameters $\lambda^c_0$ and $\lambda^c_1$ can be effectively estimated by a maximum likelihood estimator [36]. For convenience, let us define a binary random variable as follows to indicate whether sub-channel $c$ is being occupied by the PU (ON state) or not (OFF state) at time $\varsigma$, for $\varsigma \in [0, T]$, i.e.,

$$S_c(\varsigma) = \begin{cases} 1, & \text{if sub-channel } c \text{ is busy at time } \varsigma, \\ 0, & \text{if sub-channel } c \text{ is idle at time } \varsigma. \end{cases} \tag{3}$$

Thereby, the prior probabilities of sub-channel $c$ being idle or occupied by the PU at time $\varsigma$ can be respectively given by

$$\Pr\{S_c(\varsigma) = 0\} = \frac{\lambda^c_1}{\lambda^c_0 + \lambda^c_1}, \tag{4}$$

$$\Pr\{S_c(\varsigma) = 1\} = \frac{\lambda^c_0}{\lambda^c_0 + \lambda^c_1}. \tag{5}$$

We then easily have $\Pr\{S_c(\varsigma) = 0\} + \Pr\{S_c(\varsigma) = 1\} = 1$.

### B. CU-SWARM Network Model

*1) Mobility Model:* We consider the *underlay* CU-SWARM consisting of $N$ UAVs, denoted by a set $\mathcal{N} = \{1, 2, \cdots, N\}$, to undertake the disaster sensing missions in the aerial sensing layer. Without loss of generality, a 3D Cartesian coordinate system is considered. During the disaster sensing phase, each UAV is assumed to fly a random trajectory within a 2D rectangular area $\Upsilon(\mathfrak{L}_l, \mathfrak{L}_w)$ at a fixed altitude $H_u$ above the ground, following the GM mobility model[2] [37], [38]. Here, $\mathfrak{L}_l$ and $\mathfrak{L}_w$ denote the length and the width of the 2D rectangular area. The reason for using the GM mobility model is that the UAV's trajectory as for disaster responders is generally smooth and without sudden stops and sharp turns, and the flying speed and direction at the current time is always highly correlated to

[2]Note that in general, the UAV can fully take advantage of the flexible 3D mobility via trajectory design [4], [8], [9] or location planning [22], [23] for improving the system performance. In this work, we focus on a random trajectory for the UAV during disaster sensing phase, by setting up a proper altitude for terrain or building avoidance. Particularly, the UAV's utility as shown later obtained from selecting the potential next-hop UAV depends on both the residual energy of UAV and the achievable rate between UAV pair. Thus, the routing decision taken by the agent is mainly associated with the distance between UAV pair at time slot $K$. Therefore, the trajectory design or location planning for the UAV will not affect the representation of reward function in our intelligent routing framework via $Q$-learning.

the speed and direction at the previous time [38]. In practice, it can be used for random search on a specified target area, which is better suited for UAV disaster sensing in disaster-affected areas. Note that the GM mobility model can be designed to adapt to different levels of randomness for UAV via an adopted tuning parameter. We set the fixed altitude $H_u$ for each UAV due to the minimum altitude requirement for terrain or building avoidance without frequent aircraft ascending and descending over a terrestrial disaster scenario.

Through the discretization of disaster sensing phase with duration $\tau$, the trajectory of UAV $n$ at time $\varsigma$ in the 3D coordinate system can be approximated by a $K$-length sequence composed of $K$ discrete points within $\Upsilon(\mathfrak{L}_l, \mathfrak{L}_w)$, i.e., $\mathbf{q}_n(\varsigma) \approx \{(\mathbf{q}_n[k], H_u)^\tau\}_{k=1}^K$, where $\mathbf{q}_n[k] = (x_n[k], y_n[k]) \in \mathbb{R}^{2\times 1}$ refers to the horizontal location of UAV $n$ at time slot $k$, for $n \in \mathcal{N}$ and $\varsigma \in [0, \tau]$. We assume that the horizontal location $\mathbf{q}_n[K] = (x_n[K], y_n[K])$ of UAV $n$ will be kept unchanged at time slot $K$, which continues for the whole duration of the disaster information fusion phase with duration $T - \tau$. With the GM mobility model, the flying speed $v_n[k]$ and direction $\varphi_n[k]$ of UAV $n$ at time slot $k$ are calculated based on the flying speed $v_n[k-1]$ and direction $\varphi_n[k-1]$ at time slot $k-1$ using the following equations

$$v_n[k] = \beta v_n[k-1] + (1-\beta)\overline{v} + \sqrt{1-\beta^2} G_v[k-1], \quad (6)$$

$$\varphi_n[k] = \beta \varphi_n[k-1] + (1-\beta)\overline{\varphi} + \sqrt{1-\beta^2} G_\varphi[k-1], \quad (7)$$

where $\beta \in [0,1]$ is a tuning parameter used to vary the degrees of randomness, $\overline{v}$ and $\overline{\varphi}$ are the mean values of flying speed and direction for UAV, respectively, when $k \to \infty$, and $G_v[k-1]$ and $G_\varphi[k-1]$ are two random variables that follow the independent and unrelated Gaussian distributions. Given the flying speed and direction of UAV $n$ in (6) and (7), the horizontal location $\mathbf{q}_n[k]$ of UAV $n$ at time slot $k$ within $\Upsilon(\mathfrak{L}_l, \mathfrak{L}_w)$ can be determined by

$$x_n[k] = x_n[k-1] + v_n[k-1]\cos\varphi_n[k-1], \quad (8)$$

$$y_n[k] = y_n[k-1] + v_n[k-1]\sin\varphi_n[k-1]. \quad (9)$$

*2) Spectrum Sensing Model:* To identify the occupation status of the licensed sub-channels for PUs during the disaster information fusion phase, each UAV performs the SS via the energy detection policy [39]. This can be modeled as a binary hypothesis testing problem, which distinguishes between two hypotheses $\mathcal{H}_{n,c}^I$ and $\mathcal{H}_{n,c}^B$ for UAV $n$ corresponding to the idle and busy states of the PU on sub-channel $c$ at time $\varsigma$ respectively, for $n \in \mathcal{N}$ and $\varsigma \in [\tau, \tau + \tau_s]$, i.e.,

$$\begin{cases} \mathcal{H}_{n,c}^I: & S_c(\varsigma) = 0 \text{ (idle)}, \\ \mathcal{H}_{n,c}^B: & S_c(\varsigma) = 1 \text{ (busy)}. \end{cases} \quad (10)$$

The performance of SS is described by two metrics, namely, the detection probability and the false alarm probability [40]. Note that a higher detection probability brings about better protection to UAVs, whereas a lower false alarm probability results in efficient utilization of sub-channels. Denote $\xi_{n,c}$ as the decision threshold for UAV $n$ to decide whether sub-channel $c$ is occupied by the PU. For simplicity, we set the detection threshold $\xi_{n,c}$ for sub-channel $c$ to be the same for all the UAVs during the SS subphase. Then, the detection probability and false alarm probability of UAV $n$ for sub-channel $c$ can be respectively obtained as

$$\rho_{n,c}^D = \mathcal{Q}\left(\frac{\xi_{n,c} - 2\vartheta_c(\Gamma_{n,c}+1)}{\sqrt{4\vartheta_c(2\Gamma_{n,c}+1)}}\right), \quad (11)$$

$$\rho_{n,c}^F = \mathcal{Q}\left(\frac{\xi_{n,c} - 2\vartheta_c}{\sqrt{4\vartheta_c}}\right), \quad (12)$$

where $\Gamma_{n,c}$ is the received SNR of the primary signal at UAV $n$ on sub-channel $c$, $\vartheta_c$ is the bandwidth-time product for sub-channel $c$, and $\mathcal{Q}(x)$ is the cumulative distribution function of the standard Gaussian distribution, which can be defined by $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-\frac{y^2}{2}} dy$.

TABLE I
FOUR DIFFERENT CASES OF IMPERFECT SS.

| Case | Actual state | Sensing result | Probability |
|---|---|---|---|
| 1 | $\mathcal{H}_{n,c}^I$ (Idle) | $\tilde{\mathcal{H}}_{n,c}^I$ (Vacant) | $\rho_{n,c}^{(1)} = \Pr\left\{\mathcal{H}_{n,c}^I \middle\vert \tilde{\mathcal{H}}_{n,c}^I\right\}$ |
| 2 | $\mathcal{H}_{n,c}^I$ (Idle) | $\tilde{\mathcal{H}}_{n,c}^B$ (Occupied) | $\rho_{n,c}^{(2)} = \Pr\left\{\mathcal{H}_{n,c}^I \middle\vert \tilde{\mathcal{H}}_{n,c}^B\right\}$ |
| 3 | $\mathcal{H}_{n,c}^B$ (Busy) | $\tilde{\mathcal{H}}_{n,c}^I$ (Vacant) | $\rho_{n,c}^{(3)} = \Pr\left\{\mathcal{H}_{n,c}^B \middle\vert \tilde{\mathcal{H}}_{n,c}^I\right\}$ |
| 4 | $\mathcal{H}_{n,c}^B$ (Busy) | $\tilde{\mathcal{H}}_{n,c}^B$ (Occupied) | $\rho_{n,c}^{(4)} = \Pr\left\{\mathcal{H}_{n,c}^B \middle\vert \tilde{\mathcal{H}}_{n,c}^B\right\}$ |

Due to the varying ground-to-air channel conditions and the physical hardware impairments [41], the UAVs cannot completely ensure the perfect SS. Imperfect SS should be fully considered. Particularly, two errors of SS are inevitable, namely, the false alarm and miss detection. The false alarm indicates sub-channel $c$ is detected as occupied by the PU but it is actually idle, while the miss detection means sub-channel $c$ is detected as vacant when it is truly busy. Combining the relationship between actual state and sensing result in imperfect sensing, we thus focus on four different cases, which are listed in Table I. For Case 1 and Case 4, the UAV makes the correct decision. Besides, Case 2 is a false alarm, and Case 3 is a miss-detection. We denote $\tilde{\mathcal{H}}_{n,c}^B$ as the sensing result that sub-channel $c$ is occupied by the PU for UAV $n$, and denote $\tilde{\mathcal{H}}_{n,c}^I$ as the sensing result that sub-channel $c$ is vacant for UAV $n$. For clarity, let $\rho_c$ be the prior probability of sub-channel $c$ being idle at time $\varsigma$, i.e., $\rho_c \triangleq \Pr\{S_c(\varsigma) = 0\} = \frac{\lambda_0^c}{\lambda_0^c + \lambda_1^c}$. With the Bayes' rule, the probabilities of Cases 1, 2, 3, and 4 for UAV $n$ on sub-channel $c$ can be written by (13), (14), (15), and (16), respectively, shown at the top of the next page.

Given the probabilities of Cases 1, 2, 3, and 4 as listed in Table I, it can be easy to verify that only Case 1 ensures the normal access to the sub-channel by the UAV for forwarding the sensing data. Note that in Case 3 of miss-detection, the UAV can also transmit the sensing data to the next-hop UAV. However, the transmitted sensing data will be lost due to the collision with the PU transmission. Different from [42], that combines all of four different cases to represent the throughput of secondary transmission, we in this work focus on the actual achievable rate of the UAV, i.e., the successfully transmitted bits, under the imperfect SS. In this case, the number of actual transmitted bits will be zero, and thus, we ignore Case 3 of miss-detection during the disaster information fusion phase.

$$\rho_{n,c}^{(1)}=\Pr\left\{\mathcal{H}_{n,c}^{I}\left|\tilde{\mathcal{H}}_{n,c}^{I}\right.\right\}=\frac{\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{I}\right.\right\}}{\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{I}\right.\right\}+\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{B}\right.\right\}}=\frac{\rho_c\left(1-\rho_{n,c}^{F}\right)}{\rho_c\left(1-\rho_{n,c}^{F}\right)+\left(1-\rho_c\right)\left(1-\rho_{n,c}^{D}\right)}, \quad (13)$$

$$\rho_{n,c}^{(2)}=\Pr\left\{\mathcal{H}_{n,c}^{I}\left|\tilde{\mathcal{H}}_{n,c}^{B}\right.\right\}=\frac{\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{I}\right.\right\}}{\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{I}\right.\right\}+\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{B}\right.\right\}}=\frac{\rho_c\rho_{n,c}^{F}}{\rho_c\rho_{n,c}^{F}+\left(1-\rho_c\right)\rho_{n,c}^{D}}, \quad (14)$$

$$\rho_{n,c}^{(3)}=\Pr\left\{\mathcal{H}_{n,c}^{B}\left|\tilde{\mathcal{H}}_{n,c}^{I}\right.\right\}=\frac{\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{B}\right.\right\}}{\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{B}\right.\right\}+\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{I}\left|\mathcal{H}_{n,c}^{I}\right.\right\}}=\frac{\left(1-\rho_c\right)\left(1-\rho_{n,c}^{D}\right)}{\left(1-\rho_c\right)\left(1-\rho_{n,c}^{D}\right)+\rho_c\left(1-\rho_{n,c}^{F}\right)}, \quad (15)$$

$$\rho_{n,c}^{(4)}=\Pr\left\{\mathcal{H}_{n,c}^{B}\left|\tilde{\mathcal{H}}_{n,c}^{B}\right.\right\}=\frac{\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{B}\right.\right\}}{\Pr\left\{\mathcal{H}_{n,c}^{B}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{B}\right.\right\}+\Pr\left\{\mathcal{H}_{n,c}^{I}\right\}\Pr\left\{\tilde{\mathcal{H}}_{n,c}^{B}\left|\mathcal{H}_{n,c}^{I}\right.\right\}}=\frac{\left(1-\rho_c\right)\rho_{n,c}^{D}}{\left(1-\rho_c\right)\rho_{n,c}^{D}+\rho_c\rho_{n,c}^{F}}. \quad (16)$$

*3) Transmission Model:* To enable the sensing data forwarding in the disaster information fusion phase, the UAV and its next-hop UAV set up the air-to-air (A2A) channel between them by a short-range LoS link. For practical consideration, we incorporate both path loss exponent and shadow fading into the A2A channel model for low-altitude UAV swarm application. Thus, we utilize the extension of the classical log-distance path loss model as in [43] to describe the propagation path loss of A2A channel between the UAV pair in CU-SWARM. Specifically, the path loss (in dB) of A2A channel between UAV $n$ and its next-hop UAV $j$ on sub-channel $c$ at time slot $K$, for $n, j \in \mathcal{N}$ and $n \neq j$, can be given as

$$L_{n,j}^{c}\left[K\right]=L_c\left(d_0\right)+10\mu\log_{10}\left(\frac{d_{n,j}\left[K\right]}{d_0}\right)+X_{\sigma}^{c}+X_{A}^{c}, \quad (17)$$

where $L_c\left(d_0\right)$ is the path loss of sub-channel $c$ at the reference distance of $d_0 = 1$ m, $\mu \geq 2$ is the path loss exponent, $d_{n,j}\left[K\right]$ is the distance between UAV $n$ and its next-hop UAV $j$ at time slot $K$, $X_{\sigma}^{c}$ is the zero-mean Gaussian distributed random variable with standard deviation $\sigma$ used to describe the shadow fading of sub-channel $c$, and $X_{A}^{c}$ is the additional fading of sub-channel $c$ caused by increasing altitude of the UAV [43].

Then, the channel power gain between UAV $n$ and its next-hop UAV $j$ on sub-channel $c$ at time slot $K$ is modeled by

$$g_{n,j}^{c}\left[K\right]=\left(10^{\frac{L_{n,j}^{c}\left[K\right]}{10}}\right)^{-1}. \quad (18)$$

Combining with the imperfect SS, the achievable rate from UAV $n$ to its next-hop UAV $j$ on sub-channel $c$ at time slot $K$ is obtained as

$$R_{n,j}^{c}=B_P\log_2\left(1+\frac{p_u g_{n,j}^{c}\left[K\right]}{\sigma_j^2}\right)\rho_{n,c}^{(1)}, \quad (19)$$

where $p_u$ is the transmit power of the UAV for forwarding the sensing data, and $\sigma_j^2$ is the noise power spectral density at next-hop UAV $j$.

*4) Energy Consumption Model:* We focus on the energy consumption of the UAV in a rotary-wing configuration before the routing decision and forwarding subphase, i.e., the energy consumed in both the disaster sensing phase with duration $\tau$ and the SS subphase with duration $\tau_s$ during the disaster information fusion phase, as indicated in Fig. 2.

As for the disaster sensing phase, determined by the GM mobility model, we consider that each UAV is capable of hovering at the time-dependent horizontal location for disaster sensing within one time slot, and flying to the next horizontal location within the successive time slot. Therefore, the total energy consumption of UAV during the disaster sensing phase includes three major components, namely, the energy consumed in sensing, hovering, and flying, respectively. The sensing related energy consumption of UAV is aimed at the energy consumed at the operation of disaster sensing, while hovering at the horizontal location. For simplicity, we assume that the sensing related power for UAV is regarded as a constant within each time slot, but may continuously vary over the different UAVs. We thus let $P_n^S\left[k\right]$ denote the disaster sensing related power of UAV $n$ at time slot $k$.

On the other hand, the required power of hovering for each UAV primarily includes the induced power and profile power [44]. For the UAV in hovering status, the induced power produces thrust by propelling air downward, while the profile power overcomes the rotational drag encountered by rotating propeller blades. Thus, the hovering power consumption of UAV $n$ at time slot $k$ can be expressed by

$$P_n^H\left[k\right]=\underbrace{\left(1+\varpi\right)\frac{W_n^{\frac{3}{2}}}{\sqrt{2\varrho_a A_n}}}_{\triangleq P_n^I,\ \text{induced power}}+\underbrace{\frac{M_n\kappa c_d\varrho_a R_b^4}{8}\Omega^3}_{\triangleq P_n^P,\ \text{profile power}}, \quad (20)$$

where $\varpi$ is the incremental correction factor to induced power, $W_n$ is the weight of UAV $n$, $\varrho_a$ is the density of air, $A_n$ is the rotor disc area of UAV $n$, $M_n$ is the total number of blades of UAV $n$, $\kappa$ is the blade chord width, $c_d$ is the drag coefficient of the blade, $R_b$ is the radius of rotor blade, and $\Omega$ is the angular speed of rotor blade.

Besides, the flying energy consumption of UAV is mainly based on its propulsion energy consumption to maintain the airborne and to support the flying speed. As in [44], the propulsion power consumption of UAV is decomposed into three components, i.e., the induced power, profile power, and parasite power. Note that the parasite power resists UAV's body drag when there exists relative translational motion between the aircraft and wind. To be specific, the required power of UAV $n$ for flying at time slot $k$ can be calculated as

$$P_n^F[k] = P_n^I \sqrt{\sqrt{1+\frac{v_n^4[k]}{4v_0^4}} - \frac{v_n^2[k]}{2v_0^2}} + P_n^P\left(1+\frac{3v_n^2[k]}{V_t^2}\right) + \frac{f_0 \varrho_a r_0 A_n v_n^3[k]}{2}, \tag{21}$$

where $v_0$ is the mean rotor induced velocity in hovering, $V_t$ is the tip speed of the rotor blade, $f_0$ is the fuselage drag ratio, and $r_0$ is the ratio of blade area to the disc area.

So far, we have obtained the required power for sensing, hovering, and flying of rotary-wing UAV during the disaster sensing phase. To derive the total energy consumption, we then turn to derive the required time to sense, hover, and fly for the UAV within the time slot, respectively. Given the horizontal location $\mathbf{q}_n[k]$ of UAV $n$ as in (8) and (9), the flying distance with respect to UAV $n$ from previous time slot $k-1$ to current time slot $k$ within $\Upsilon(\mathfrak{L}_l, \mathfrak{L}_w)$ can be specified by

$$\Delta d_n[k] = \sqrt{(x_n[k]-x_n[k-1])^2 + (y_n[k]-y_n[k-1])^2}. \tag{22}$$

Then, the required time to sense, hover, and fly for UAV $n$ at time slot $k$ can be expressed as

$$\begin{cases} T_n^S[k] = T_n^H[k] = \dfrac{\tau}{K} - \dfrac{\Delta d_n[k]}{v_n[k]}, & \text{for sensing or hovering,} \\ T_n^F[k] = \dfrac{\Delta d_n[k]}{v_n[k]}, & \text{for flying.} \end{cases} \tag{23}$$

Thereby, the total energy consumption of UAV $n$ during the disaster sensing phase is given by

$$\begin{aligned} E_n^{DS} &= \sum_{k=1}^{K} T_n^S[k]\, P_n^S[k] + \sum_{k=1}^{K} T_n^H[k]\, P_n^H[k] + \sum_{k=1}^{K-1} T_n^F[k]\, P_n^F[k] \\ &= \sum_{k=1}^{K}\left(\frac{\tau}{K} - \frac{\Delta d_n[k]}{v_n[k]}\right)\left(P_n^S[k] + P_n^H[k]\right) + \sum_{k=1}^{K-1} \frac{\Delta d_n[k]}{v_n[k]} P_n^F[k]. \end{aligned} \tag{24}$$

For the SS subphase, the total energy consumption of UAV includes two major parts, namely, the energy consumed in SS and hovering, respectively. Let $P_n^{SP}$ be the SS associated power of UAV $n$ during the SS subphase with duration $\tau_s$. Recall that the horizontal location $\mathbf{q}_n[K]$ of UAV $n$ will be kept unchanged at time slot $K$, which continues for the whole duration of the disaster information fusion phase. Then, the hovering power consumption of UAV $n$ during the SS subphase is equal to the power consumed at time slot $K$ for hovering. Therefore, the total energy consumption of UAV $n$ during the SS subphase is denoted as

$$E_n^{SS} = \left(P_n^{SP} + P_n^H[K]\right)\tau_s. \tag{25}$$

We denote the initial energy of UAV $n$ by $E_n^I$, which is determined by the nature of UAV's energy storage (e.g., on-board lithium-ion polymer battery) and is further assumed to be equal for each UAV. The residual energy of UAV $n$ during the routing decision and forwarding subphase can be given by

$$E_n = E_n^I - E_n^{DS} - E_n^{SS}. \tag{26}$$

## V. $Q$-Learning Aided Intelligent Routing Scheme

In this section, we will present the $Q$-learning framework of intelligent routing with maximum utility for CU-SWARM in Subsection V-A, based on which we particularly design the reward function by combining the routing metric and candidate UAV selection optimization in Subsection V-B. Then, we propose an overall iterative algorithm to achieve the training procedure through the analytical efforts in Subsection V-C, and also analyze the complexity of the proposed algorithm in Subsection V-D.

### A. $Q$-Learning Framework

Under the above setup, we pursue an intelligent routing design aided by $Q$-learning for improving the system's performance of utility maximization in CU-SWARM. The proposed $Q$-learning framework of intelligent routing considers each UAV as a learning agent in RL. Each agent maintains its own $Q$-table, while the other UAVs, HAP, PUs, and licensed sub-channels over the RB constitute the external environment for the agent. Particularly, the learning process for the agent by interacting with the environment is modeled as a Markov decision process (MDP) with discrete time-steps, which is formulated by 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}: \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the finite set of rewards that contains all the immediate rewards when moving from one state to next state resulting from the actions taken by the agents, and $\gamma \in [0,1]$ is the discount factor, which determines the effect of future rewards on the current action. Note that a higher discount factor generally contributes to more accumulated rewards. The state, action, and reward of the $Q$-learning framework are defined as follows.

- **State**: The state observed by an agent is determined by a combination of the horizontal location and residual energy of a UAV during the routing decision and forwarding subphase. Thus, we define the state of UAV $n$ at time-step $t$, for $n \in \mathcal{N}$, as follows
$$s_n^{(t)} = \left(\mathbf{q}_n[K], E_n\right)^{(t)}. \tag{27}$$
The state space is given by the discrete set of all possible states at time-step $t$, i.e., $\mathcal{S} = \left\{s_1^{(t)}, s_2^{(t)}, \cdots, s_n^{(t)}, \cdots, s_N^{(t)}\right\}$.
- **Action**: The action taken by the agent at each epoch consists of two components, namely, obtaining an idle sub-channel and choosing a next-hop UAV. Thus, each agent needs to consider at least the following two aspects before taking a new action: i) to choose the sub-channel by identifying whether it is being occupied by the PU or not, and ii) to select the next-hop UAV as the relay to forward the sensing data. Thus, the total number of actions is given by $N \times C$. We define the action performed in time-step $t$ as $a_\ell^{(t)}$, for $\ell = 1, 2, \cdots, N\times C$. The action space is denoted by the discrete set of all possible actions at time-step $t$, i.e., $\mathcal{A} = \left\{a_1^{(t)}, a_2^{(t)}, \cdots, a_\ell^{(t)}, \cdots, a_{N\times C}^{(t)}\right\}$.
- **Reward**: The immediate reward is the utility achieved by the agent in response to the state transition from current state $s_n^{(t)} \in \mathcal{S}$ to next state $s_{n'}^{(t+1)} \in \mathcal{S}$ by executing an action $a_\ell^{(t)} \in \mathcal{A}$, for $n \neq n' \in \mathcal{N}$. With this transition, the agent receives an immediate reward

of $r^{(t+1)} \triangleq r\left(s^{(t)}, a^{(t)}\right) \in \mathcal{R}$ that describes its benefit from taking action $a^{(t)} \in \mathcal{A}$ at state $s^{(t)} \in \mathcal{S}$. Note that $r\left(s^{(t)}, a^{(t)}\right)$ is the reward function that is designed based on the adopted routing metric and the available sub-channel we will elaborate on in the following subsection.

- **Action selection algorithm**: The $\epsilon$-greedy policy is applied in the learning process to enable the agent to explore and exploit available action for current state. When the agent explores, it will select an action at random, targeting a higher long-term reward. And when exploiting, it selects the greedy action to gain the most rewards immediately, even if it is a sub-optimal behavior. The aim of the $\epsilon$-greedy policy is thus to seek a trade-off between exploration and exploitation for the agent by performing the exploration with probability $\epsilon \in (0,1)$, referred to as the exploration rate. More specifically, conditioned on the current state being $s^{(t)}$ in time-step $t$, the agent chooses action $a^{*(t)} \in \mathcal{A}$ that maximizes the $Q$-value with probability $1-\epsilon$ for exploitation, and a random action $a'^{(t)} \in \mathcal{A}$ with probability $\epsilon$ for exploration, i.e.,

$$a^{(t)} = \begin{cases} \text{random action } a'^{(t)}, & \text{with probability } \epsilon, \\ \arg\max\limits_{a^{*(t)}} Q\left(s^{(t)}, a^{(t)}\right), & \text{with probability } 1-\epsilon, \end{cases} \quad (28)$$

where $Q\left(s^{(t)}, a^{(t)}\right)$ is the $Q$-value associate with action $a^{(t)}$ taken by the agent under state $s^{(t)}$ in time-step $t$.

### B. Reward Function Design

*1) Design of Routing Metric:* Since different UAVs may have different energy states and different rate requirements for forwarding sensing data in CU-SWARM, both the achievable rate and residual energy are essential for UAVs. The routing metric for identifying the potential next-hop UAV should be carefully designed by jointly capturing the achievable rate from current UAV to its next-hop UAV and the residual energy of the potential next-hop UAV. Thus, the utility received from selecting the potential next-hop UAV for current UAV contains two parts: one is proportional to the achievable rate of UAV pair, and the other is earned by considering the residual energy of the potential next-hop UAV. Hence, combined with (19) and (26), the utility of UAV $n$ obtained from selecting the potential next-hop UAV $j$ can be defined as

$$U_n\left(R_{n,j}^c, E_j\right) = w_n^R R_{n,j}^c + w_n^E E_j, \quad (29)$$

where $w_n^R$ and $w_n^E$ are adjustable weighting factors announced from UAV $n$, indicating its desire to improve the utility by well balancing the residual energy of next-hop UAV $j$ and the achievable rate from UAV $n$ to next-hop UAV $j$ on sub-channel $c$, and $w_n^R + w_n^E = 1$.

We then readily define the routing metric as UAV $n$ chooses UAV $j^*$ serving as the optimal next-hop UAV that maximizes the utility $U_n\left(R_{n,j}^c, E_j\right)$ given in (29), i.e.,

$$j^* = \arg\max_{j \in \mathcal{N},\, j \neq n} U_n\left(R_{n,j}^c, E_j\right). \quad (30)$$

*2) Candidate UAV Selection Strategy:* Due to the large-scale deployment of UAVs in the aerial sensing layer, it is challenging to efficiently identify the next-hop UAV $j^*$ for UAV $n$ using (30). In addition, the massive deployment of UAVs enlarges the scale of action selection taken by the agent, and consequently, the convergence speed of $Q$-learning algorithm will become much slower. Therefore, it is necessary to optimize the action selection taken by the agent for reducing the complexity of the next-hop UAV selection. With this in mind, we present a paradigm of *candidate UAV set* to further narrow the search scope for the next-hop UAVs. We denote the maximum transmission range of UAV $n$ by $R_u$, which is assumed to be equal for each UAV. Then, the UAVs distributed within $R_u$ of UAV $n$ over $\Upsilon\left(\mathfrak{L}_l, \mathfrak{L}_w\right)$ can be regarded as the neighboring UAVs of UAV $n$, as illustrated in Fig. 3(a). To facilitate the following analysis, let $\mathcal{N}_n[K]$ correspond to the set of neighboring UAVs of UAV $n$. The key notations used in the candidate UAV selection strategy are listed in Table II.

We denote the horizontal location of the HAP mapped onto $\Upsilon\left(\mathfrak{L}_l, \mathfrak{L}_w\right)$ by $\mathbf{q}_H = (x_0, y_0)$, which is also referred to as the HAP's projection point, as depicted in Fig. 3(b). For brevity, we use $\mathcal{F}_H[K]$ to denote the set of the UAVs located inside the HAP coverage area[3]. Therefore, the HAP coverage area can be described as a circular area $\Psi_H\left(\mathbf{q}_H, \mathcal{F}_H[K], R_H\right)^+$ of radius $R_H$ centered at the projection point $\mathbf{q}_H$ of the HAP over the 2D rectangular area $\Upsilon\left(\mathfrak{L}_l, \mathfrak{L}_w\right)$, as shown in Fig. 3(b).

Accordingly, we have the potential to devise a circular sector $\mathcal{J}_n\left(\mathbf{q}_n[K], R_u, \psi_n\right)^+$ with a radius of length $R_u$ and a central angle (in rad) of $\psi_n$ for UAV $n$ within $\Upsilon\left(\mathfrak{L}_l, \mathfrak{L}_w\right)$. As can be seen from Fig. 3(a), the circular sector $\mathcal{J}_n(\cdot)^+$ can also be viewed as a pie-shaped part of a circle enclosed by two radii and a minor arc between them. Three constraints are considered for designing the circular sector $\mathcal{J}_n(\cdot)^+$:

- **Location constraint**: The center of the circle associated with the circular sector is the horizontal location of UAV $n$, i.e., $\mathbf{q}_n[K] = (x_n[K], y_n[K])$.
- **Arc constraint**: The minor arc of the circular sector is divided equally by the connection line between $\mathbf{q}_n[K]$ of UAV $n$ and $\mathbf{q}_H$ of the HAP.
- **Direction constraint**: The central angle of the circular sector is always pointing towards $\mathbf{q}_H$ of the HAP.

With that in mind, we then consider neighboring UAV $j$ of UAV $n$ within $\Upsilon\left(\mathfrak{L}_l, \mathfrak{L}_w\right)$, for $j \in \mathcal{N}_n[K]$. The distance from UAV $n$ to UAV $j$, the distance from UAV $n$ to the HAP's projection point, and the distance from UAV $j$ to the HAP's projection point, can be respectively expressed as

$$d_{n,j}[K] = \sqrt{\left(x_n[K] - x_j[K]\right)^2 + \left(y_n[K] - y_j[K]\right)^2}, \quad (31)$$

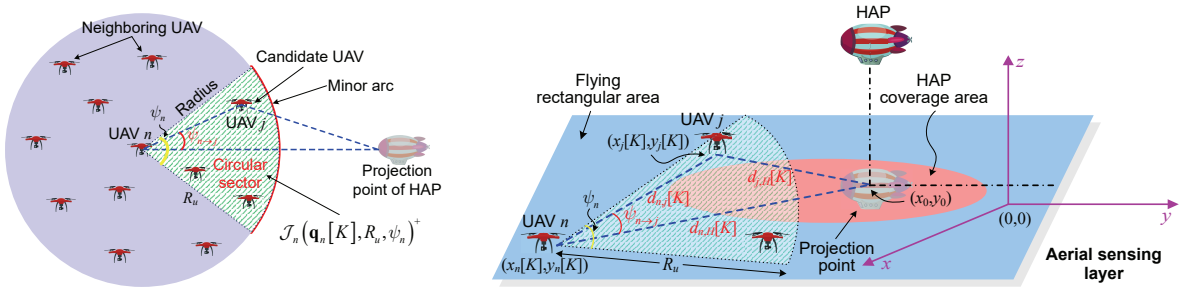
$$d_{n,H}[K] = \sqrt{\left(x_n[K] - x_0\right)^2 + \left(y_n[K] - y_0\right)^2}, \quad (32)$$

$$d_{j,H}[K] = \sqrt{\left(x_j[K] - x_0\right)^2 + \left(y_j[K] - y_0\right)^2}. \quad (33)$$

The angle (in rad) of the triangle formed by UAV $n$, UAV $j$, and the projection point $\mathbf{q}_H$ of the HAP is obtained by

$$\psi_{n \to j} = \arccos \frac{d_{n,j}^2[K] + d_{n,H}^2[K] - d_{j,H}^2[K]}{2 d_{n,j}[K]\, d_{n,H}[K]}. \quad (34)$$

[3]Note that the HAP can create $\mathcal{F}_H[K]$ via the periodic exchange of "hello" message with the UAVs located inside the HAP coverage area, while the associated UAVs can also obtain $\mathcal{F}_H[K]$ through the periodic exchange of "hello" message with the HAP [45]. In practice, the "hello" message contains the coverage radius and location information of the HAP as well as the velocity, timestamp, and position information of UAVs. Remark that a detailed discussion on the control message format is beyond the scope of this work.




(a) (b)

Fig. 3. An illustration of the selection strategy for candidate UAVs during the disaster information fusion phase: (a) 2D plane view, and (b) 3D view.

TABLE II
LIST OF KEY NOTATIONS USED IN CANDIDATE UAV SELECTION STRATEGY.

| Notation | Description |
|---|---|
| $\Upsilon(\mathfrak{L}_l, \mathfrak{L}_w)$ | 2D rectangular area with a length of $\mathfrak{L}_l$ and a width of $\mathfrak{L}_w$ |
| $\Psi_H(\mathbf{q}_H, \mathcal{F}_H[K], R_H)^+$ | HAP coverage area with a radius of $R_H$ centered at HAP projection point $\mathbf{q}_H$ |
| $\mathcal{J}_n(\mathbf{q}_n[K], R_u, \psi_n)^+$ | Circular sector with a radius of length $R_u$ and a central angle of $\psi_n$ for UAV $n$ |
| $\mathcal{N}_n[K]$ | Set of the neighboring UAVs of UAV $n$ at time slot $K$ |
| $\mathcal{F}_H[K]$ | Set of the UAVs located inside the HAP coverage area at time slot $K$ |
| $\mathcal{G}_n[K]$ | Set of the candidate UAVs of UAV $n$ at time slot $K$ |
| $\mathbf{q}_n[K] = (x_n[K], y_n[K])$ | Horizontal location of UAV $n$ at time slot $K$ |
| $\mathbf{q}_H = (x_0, y_0)$ | Horizontal location of HAP |
| $\psi_{n\to j}$ | Angle of the triangle formed by UAV $n$, UAV $j$, and the HAP projection point $\mathbf{q}_H$ |

To reduce the complexity for choosing the next-hop UAVs, and to guarantee the efficiency for forwarding the sensing data, only the candidate UAVs inside the circular sector $\mathcal{J}_n(\cdot)^+$ (see Fig. 3(a)) have the opportunity to serve as the potential next-hop UAVs. With this insight, we can then readily establish the following constraints that must be simultaneously satisfied for identifying the candidate UAVs with respect to UAV $n$, i.e.,

- **Central angle constraint**:

$$\psi_{n\to j} \le 0.5\psi_n. \tag{35}$$

- **SNR constraint**:

$$p_u g_{n,j}^c[K]/\sigma_j^2 \ge \Gamma_{\min}, \tag{36}$$

where $\Gamma_{\min}$ is the acceptable SNR for UAV.

It can be noted that the central angle constraint (35) is designed to narrow the scale for searching the next-hop UAVs from the physical location perspective, while the SNR constraint (36) is considered based on the transmission quality of A2A channel for the sensing data forwarding.

Although the circular sector $\mathcal{J}_n(\cdot)^+$ and the constraints in (35) and (36) give a solution for the selection of candidate UAVs for each UAV, it still remains to develop an algorithm to indicate the execution structure for the equations and constraints. Let $\mathcal{G}_n[K]$ be the set of the candidate UAVs of UAV $n$. We then propose Algorithm 1, which gives the procedures of the implementation. Algorithm 1 can be implemented by each UAV via only local information and limited interaction with other neighboring UAVs[4], and consequently, Algorithm 1 is distributed and the practicability is guaranteed.

[4]Note that in Algorithm 1, the horizontal location information, e.g., $\mathbf{q}_n[K]$, $\mathbf{q}_j[K]$, and $\mathbf{q}_H$, can be obtained at each UAV through an onboard GPS device. In addition, the neighbor table (e.g., neighboring UAV set $\mathcal{N}_n[K]$) for each UAV can be created by periodic exchange of "hello" message containing the velocity, timestamp, and position information of neighboring UAVs [45].

**Algorithm 1** Distributed Candidate UAV Selection Algorithm

**Input:** $R_u$, $\Gamma_{\min}$, $\mathcal{N}_n[K]$, $\mathbf{q}_n[K]$, $\mathbf{q}_j[K]$, $\mathbf{q}_H$, $\psi_n$, $\forall n, j$.
**Output:** $\mathcal{G}_1[K]$, $\mathcal{G}_2[K]$, $\cdots$, $\mathcal{G}_N[K]$.

```
for n = 1 to N do
    Set G_n[K] = ∅;
    Calculate d_{n,H}[K] according to (32);
    for j = 1 to |N_n[K]| do
        Calculate d_{n,j}[K] according to (31);
        Calculate d_{j,H}[K] according to (33);
        Calculate ψ_{n→j} according to (34);
        if ψ_{n→j} ≤ 0.5ψ_n then
            if p_u g^c_{n,j}[K]/σ_j^2 ≥ Γ_min then
                Set G_n[K] = G_n[K] ∪ {j};
            end if
        end if
    end for
    Return G_n[K];
end for
```

With the output of Algorithm 1, the action space of each agent at time-step $t$ can be interpreted as

$$\mathcal{A} = \left\{a_1^{(t)}, a_2^{(t)}, \cdots, a_\ell^{(t)}, \cdots, a_{|\mathcal{G}_n[K]|\times C}^{(t)}\right\}, \tag{37}$$

where $|\cdot|$ is the cardinality of a set (or a space).

*3) Reward Function Representation:* The reward function is formulated by the immediate utility of the agent in response to the state transition from $s_n^{(t)}$ to $s_{n'}^{(t+1)}$ by carrying out an action $a_\ell^{(t)}$, for $n \ne n' \in \mathcal{N}$. The design principle of reward function are shown as follows:

- **Single-hop case**: Since UAV $j^*$ can directly transmit the sensing data to the HAP via single hop, we calculate the reward function obtained by UAV $n$ through setting a larger positive reward of $\mathscr{R}_H \in \mathbb{R}^+$, for $j^* \in \mathcal{F}_H[K]$, $\mathscr{R}_H \gg U_n\left(R_{n,j'}^c, E_{j'}\right)$, $j' \in \mathcal{N}_n[K]$, and $j^* \ne j'$.

- **Multi-hop case choosing candidate UAV**: As the routing metric is to maximize the utility $U_n\left(R^c_{n,j^*}, E_{j^*}\right)$, we represent the reward function as the instantaneous utility received by UAV $n$ via choosing UAV $j^*$ as the optimal next-hop UAV from $\mathcal{G}_n[K]$, for $j^* \in \mathcal{G}_n[K]$.
- **Multi-hop case choosing non-candidate UAV**: We set the reward to be zero when UAV $n$ chooses UAV $j^*$ as the next-hop UAV that does not belong to the candidate UAV set $\mathcal{G}_n[K]$, for $j^* \in \mathcal{N}_n[K] - \mathcal{G}_n[K]$.
- **Overlapped sub-channel usage**: The reward received by UAV $n$ would be a penalty value via setting a negative reward of $\mathscr{R}_P \in \mathbb{R}^-$ when UAV $n$ chooses UAV $j^*$ as the next-hop UAV via sub-channel $c$ overlapping with the same sub-channel $c$ through which the previous-hop UAV $n'$ chooses UAV $n$ as the next-hop UAV, for $j^* \in \mathcal{F}_H[K]$ or $j^* \in \mathcal{G}_n[K]$ or $j^* \in \mathcal{N}_n[K] - \mathcal{G}_n[K]$.

To sum up, based on the above design principle, the reward function can be specifically determined by

$$r^{(t+1)} \triangleq r\left(s^{(t)}, a^{(t)}\right) = \begin{cases} \mathscr{R}_H, & j^* \in \mathcal{F}_H[K], \\ w_n^R R^c_{n,j^*} + w_n^E E_{j^*}, & j^* \in \mathcal{G}_n[K], \\ 0, & j^* \in \mathcal{N}_n[K] - \mathcal{G}_n[K], \\ \mathscr{R}_P, & \text{Overlapped sub-channel usage}. \end{cases} \tag{38}$$

### C. Proposed Algorithm

We summarize the training procedure as an overall iterative algorithm for achieving the $Q$-learning aided intelligent routing. We use $\mathscr{E}_{\max}$ and $\mathscr{T}_{\max}$ to denote the maximum number of iterations and time-steps for training, respectively. Let $\Xi_{s^{(t)}}$ be the set of relay UAVs of state $s^{(t)}$. We define a temporary variable $r_{\text{acc}}$, which records the accumulated reward values during the learning process. At the very beginning, all $Q$-values in $Q$-table are initialized to zero, and the maximum number of iterations and time-steps are respectively set to be positive integers. Then, at each iteration $\zeta$, the accumulated rewards are also set as zero and an initial state is randomly selected as source UAV in time-step $t = 1$. In time-step $t$, by taking the current state, the agent obtains the candidate UAV set to generate the action space in (37) using Algorithm 1, and calculates the channel availability probability of the UAV due to the imperfect SS according to (13). Conditioned on the current state in this time-step, the action will be chosen via the $\epsilon$-greedy policy as discussed in (28). Subsequently, based on the design principle of the reward function in (38) by observing the next state, the accumulated rewards will be updated, and the $Q$-value will be further updated depending on the current state, the selected action and the next state. To be precise, the $Q$-value at each state can be calculated through the following iterative procedure

$$\begin{aligned} &Q\left(s^{(t)}, a^{(t)}\right) \leftarrow Q\left(s^{(t)}, a^{(t)}\right) \\ &+ \alpha\left\{r^{(t+1)} + \gamma \max_{a \in \mathcal{A}} Q\left(s^{(t+1)}, a\right) - Q\left(s^{(t)}, a^{(t)}\right)\right\}, \end{aligned} \tag{39}$$

where $\alpha \in (0, 1]$ is the learning rate. Here, we are interested in the mode of dynamic learning rate, which has a beneficial effect on the rate of convergence for all $Q$-values in $Q$-table by modifying the learning rate over time according to the environment change, instead of keeping it fixed. As in [46], the dynamic learning rate $\alpha$ can be obtained by

$$\alpha = \alpha_{\max}\left(1 - e^{-\frac{(\eta-1)^2}{\phi^2}}\right), \tag{40}$$

where $\alpha_{\max}$ is the maximum value of $\alpha$, $\phi^2$ is the updating rate, and $\eta$ is the environment change parameter used to measure the change rate of environment state. In practice, the environment change parameter $\eta$ can be specifically given as

$$\eta = \frac{Q\left(s^{(t)}, a^{(t)}\right) - \max_{a \in \mathcal{A}} Q\left(s^{(t+1)}, a\right)}{r^{(t+1)}}. \tag{41}$$

Note that the learning rate in (40) can be dynamically adjusted based on the feedback of the change of environment state, and the convergence rate for $Q$-table can also change adaptively. With this setup, the $Q$-table will be accordingly updated with $Q$-value at each iteration. The training process continues till the number of iterations reaches upper bound $\mathscr{E}_{\max}$, which indicates the training process achieving routing selection with maximum utility for each UAV. The detailed process of the proposed algorithm is shown in Algorithm 2.

### D. Complexity Analysis

We now analyze the complexity of Algorithm 2 in terms of the space and the computational complexity as follows.

*1) Space Complexity:* For Algorithm 2, the space complexity of the training process is given in Theorem 1.

**Theorem 1.** *For each agent, the space complexity of Algorithm 2 to keep track of the Q-table is* $\mathcal{O}(|\mathcal{S}| \cdot |\mathcal{A}|)$.

*Proof:* For Algorithm 2, the state space and action space of each agent are respectively determined by $|\mathcal{S}|$ and $|\mathcal{G}_n[K]| \cdot C$ (see the output of Algorithm 1). Then, each agent requires at least a memory of $|\mathcal{S}| \cdot |\mathcal{G}_n[K]| \cdot C$ to keep track of the $Q$-table, and the maximum of which is $|\mathcal{S}| \cdot |\mathcal{A}|$. Therefore, the space complexity of Algorithm 2 in the worst case would be $\mathcal{O}(|\mathcal{S}| \cdot |\mathcal{A}|)$, which is proportional to the size of the state space and action space for each agent. ■

Note that, for a larger size of state and action space, the space complexity accordingly is higher. In practice, we tend to reduce the space complexity by designing the candidate UAV selection strategy as previously introduced to cut down the actions taken by the agent. Therefore, the space complexity of Algorithm 2 can be reduced, and the candidate UAV selection optimization is implementable in practical scenarios.

*2) Computational Complexity:* In Algorithm 2, the computationally most expensive part is mainly incurred from the implementation of the selection of candidate UAVs for the agent in Algorithm 1 (*the inner loop*), as well as the $Q$-learning algorithm through the determination of each action by the $\epsilon$-greedy policy in (28) along with the update of the $Q$-value for each agent (*the main loop*). The computational complexity of Algorithm 2 is given in Theorem 2.

**Theorem 2.** *Given the UAV $n$'s neighboring UAV set $|\mathcal{N}_n[K]|$, the computational complexity of Algorithm 2 for the training process is* $\mathcal{O}(\mathscr{E}_{max} \cdot \mathscr{T}_{max} \cdot N \cdot |\mathcal{N}_n[K]|)$.

*Proof:* For the inner loop, in the step of obtaining $\mathcal{G}_n[K]$ at each iteration, the determination process requires $|\mathcal{N}_n[K]|$

**Algorithm 2** $Q$-Learning Aided Intelligent Routing Algorithm

**Input:** $\alpha_{\max}$, $\gamma$, $\epsilon$, $\mathcal{F}_H[K]$, $\mathscr{E}_{\max}$, $\mathscr{T}_{\max}$, $R_u$, $\Gamma_{\min}$, $\mathcal{N}_n[K]$, $\mathbf{q}_n[K]$, $\mathbf{q}_j[K]$, $\mathbf{q}_H$, $\psi_n$, $\forall n, j$.
**Output:** $Q\left(s^{(t)}, a^{(t)}\right)$.

**Initialize** $Q\left(s^{(t)}, a^{(t)}\right) = 0$, $\forall s^{(t)} \in \mathcal{S}$, $a^{(t)} \in \mathcal{A}$;
**repeat**
  **Initialize** $r_{\text{acc}} = 0$;
  Set $t = 1$, $\Xi_{s^{(t)}} = \varnothing$;
  Randomly select an initial state $s^{(t)} \in \mathcal{S}$ as source UAV;
  **while** $t \leq \mathscr{T}_{\max}$ **do**
    Observe current state $s^{(t)}$;
    Obtain candidate UAV set $\mathcal{G}_{s^{(t)}}[K]$ via Algorithm 1;
    Calculate channel availability probability $\rho^{(1)}_{s^{(t)},c}$ according to (13);
    Choose action $a^{(t)} \in \mathcal{A}$ via $\epsilon$-greedy policy in (28);
    Observe next state $s'^{(t)} \in \mathcal{S}$;
    **if** $s'^{(t)} \in \mathcal{F}_H[K]$ **then**
      Calculate reward $r^{(t+1)} = R_H \in \mathbb{R}^+$; Break;
    **else if** $s'^{(t)} \in \mathcal{G}_{s^{(t)}}[K]$ **then**
      Calculate $r^{(t+1)} = w^R_{s^{(t)}} R^c_{s^{(t)},s'^{(t)}} + w^E_{s^{(t)}} E_{s'^{(t)}}$;
      Set $\Xi_{s^{(t)}} = \Xi_{s^{(t)}} \cup \left\{s'^{(t)}\right\}$;
    **else if** UAV $s^{(t)}$ chooses UAV $s'^{(t)}$ via sub-channel $c$ through which previous-hop UAV $s'$ chooses UAV $s^{(t)}$ **then**
      Calculate reward $r^{(t+1)} = R_P \in \mathbb{R}^-$;
    **else**
      Calculate reward $r^{(t+1)} = 0$;
    **end if**
    Update accumulated reward $r_{\text{acc}} = r_{\text{acc}} + r^{(t+1)}$;
    Update $Q$-value $Q\left(s^{(t)}, a^{(t)}\right)$ according to (39);
    Update learning rate $\alpha$ according to (40) and (41);
    Update state $s^{(t)} \leftarrow s'^{(t)}$;
    Set $t = t + 1$;
  **end while**
  Set $\zeta = \zeta + 1$;
**until** $\zeta < \mathscr{E}_{\max}$

calculations at most. With the total number of iterations $N$ for all the UAVs, the computational complexity of the inner loop is given by the order of $\mathcal{O}\left(N \cdot |\mathcal{N}_n[K]|\right)$. For the main loop, the maximum number of iterations for training is specified by $\mathscr{E}_{\max}$, and in each iteration, the maximum number of time-steps is given as $\mathscr{T}_{\max}$. In a worst-case scenario, at each iteration, the update process of the $Q$-value and state for each agent entails $\mathscr{T}_{\max}$ operations at most. Then, the computational complexity of the main loop is determined by the order of $\mathcal{O}\left(\mathscr{E}_{\max} \cdot \mathscr{T}_{\max}\right)$. In summary, the total computational complexity of Algorithm 2 is $\mathcal{O}\left(\mathscr{E}_{\max} \cdot \mathscr{T}_{\max} \cdot N \cdot |\mathcal{N}_n[K]|\right)$. ■

Note that the computational complexity of Algorithm 2 is linear in the number of UAVs and the number of neighboring UAVs for any UAV. We wish to remark that, Algorithm 2 is suitable for practical applications of UAV swarm, since it has an acceptable complexity in polynomial time.

## VI. Performance Evaluation

In this section, we numerically evaluate the performance of the proposed $Q$-learning framework of intelligent routing by implementing simulations. For CU-SWARM, all the UAVs are uniformly distributed in their initial horizontal locations within $\Upsilon(10\,\text{km}, 8\,\text{km})$ at the fixed altitude $H_u = 50\,\text{m}$. We set the maximum transmission range of UAV to be $R_u = 300\,\text{m}$. The HAP's horizontal location mapped onto the 2D rectangular area is set to $\mathbf{q}_H = (0\,\text{m}, 0\,\text{m})$, and the radius of the HAP coverage area is set to $R_H = 500\,\text{m}$. For the time-slotted frame structure, the duration of cognitive frame is given by $T = 80\,\text{s}$, of which the duration of the disaster sensing phase is set as $\tau = 50\,\text{s}$. We set $K = 50$ equally spaced time slots to divide the duration of the disaster sensing phase, and set the duration of the SS subphase to be $\tau_s = 5\,\text{ms}$. The whole RB is divided into $C = 600$ licensed sub-channels, each having an equally-sized bandwidth of $B_P = 10\,\text{MHz}$. The acceptable SNR for UAV is set to $\Gamma_{\min} = 50\,\text{dB}$.

TABLE III
Simulation Parameters.

| Parameter | Description | Value |
|---|---|---|
| $\varpi$ | Incremental correction factor | 0.1 |
| $W_n$ | UAV's weight | $20\,\text{kg}\cdot\text{m/s}^2$ |
| $A_n$ | UAV's rotor disc area | $0.503\,\text{m}^2$ |
| $M_n$ | UAV's total number of blades | 4 |
| $\kappa$ | Blade chord width | 0.0157 |
| $c_d$ | Drag coefficient of the blade | 0.012 |
| $R_b$ | Radius of rotor blade | $0.4\,\text{m}$ |
| $\Omega$ | Angular speed of rotor blade | $300\,\text{rad/s}$ |
| $V_t$ | Tip speed of the rotor blade | $120\,\text{m/s}$ |
| $\varrho_a$ | Density of air | $1.225\,\text{kg/m}^3$ |
| $v_0$ | Mean rotor induced velocity | $4.03\,\text{m/s}$ |
| $f_0$ | Fuselage drag ratio | 0.6 |
| $r_0$ | Ratio of blade area to disc area | 0.05 |

Regarding the GM mobility model, the tuning parameter of UAV is given by $\beta = 0.5$, and the mean values of flying speed and direction of UAV are set as $\overline{v} = 12\,\text{m/s}$ and $\overline{\varphi} = \frac{1}{5}\pi\,\text{rad}$, respectively. In our simulations, the detection probability $\rho^D_{n,c}$ and false alarm probability $\rho^F_{n,c}$ of UAV, and the prior probability $\rho_c$ of sub-channel being idle are uniformly distributed over $[0.9, 0.95]$, $[0.05, 0.1]$, and $[0.5, 0.75]$, respectively. For the transmission model, we set the default parameters as $d_0 = 1\,\text{m}$, $L_c(d_0) = 46.42\,\text{dB}$, $\mu = 2$, $\sigma^2_j = -46\,\text{dBm}$. As analyzed in [43], the shadow fading of sub-channel is defined as the zero-mean Gaussian distributed random variable $X^c_\sigma = 1.9144\,\text{dB}$. Adapted from [43], the additional fading for sub-channel due to increasing altitude of UAV is calculated by

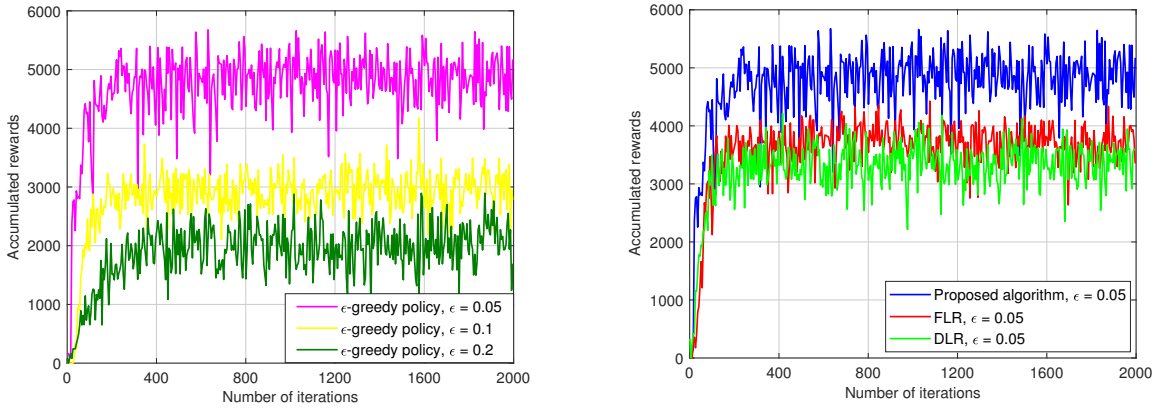
$$X^c_A = \hbar_1 \cdot H_u + \hbar_2, \tag{42}$$

where $\hbar_1 = -0.09393$ and $\hbar_2 = 4.702$ are adopted for simulations [43].

Additionally, the weighting factors for the utility in (29) are set to be $w^R_n = w^E_n = 0.5$. As for the energy consumption model, we set the initial energy of UAV as $E^I_n = 8800\,\text{Joule}$. The disaster sensing related power of UAV at each time slot is set to $P^S_n[k] = 1\,\text{W}$, and the SS associated power of UAV during the SS subphase is set to $P^{SP}_n = 250\,\text{mW}$. Other simulation parameters for UAV's energy consumption are summarized in Table III, unless otherwise specified, which are set based on the results in [44].

To verify the performance of the proposed algorithm, we consider four benchmarks listed as follows for comparison:

- **Fixed learning rate (FLR)**: Different from the dynamic learning rate adapted to the change of environment state, we in this scheme resort to the fixed learning rate by




(a) (b)

Fig. 4. Convergence behavior of the proposed algorithm, when $N = 900$, $p_u = 900$ mW, and $\psi_n = \frac{1}{3}\pi$ rad. (a) Proposed algorithm under different exploration rates $\epsilon$=0.05, $\epsilon$=0.1, and $\epsilon$=0.2, respectively. (b) Comparison between the proposed algorithm and the benchmarks of FLR and DLR with $\epsilon$=0.05.

setting the step size fixed at each iteration [46], [47]. In the simulations, we assume the constant learning rate $\alpha$=0.1 which is fixed during the training process.

- **Decaying learning rate (DLR)**: The decaying learning rate starts with a relatively large learning rate and then gradually decreases to a lower value with the iteration process [46]. From this perspective, we adopt an initially large learning rate $\alpha = 0.5$ in the simulations and then decay it to 0.05 after pre-defined number of iterations.
- **Intelligence routing without candidate UAV selection (IR-CUS)**: This scheme adopts the same training process as presented in Algorithm 2 under the same $Q$-learning framework of intelligent routing. The dynamic learning rate in (40) is also used for updating $Q$-values during the training process. At each time-step, this scheme does not adopt the candidate UAV selection strategy to optimize the action space of the agent by Algorithm 1.
- **Cognitive radio shortest path routing (CR-SPR)**: This scheme combines the traditional shortest-path metric and the spectrum-aware policy jointly to choose the next-hop UAV. It always chooses the potential next-hop UAV with higher probability of the associated sub-channel being idle (OFF state) via the energy detection technique, while satisfying the shortest-distance constraint.

Throughout the simulations, unless otherwise specified, we set the discount factor as $\gamma = 0.9$, and set the number of iterations to be $\mathscr{E}_{\max} = 2000$, for both the proposed algorithm and these four benchmarks. We obtained the final results by averaging every five previous simulated points to avoid the larger fluctuation of the curves.

### A. Convergence Speed and Analysis

To investigate the convergence behavior of the proposed algorithm, we start with illustrating the accumulated rewards of CU-SWARM versus the number of iterations in Fig. 4(a) and Fig. 4(b), respectively. In obtaining the convergence speed, the number of UAVs is set to $N = 900$, and meanwhile, the transmit power and the central angle are set as $p_u = 900$ mW and $\psi_n = \frac{1}{3}\pi$ rad, respectively.

Fig. 4(a) shows the convergence speed of the proposed algorithm over the number of iterations under different exploration rates $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 0.2$, respectively. We observe that the accumulated rewards of CU-SWARM for the proposed algorithm increase quickly in less than 200 iterations, and then gradually converge to the optimal values for varying exploration rates, which ensures that the proposed algorithm is practical. As can be also seen from Fig. 4(a), the accumulated rewards of the proposed algorithm with smaller exploration rate are obviously higher than the rewards with bigger exploration rate. The reason is that as the exploration rate decreases, the action can be chosen by the agent based on the maximum $Q$-value with a higher probability. These results provide a hint to choose a proper exploration rate for the training process to further improve the accumulated rewards.

In Fig. 4(b), we further report the convergence behavior of the proposed algorithm in terms of the accumulated rewards of CU-SWARM against the benchmark schemes of FLR and DLR with $\epsilon = 0.05$. It is clear that the accumulated rewards for the proposed algorithm and these two benchmarks increase consistently and converge rapidly within approximately 200 iterations, and then gradually reach the optimal values for the remaining iterations of training. We also find that the proposed algorithm outperforms two benchmarks in terms of the accumulated rewards, and the gap between them continues with the relatively fixed values. This is because the dynamic learning rate adopted in the proposed algorithm can modify the learning rate over time according to the change of external environment, which makes the training process more stable compared to the fixed and decaying learning rates in these two benchmarks. Such observations bolster the importance of choosing an appropriate mode of learning rate for the training process to improve the accumulated rewards.

### B. Performance Comparison and Analysis

The accumulated rewards of CU-SWARM are plotted against varying number of UAVs, $N$, in Fig. 5, for the proposed algorithm and four benchmarks, when $p_u = 900$ mW, $\psi_n = \frac{1}{3}\pi$ rad, and $\epsilon = 0.05$. It can be seen that as $N$ grows, the accumulated rewards present a markedly increasing trend for both the proposed algorithm and four benchmark schemes. With respect to the proposed algorithm and the benchmarks of FLR and DLR, an intuitive explanation for this trend is interpreted as follows. With the increasing number of UAVs, each UAV is more likely to have more action sets to choose,

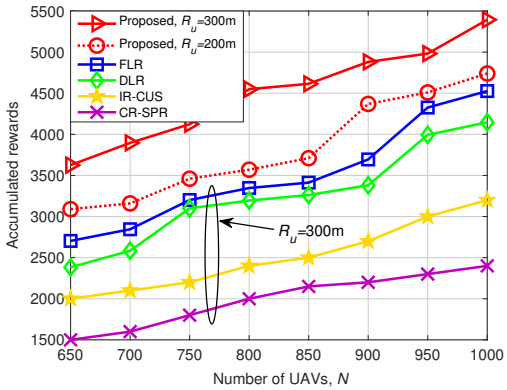



Fig. 5. Accumulated rewards versus number of UAVs $N$, when $p_u = 900\,\text{mW}$, $\psi_n = \frac{1}{3}\pi\,\text{rad}$, and $\epsilon = 0.05$.

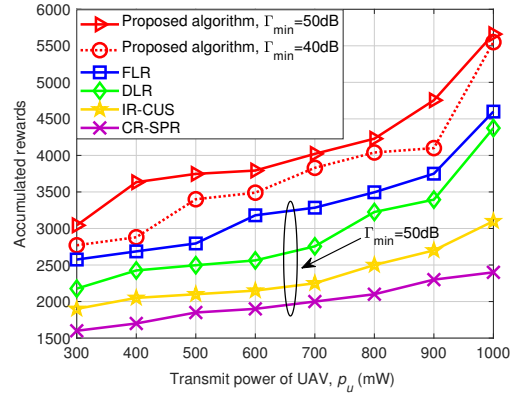



Fig. 6. Accumulated rewards versus transmit power of UAV $p_u$, when $N = 900$, $\psi_n = \frac{1}{3}\pi\,\text{rad}$, and $\epsilon = 0.05$.

and consequently, it has a greater chance to obtain the optimal routing compared to the case of a relatively small number of UAVs. This implies that more accumulated rewards can be received by CU-SWARM. As for the IR-CUS scheme, it can be explained that, although the optimized action space for the agent cannot be guaranteed, the IR-CUS scheme still has more opportunities to select the optimal actions, as $N$ keeps growing. While for the CR-SPR scheme, an increment of $N$ contributes to more chances to select the next-hop UAV with higher probability of the associated sub-channel being idle, thus resulting in more accumulated rewards.

From Fig. 5, we can also observe that the proposed algorithm shows performance gains over the benchmarks of FLR and DLR in terms of the accumulated rewards, under $R_u = 300\,\text{m}$. This is due to the advantage of our proposed algorithm by adopting the dynamic learning rate, adapting to the dynamically changing environment state in the training process for enhancing the accumulated rewards. As can be expected, under $R_u = 300\,\text{m}$, our proposed algorithm significantly outperforms the benchmarks of IR-CUS and CR-SPR regarding the accumulated rewards. This observation confirms our findings in the candidate UAV selection strategy via optimizing the action selection taken by the agent to obtain considerable performance gains over the IR-CUS scheme. However, by balancing the shortest-path and the spectrum selection, the CR-SPR scheme only considers the achievable rate between the UAV pair without capturing the benefits brought by the $Q$-learning process. Another important observation is that the performance of the proposed algorithm with $R_u = 300\,\text{m}$ in terms of the accumulated rewards always outperforms that of $R_u = 200\,\text{m}$. To explain, for the current UAV, the larger $R_u$ is, the more next-hop UAVs and actions are available, thus enabling the current UAV to find the optimal next-hop UAV to improve the accumulated rewards. The results show the performance gains of the accumulated rewards benefit a lot from choosing larger maximum transmission range of UAV.

In Fig. 6, we compare the accumulated rewards of CU-SWARM over the UAV's transmit power, $p_u$, between the proposed algorithm (with $\Gamma_{\min} = 40\,\text{dB}$ and $\Gamma_{\min} = 50\,\text{dB}$) and four benchmarks, when $N = 900$, $\psi_n = \frac{1}{3}\pi\,\text{rad}$, and $\epsilon = 0.05$. It is evident that the simulated accumulated rewards obviously increase with $p_u$ for the proposed algorithm and the benchmarks. To explain, the higher the UAV's transmit power is, the larger the achievable rate is obtained by UAV, yielding the increased accumulated rewards thanks to the routing metric directly proportional to the achievable rate.

Moreover, the results in Fig. 6 also show that the proposed algorithm obviously improve the accumulated rewards compared to the benchmarks of FLR and DLR with evolution of $p_u$, under $\Gamma_{\min} = 50\,\text{dB}$. Noteworthy, the simulation results for the effect of UAV's transmit power on the accumulated rewards provide similar insights to those for the effect of number of UAVs. This implies that we need to properly set up the learning rate to receive more accumulated rewards. Under $\Gamma_{\min} = 50\,\text{dB}$, the results of Fig. 6 further illustrate significant performance gains of the proposed algorithm over the benchmarks of IR-CUS and CR-SPR regarding the accumulated rewards. For the IR-CUS scheme, the reason is that the current UAV is difficult to choose the optimized next-hop UAV yielding more accumulated rewards, due to the lack of location, arc, and direction constraints when selecting the next-hop UAV. Since the CR-SPR scheme only considers the achievable rate between the UAV pair via the shortest-path metric, it cannot ensure that the current UAV is able to select the proper next-hop UAV that could improve the accumulated rewards. Besides, the accumulated rewards for the proposed algorithm with $\Gamma_{\min} = 50\,\text{dB}$ are observed to be always larger than those of $\Gamma_{\min} = 40\,\text{dB}$. The results manifest the crucial role of the acceptable SNR for UAV on the performance of accumulated rewards with increasing UAV's transmit power.

### C. The Effect of Central Angle and Discount Factor

Finally, in Fig. 7, we examine the accumulated rewards of CU-SWARM for the proposed algorithm with discount factors $\gamma = 0.9$ and $\gamma = 0.95$ against different values of central angle, $\psi_n$, of circular sector $\mathcal{J}_n(\cdot)^+$, when $N = 900$, $p_u = 900\,\text{mW}$, $\epsilon = 0.05$, and $B_P = 25\,\text{MHz}$. We can observe that with the increase of $\psi_n$ varying from $\frac{1}{6}\pi\,\text{rad}$ to $\pi\,\text{rad}$, the accumulated rewards for the proposed algorithm gradually increase, and basically do not change when $\psi_n = \pi\,\text{rad}$. This can be explained by the fact that, when $\psi_n$ increases from $\frac{1}{6}\pi\,\text{rad}$ to $\pi\,\text{rad}$, the current UAV will have more candidate UAVs as the potential next-hop UAVs. Thus, there will be more candidate actions to choose for the current UAV during

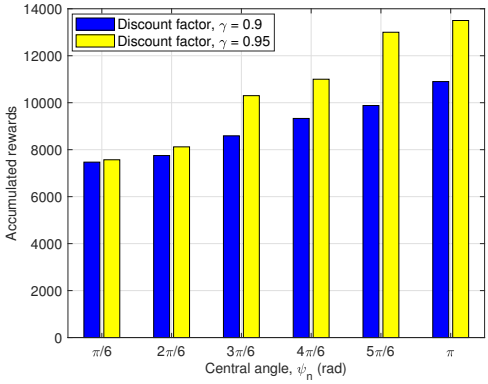



Fig. 7. Accumulated rewards versus central angle $\psi_n$ of circular sector $\mathcal{J}_n(\cdot)^+$, when $N = 900$, $p_u = 900\,\text{mW}$, $\epsilon = 0.05$, and $B_P = 25\,\text{MHz}$.

the training process, thereby resulting in notable performance gains of the accumulated rewards. Moreover, as shown in Fig. 7, the accumulated rewards for the proposed algorithm with $\gamma = 0.95$ are observed to be always larger than that of $\gamma = 0.9$. This implies that as $\gamma$ increases, the accumulated rewards clearly show an obvious improvement. Such behavior can be interpreted as follows. The discount factor reflects the ratio of future rewards to immediate rewards in the proposed $Q$-learning framework. The larger the discount factor is, the greater the importance of future rewards is shown on the current action. In other words, it puts more emphasis on the $Q$-values previously learned and stored in the $Q$-table. In this way, it is easier to get the global optimal routing. Therefore, we can conclude that designing appropriate discount factor gives non-negligible improvement of the accumulated rewards during the learning process, which validates our analysis.

## VII. Conclusion

In this paper, we have investigated the intelligent routing with maximum utility via $Q$-learning in CU-SWARM for emergency communications. We integrate the CR with UAV swarm to build the CU-SWARM in the aerial sensing layer of the three-layer hierarchical architecture, which consists of a UAV swarm for aerial sensing and an HAP for aerial access. We combine both the routing metric and the candidate UAV selection optimization policy to formulate the reward function. In particular, we characterize the routing metric by maximizing the utility, which is derived by balancing the achievable rate and the residual energy of UAV. We present the circular sector with the location, arc, and direction constraints by setting the central angle to optimize the candidate UAV selection. Finally, we develop a low-complexity iterative algorithm via the dynamic learning rate for updating $Q$-values during the training process to achieve a fast convergence speed. Our results have demonstrated that the proposed algorithm is convergent, and also shown that significant gains can be brought by the proposed algorithm over the benchmark schemes in terms of the accumulated rewards.

## References


[1] L. Zhang, H. Zhang, C. Guo, H. Xu, L. Song, and Z. Han, "Satellite-aerial integrated computing in disasters: User association and offloading decision," in *Proc. IEEE ICC*, Dublin, Ireland, Jun. 2020.

[2] S. Zhang and J. Liu, "Analysis and optimization of multiple unmanned aerial vehicle-assisted communications in post-disaster areas," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12049–12060, Dec. 2018.

[3] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Stochastic geometry study on device-to-device communication as a disaster relief solution," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3005–3017, May 2016.

[4] D.-H. Tran, V.-D. Nguyen, S. Chatzinotas, T. X. Vu, and B. Ottersten, "UAV relay-assisted emergency communications in IoT networks: Resource allocation and trajectory optimization," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 3, pp. 1621–1637, Mar. 2022.

[5] D. Wu, X. Sun, and N. Ansari, "An FSO-based drone assisted mobile access network for emergency communications," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1597–1606, Jul.-Sept. 2020.

[6] M. Casoni, C. A. Grazia, M. Klapez, N. Patriciello, A. Amditis, and E. Sdongos, "Integration of satellite and LTE for disaster recovery," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 47–53, Mar. 2015.

[7] B. Wang, Y. Sun, Z. Sun, L. D. Nguyen, and T. Q. Duong, "UAV-assisted emergency communications in social IoT: A dynamic hypergraph coloring approach," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7663–7677, Aug. 2020.

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

[9] 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, Sept. 2021.

[10] Y. Xu, Z. Liu, C. Huang, and C. Yuen, "Robust resource allocation algorithm for energy-harvesting-based D2D communication underlaying UAV-assisted networks," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 17161–17171, Dec. 2021.

[11] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervasive Comput.*, vol. 16, no. 1, pp. 24–32, Jan.-Mar. 2017.

[12] W. Chen, J. Liu, H. Guo, and N. Kato, "Toward robust and intelligent drone swarm: Challenges and future directions," *IEEE Netw.*, vol. 34, no. 4, pp. 24–32, Jul.-Aug. 2020.

[13] Y. Saleem, M. H. Rehmani, and S. Zeadally, "Integration of cognitive radio technology with unmanned aerial vehicles: Issues, opportunities, and future research challenges," *J. Netw. Comput. Appl.*, vol. 50, pp. 15–31, Apr. 2015.

[14] B. Shang, V. Marojevic, Y. Yi, A. S. Abdalla, and L. Liu, "Spectrum sharing for UAV communications: Spatial spectrum sensing and open issues," *IEEE Veh. Technol. Mag.*, vol. 15, no. 2, pp. 104–112, Jun. 2020.

[15] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 2, pp. 668–695, Second Quarter 2021.

[16] A. Paul and S. P. Maity, "Outage analysis in cognitive radio networks with energy harvesting and $Q$-routing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6755–6765, Jun. 2020.

[17] E. Nisioti and N. Thomos, "Fast $Q$-learning for improved finite length performance of irregular repetition slotted ALOHA," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 844–857, Jun. 2020.

[18] L. Zhang *et al.*, "A survey on 5G millimeter wave communications for UAV-assisted wireless networks," *IEEE Access*, vol. 7, pp. 117460–117504, Sept. 2019.

[19] N. Cheng, W. Xu, W. Shi, Y. Zhou, N. Lu, H. Zhou, X. Shen, "Air-ground integrated mobile edge networks: Architecture, challenges, and opportunities," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 26–32, Aug. 2018.

[20] Y. Zhou, N. Cheng, N. Lu, and X. S. Shen, "Multi-UAV-aided networks: Aerial-ground cooperative vehicular networking architecture," *IEEE Veh. Technol. Mag.*, vol. 10, no. 4, pp. 36–44, Dec. 2015.

[21] H. Ahmadinejad and A. Falahati, "Forming a two-tier heterogeneous air-network via combination of high and low altitude platforms," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1989–2001, Feb. 2022.

[22] Y. He *et al.*, "A NOMA-enabled framework for relay deployment and network optimization in double-layer airborne access VANETs," *IEEE Trans. Intell. Transp. Syst.*, Feb. 2022, Early Access.

[23] P. Qin, Y. Zhu, X. Zhao, X. Feng, J. Liu, and Z. Zhou, "Joint 3D-location planning and resource allocation for XAPS-enabled C-NOMA in 6G heterogeneous Internet of Things," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10594–10609, Oct. 2021.

[24] D. S. Lakew, A.-T. Tran, N.-N. Dao, and S. Cho, "Intelligent offloading and resource allocation in heterogeneous aerial access IoT networks," *IEEE Internet Things J.*, Mar. 2022, Early Access.

[25] J. Liu, L. Li, F. Yang, X. Liu, X. Li, X. Tang, and Z. Han, "Minimization of offloading delay for two-tier UAV with mobile edge computing," in *Proc. IEEE IWCMC*, Tangier, Morocco, Jun. 2019.

[26] A. Mukherjee, S. Misra, V. S. P. Chandra, and M. S. Obaidat, "Resource-optimized multiarmed bandit-based offload path selection in edge UAV swarms," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4889–4896, Jun. 2019.

[27] B. Liu, W. Zhang, W. Chen, H. Huang, and S. Guo, "Online computation offloading and traffic routing for UAV swarms in edge-cloud computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8777–8791, Aug. 2020.

[28] H. Song, L. Liu, B. Shang, S. Pudlewski, and E. S. Bentley, "Enhanced flooding-based routing protocol for swarm UAV networks: Random network coding meets clustering," in *Proc. IEEE INFOCOM*, Vancouver, BC, Canada, May 2021.

[29] 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. 3, pp. 8958–8976, Oct. 2019.

[30] T. Li *et al.*, "A mean field game-theoretic cross-layer optimization for multi-hop swarm UAV communications," *J. Commun. Netw.*, vol. 24, no. 1, pp. 68–82, Feb. 2022.

[31] Z. Cao *et al.*, "Using reinforcement learning to minimize the probability of delay occurrence in transportation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2424–2436, Mar. 2020.

[32] F. Li, X. Song, H. Chen, X. Li, and Y. Wang, "Hierarchical routing for vehicular ad hoc networks via reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1852–1865, Feb. 2019.

[33] L. Luo, L. Sheng, H. Yu, and G. Sun, "Intersection-based V2X routing via reinforcement learning in vehicular Ad Hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5446–5459, Jun. 2022.

[34] W. Zhang and Y. Ye, "A table-free approximate $Q$-learning-based thermal-aware adaptive routing for optical NoCs," *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, vol. 40, no. 1, pp. 199–203, Jan. 2021.

[35] C. Jiang, Y. Chen, K. R. Liu, and Y. Ren, "Renewal-theoretical dynamic spectrum access in cognitive radio network with unknown primary behavior," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 406–416, Mar. 2013.

[36] H. Kim and K. Shin, "Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 5, pp. 533–545, May 2008.

[37] Z. Ma, B. Ai, R. He, G. Wang, Y. Niu, and Z. Zhong, "A wideband non-stationary air-to-air channel model for UAV communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1214–1226, Feb. 2020.

[38] M. Gao, B. Zhang, and L. Wang, "A dynamic priority packet scheduling scheme for post-disaster UAV-assisted mobile ad hoc network," in *Proc. IEEE WCNC*, Nanjing, China, Mar.-Apr. 2021.

[39] G. Yang *et al.*, "Cooperative spectrum sensing in heterogeneous cognitive radio networks based on normalized energy detection," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1452–1463, Mar. 2016.

[40] L. Zhang, J. Hu, C. Guo, and H. Xu, "Dynamic power optimization for secondary wearable biosensors in e-healthcare leveraging cognitive WBSNs with imperfect spectrum sensing," *Future Gener. Comput. Syst.*, vol. 112, pp. 67–92, Nov. 2020.

[41] Y. Xu, H. Xie, Q. Wu, C. Huang, and C. Yuen, "Robust max-min energy efficiency for RIS-aided HetNets with distortion noises," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1457–1471, Feb. 2022.

[42] Y. Wu, F. Zhou, Q. Wu, Y. Huang, and R. Q. Hu, "Resource allocation for IRS-assisted sensing-enhanced wideband CR networks," in *Proc. IEEE ICC Workshops*, Montreal, QC, Canada, Jun. 2021.

[43] T. Liu *et al.*, "Measurement-based characterization and modeling for low-altitude UAV air-to-air channels," *IEEE Access*, vol. 7, pp. 98832–98840, Jul. 2019.

[44] 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.

[45] D. S. Lakew, U. Sa'ad, N.-N. Dao , W. Na , and S. Cho, "Routing in flying ad hoc networks: A comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 2, pp. 1071–1120, Second Quarter 2020.

[46] X.-L. Huang, Y.-X. Li, Y. Gao, and X.-W. Tang, "$Q$-learning-based spectrum access for multimedia transmission over cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 110–119, Mar. 2021.

[47] J. Wang, C. Jiang, K. Zhang, X. Hou, Y. Ren, and Y. Qian, "Distributed $Q$-learning aided heterogeneous network association for energy-efficient IIoT," *IEEE Trans. Industr. Inform.*, vol. 16, no. 4, pp. 2756–2764, Apr. 2020.