IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 16, 2020, accepted December 26, 2020, date of publication December 30, 2020,
date of current version January 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048021

Intelligent Trajectory Design for Secure Full-Duplex MIMO-UAV Relaying Against Active Eavesdroppers: A Model-Free Reinforcement Learning Approach

MILAD TATAR MAMAGHANI, (Graduate Student Member, IEEE), AND YI HONG, (Senior Member, IEEE)
Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia
Corresponding author: Milad Tatar Mamaghani (milad.tatarmamaghani@monash.edu)

This work was supported in part by the Australian Research Council under Grant DP200100096.

ABSTRACT Unmanned aerial vehicle (UAV) assisted wireless communication has recently been recognized as an inevitably promising component of future wireless networks. Particularly, UAVs can be utilized as relays to establish or improve network connectivity thanks to their flexible mobility and likely line-of-sight channel conditions. However, this gives rise to more harmful security issues due to potential adversaries, particularly active eavesdroppers. To combat active eavesdroppers, we propose an artificial-noise beamforming based secure transmission scheme for a full-duplex UAV relaying scenario. In the considered scheme, we investigate a UAV-relay equipped with multiple antennas to securely serve multiple ground users in the presence of randomly located active eavesdroppers. We formulate a novel average system secrecy rate (ASSR) maximization problem under some quality of service (QoS) and mission time constraints. Since the ASSR optimization problem is too hard to solve by conventional optimization methods due to the unavailability of the environment's dynamics and complex model, we develop some model-free reinforcement learning-based algorithms, i.e., Q-learning, SARSA, Expected SARSA, Double Q-learning, and SARSA(λ), to efficiently solve the problem without substantial UAV-network data exchange. Using the proposed algorithms, we can maximize ASSR via finding an optimal UAV trajectory and proper resource allocation. Simulation results demonstrate that all the proposed learning-based algorithms can train the UAV-relay to learn the environment by iterative interactions, thus finding an optimal trajectory, intelligently. Particularly, we find that SARSA(λ) based proposed algorithm with $\lambda = 0.1$ outperforms the others in terms of the ASSR.

INDEX TERMS UAV communications, full-duplex relaying, physical layer security, artificial noise injection, average system secrecy rate, trajectory optimization, reinforcement learning.

I. INTRODUCTION

The emerging AI driven 6G wireless communication networks have been envisioned to be an enabling technology of IoE, wherein the networked connection between people, process, data, and things is anticipated to be autonomously determined [1], [2]. Therefore, the ever-increasing demand for seamless and ubiquitous connectivity as well as high data rate transmission serving an exponentially increasing number of users are amongst the most critical challenges. In light of this, UAVs have been recognized as one of the key components of such networks due to their unique attributes: cost-effective, flexible deployment, maneuverability, and versatility [3], [4]. As a result, UAVs can be dispatched to avoid environmental obstacles and to provide seamless connectivity and reliable communications to a massive number of users.

Wireless applications of UAVs can be categorized into the following paradigms: one is for on-demand deployment as airborne platforms such as mobile BSs or relays to expand coverage and provide wireless connectivity in

The associate editor coordinating the review of this manuscript and approving it for publication was Moayad Aloqaily.

TABLE 1. List of acronyms.

3D	Three dimensional	6G	6th generation
AE	Active eavesdropper	AG	Air-ground
AI	Artificial intelligence	ANI	Artificial noise injection
ASSR	Average system secrecy rate	AWGN	Additive white Gaussian noise
BS	Base station	DF	Decode-and-forward
DQN	Deep Q-network	DRL	Deep reinforcement learning
FD	Full-duplex	FDD	Frequency division duplexing
i.i.d	Independent and identically distributed	IoE	Internet of everything
IoT	Internet of things	ISSR	Instantaneous system secrecy rate
LoS	Line-of-sight	MDP	Markov decision process
MIMO	Multiple-input-multiple-output	ML	Machine learning
MRC	Maximum ratio combining	PLS	Physical layer security
QoS	Quality of service	RL	Reinforcement learning
SCA	Successive convex approximation	SINR	Signal-to-interference-plus-noise ratio
TD	Temporal difference	TDMA	Time division multiple access
UAV	Unmanned aerial vehicle	UR	UAV-relay

densely crowded areas where the current infrastructures are encountering with some challenges to meet all the concurrent requests [5], [6], or in hazardous environments where no communication infrastructure is in full operation [7]; another is for data collection/dissemination due to their high mobility and low-cost operation for the UAV-IoT applications [8]–[10]; and the last one is for serving as aerial users or cellular-connected UAVs, receiving service from the terrestrial stations and cooperating multiple UAVs in the sky leading to information fusion and resources complementation to fulfill a common mission [11]–[13].

Despite the aforementioned advantages, the open nature of UAVs' AG links inevitably makes such systems vulnerable to various malicious attacks [14] such as eavesdropping, particularly active eavesdropping, wherein the adversary simultaneously performs both information eavesdropping and malicious jamming. If employed by illegitimate parties, hostile UAVs can even pose, benefiting from their salient attributes, more detrimental security threats to legitimate transmissions [15]. Therefore, wireless security is of crucial requirements for such UAV-aided wireless systems, and so, there exist various significant security challenges in the design of UAV-aided wireless communications to be addressed [16], [17]. Typically, in the network layer of wireless systems, cryptography techniques have been applied for information safeguarding, but for physical layer wireless communications, PLS approaches have been widely investigated, and recognized as one of the promising security countermeasures, especially for confidentiality. Since PLS can exploit the physical characteristics of wireless media without the need for complex encryption procedure, and more importantly, employing traditional cryptography techniques may not even lead to satisfactory confidential performance in resource-constraint aerial platforms [18], [19]. The notion of PLS, first introduced by Wyner's seminal work in [20], lies in a wiretap channel model, which guarantees that confidential communication can be established between legitimate users, provided that the eavesdropper's channel capacity is a degraded version of the legitimate user's one. Since then, various PLS techniques have been developed for terrestrial wireless communications with a three-category classification: secure channel coding design, channel-based adaptation PLS, and ANI techniques (see [21] and references therein).

A. RELATED WORKS AND MOTIVATIONS

Recently, some research works have investigated PLS for secure UAV communications. For example, in [22], PLS-based secure UAV-enabled communications have been developed via joint trajectory design and power control. In [23], a secure UR-based communication scheme via destination-assisted cooperative jamming has been proposed. In [24], the authors have studied a secure multi-UAV system with wireless energy harvesting in terms of efficient trajectory design and communication resource allocations for average secrecy rate maximization. The authors in [25] have explored employing an ANI-based secure two-phase transmission protocol for a single-antenna UAV system operating as an aerial BS, and then jointly optimized UAV's trajectory, network transmission power, and power allocation factor over a given time horizon. Further, secure energy-efficient power control and trajectory co-design has been investigated for UAV-enabled direct transmission [26], and UAV-assisted mobile relaying [27], [28]. In [29], the authors have studied a joint location-based 3D beamforming and trajectory design for the downlink multiple-antenna UAV relaying, and then proposed a heuristic-based iterative algorithm to improve the secrecy outage probability of the system.

The majority of recent research works have mainly focused on simple direct transmission [24], [29] or half-duplex UAV relaying [30], [31]. Recently, FD transmissions that double spectrum efficiency have attracted notable research interests to adopt at legitimate nodes, e.g., [32], [33]. Specifically, for non-security purposes, a UR-based FD system has been considered in [32] for the joint design of beamforming and power allocation with a fixed circular UR's

trajectory while using a DF relaying protocol. We note that DF relaying refers to a type of transmission protocol used in the communications between a source and a destination aided by one (or more) intermediate relay nodes, where the relay node decodes, remodulates, and then retransmits the received signal to the destination. Another type of relaying architecture is called the amplify-and-forward relaying, wherein the relay node simply forwards a scaled version of the received signal without decoding. Further, in [33] the trajectory design and resource allocation of a similar system model has been explored to minimize outage probability. For security-based FD-operated UAV communications, in [34], the authors have considered an FD system with an untrusted UR, and then studied secrecy outage and average secrecy rate performance metrics, wherein the untrusted relaying refers to the case when the intermediate relaying conducts adversarial activity during communication facilitation [35]–[37]. It should be worth pointing out that FD malicious nodes can potentially pose severe security attacks compared to their passive counterparts. The authors in [38] have proposed an ANI-based secure uplink UAV transmission in the presence of a multiple-antenna FD-operated AE. They have analyzed a hybrid outage secrecy metric, which enables to capture the joint effect of connection and secrecy outage probabilities.

We note that in all the abovementioned research works for UAV's trajectory design, e.g., [24], [26], [27], [29], [31], [33], standard optimization techniques such as SCA have been employed under the assumption of a known network model and a perfect knowledge of flight's dynamics. However, this assumption can be somehow impractical inasmuch as a precise mathematical model can hardly be formed, owing to the fact that the UAV-network topology frequently demands information exchange between the UAV and the core network. Consequently, the expression of the objective function to be optimized or the constraints might be either unavailable or obtaining their gradients analytically becomes almost impossible [39]. Hence, other optimization approaches are required to deal with such complex problems. One promising approach can be the model-free RL techniques [40] that can reduce the online computational complexity. To that end, in [41], the authors have proposed an RL algorithm for a multi-UAV cooperative system, aiming at maximizing the sum rate metric via trajectory design and resource management. The authors in [42] have considered exploiting RL algorithms to optimize UAV's trajectory for maximum data collection in a sensor network under some QoS constraints. However, these developments have aimed at only reliability aspects of UAV communications, and PLS security aspects have not yet been fully explored.

B. OUR CONTRIBUTIONS

Driven by this demand, in this work, we propose a secure FD-operated MIMO-UAV relaying communication scheme in the presence of multiple AEs, wherein the source is a multiple-antenna BS. We assume that both BS and UR adopt ANI-based beamforming, and AEs are equipped with double antennas to perform concurrent reception and transmission. Then, our design goal is to maximize the ASSR under some QoS and mission time requirements. To achieve our goal, we develop some applicable RL-based algorithms for an adaptive trajectory design, enabling the flying UR to autonomously find the optimal path to complete the mission. Detailed contributions are summarized below.

- In our design, we target at maximizing the ASSR of the system under some conditions for fulfilling the UR's flying mission as well as combating the active eavesdropping issue. Besides, we take into account the collision avoidance between the flying UR and environmental obstacles for safety purposes.
- The original optimization problem of ASSR is, however, hard to solve due to the non-convex complex model of the objective function and some associated constraints. To tackle this problem, we devise some efficient model-free RL-based algorithms, i.e., Q-learning, SARSA, Double Q-learning, Expected SARSA, and SARSA(λ). Via the proposed algorithms, we can train the UR to find its optimal path via environmental interactions and decision-updating using the feedback/reward received for the trajectory design purpose, meanwhile forming a simple resource allocation problem that can partially contribute to generating the reward function. We can see that our approach significantly diverges from those in [22], [24], [43], where the problems were trackable and mathematical optimizations have been applied.
- Finally, we discuss the convergence and complexity of the proposed adaptive trajectory design algorithms under the considered settings. Via extensive simulations, we demonstrate that these algorithms can effectively improve the considered ASSR performance, and their convergence rates to their optimal policies are also desirable.

The rest of this paper is organized as follows. In Section II, we detail the system model and signal representations for the proposed secure UR-based FD system in the presence of multiple randomly-located AEs, wherein the BS and UR both adopt the MIMO-based ANI beamforming. Problem formulation is then given in Section III, followed by our model-free RL-based solutions in Section IV. Section V is devoted to numerical results and discussions about the performance of the developed solutions. Finally, the conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider a UAV-assisted mobile relaying system, as depicted in Fig. 1, wherein a UAV is employed as a mobile relay to provide an enhanced service and secure connectivity for multiple ground users. Particularly, we consider a BS, denoted as $\mathbb{S}$, which intends to secretly communicate with the remote ground users with the help of a UR, denoted as $\mathbb{U}$, in the presence of multiple terrestrial AEs. We assume that $\mathbb{S}$ and $\mathbb{U}$ are equipped with N_s and N_u transmitting

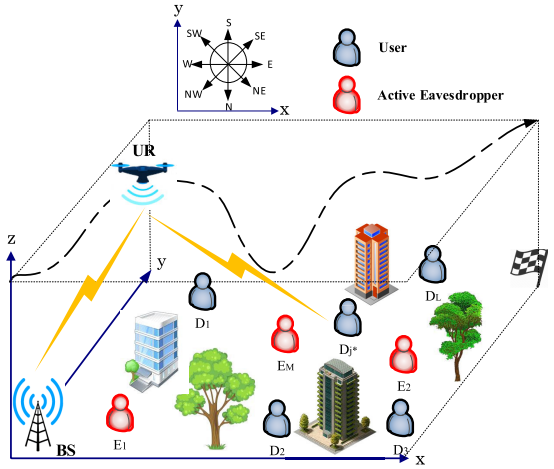

FIGURE 1. Illustration of the considered MIMO-UR system model with multiple users and in the presence of some AEs.

antennas, respectively, and $\mathbb{U}$ has also one receiving antenna. Further, we assume there are L single antenna ground users, denoted as $\mathcal{D} = \{\mathbb{D}_1, \mathbb{D}_2, \cdots, \mathbb{D}_L\}$, and M double-antenna AEs, represented by $\mathcal{E} = \{\mathbb{E}_1, \mathbb{E}_2, \cdots, \mathbb{E}_M\}$, all of which are randomly distributed across a rectangular $\mathcal{R}_w \times \mathcal{R}_l$ region. We insist that AEs, compared to conventional passive eavesdroppers, may pose stronger eavesdropping attacks as they can, in addition to overhearing transmit confidential messages, actively deteriorate the capacity of the main channel, i.e., the quality of received signals at the legitimate nodes, by malicious jamming transmissions. Further, being equipped with two antennas, we assume that each AE operates in the FD mode such that utilizes one antenna for eavesdropping purpose and the other for jamming transmission, simultaneously. To guarantee the security and reliability of the transmission in the considered system, we employ a UAV to act as a mobile DF relay. The goal of the UR is to fly over the region from a pre-specified starting location and stop at the pre-established final destination (ignoring the landing process, this point is depicted by a flag in Fig. 1) for each flight. Note that obstacles such as high-rise buildings and trees represent the forbidden region due to, for example, the possibility of collision, through which the employed low-altitude UR should avoid passing during the mission. In order for the UR to sequentially relay the data and provide service for multiple users, the total flight duration T, which should not go beyond a maximum allowed feasible mission time T_{max}, is divided into multiple sufficiently-small time slots, at each of them only one user is scheduled to receive data from $\mathbb{U}$ according to the TDMA technique, and besides that, the UR applies FDD protocol at each time slot, allocating equally-shared bandwidth for data transmission and reception. Now, we detail the channel assumptions and signal representations of the proposed system in what follows.

A. CHANNEL ASSUMPTIONS

First off, we model the location of network nodes by 3D Cartesian coordinate system. As such, without loss of generality, due to mission requirements, the UR's predetermined initial and final locations, which may refer to rising and landing sites, are represented as $Q_i = [x_I, y_I, H_u]^T$ and $Q_f = [x_F, y_F, H_u]^T$, respectively, wherein H_u denotes the operating fixed altitude of the UR. The location of flying UR at time slot t is denoted as $Q_u(t) = [x_u(t), y_u(t), H_u]^T$ with corresponding projected coordinate on the ground (x-y plane) as $q_u(t) = [x_u(t), y_u(t)]^T$. Further, the node $\mathbb{S}$ is located at the ground with x-y coordinate $q_s = [x_s, y_s]^T$, the location of randomly located ground users and AEs are also represented as $\mathbf{q}_d = [\mathbf{x}_d, \mathbf{y}_d]^T$, $\mathbf{q}_e = [\mathbf{x}_e, \mathbf{y}_e]^T$, respectively, where for $i \in \{d, e\}$, $\mathbf{x}_i \in \mathbb{R}^{N_i \times 1}$, $\mathbf{y}_i \in \mathbb{R}^{N_i \times 1}$, and so $\mathbf{q}_i \in \mathbb{R}^{2 \times N_i}$, in which N_d and N_e represent the cardinality of sets $\mathcal{D}$ and $\mathcal{E}$, respectively. Note that the j-th column (where $j = 1, 2, \cdots, N_i$) of the matrix $\mathbf{q}_i$, denoted as $\mathbf{q}_i^{(j)}$, represents the x-y coordinate of the j-th terrestrial device of type i. As a result, the instantaneous distance between the flying UR and terrestrial communication node i_j can be represented as $d_{ui_j}(t)$ which is given by

$$d_{ui_j}(t) = \sqrt{\|q_u(t) - \mathbf{q}_i^{(j)}\|^2 + H_u^2}, \tag{1}$$

Likewise, the instantaneous distance between $\mathbb{S}$ and the UR is represented as

$$d_{su}(t) = \sqrt{\|q_u(t) - q_s\|^2 + H_u^2}, \tag{2}$$

Plus, defining $G \triangleq \{\mathbb{S}\} \cup \mathcal{D} \cup \mathcal{E}$, the Euclidean distance between any pair of terrestrial nodes $a, b \in G$ is denoted as d_{ab} and calculated by

$$d_{ab} = \|q_a - q_b\|, \tag{3}$$

where $q_{a(b)} \in \{q_s, q_i^{(j)} \ \forall i, j\}$.

1) LARGE-SCALE ATTENUATION

In this work, we consider that each terrestrial device has an LoS path towards the UR with a given probability as [4]. This LoS probability is determined by the environment, locations of the terrestrial devices, and the UR. Thus, we express the LoS probability between the terrestrial node $g \in G$ and the UR as

$$P_{gu}^{\mathrm{L}}(t) = \frac{1}{1 + \omega_1 \exp(-\omega_2(\theta_{gu}(t) - \omega_1))}, \tag{4}$$

where $\theta_{gu}(t) = \frac{180}{\pi} \arcsin\left(\frac{H_u}{d_{gu}(t)}\right)$ represents the elevation angle in degree, wherein $d_{gu}(t)$ is given by (1) and (2), and the parameters $\omega_1, \omega_2 > 0$ are determined by the environment. The non-LoS probability of the link between the node g and the UR can be simply expressed as $P_{gu}^{\mathrm{N}}(t) = 1 - P_{gu}^{\mathrm{L}}(t)$. Further, we model the elevation-angle dependent probabilistic path loss component as

$$\eta_{gu} = P_{gu}^{\mathrm{L}}(t)\eta_L + P_{gu}^{\mathrm{N}}(t)\eta_N, \tag{5}$$

where $\eta_N > \eta_L \geq 2$. We note that according to (4) and (5) for fixed altitude of UR, the LoS probability will decrease

as the UR goes away from the terrestrial node g and accordingly, the channel will experience a larger path loss attenuation. Therefore, the large-scale attenuation between any two nodes m and n can be given by

$$L_{mn} = \begin{cases} L_{gu} = \beta_0 d_{gu}^{-\eta_{gu}}(t), & \forall g \in G \\ L_{ab} = \beta_0 d_{ab}^{-\eta_N}, & \forall a, b \in G \end{cases} \tag{6}$$

where β_0 represents the channel power gain at the reference distance of 1m.

2) SMALL-SCALE FADING

We further consider that the AG channels experience Rician fading for LoS propagation conditions and Rayleigh fading for non-LoS propagation conditions. As such, we express the time varying channel between the BS and the UR as

$$\mathbf{h}_{su}(t) = \sqrt{\frac{K_{su}(t)}{K_{su}(t)+1}}\mathbf{h}_{su}^{o}(t) + \sqrt{\frac{1}{K_{su}(t)+1}}\mathbf{h}_{su}^{r}(t), \tag{7}$$

where $K_{su}(t) = \omega_3 \exp(\omega_4 \theta_{su}(t))$ with ω_3 and ω_4 being constant parameters, denotes the Rician K-factor of the channel, $\mathbf{h}_{su}^{o}$ represents the LoS component of the corresponding channel, defined as

$$\mathbf{h}_{su}^{o}(t) = \frac{1}{\sqrt{N_s}}\left[1, \cdots, \exp(-\mathrm{j}2\pi\,(N_s-1)\,\delta_s \sin(\beta_{su}(t))\cos(\theta_{su}(t))\right], \tag{8}$$

where $\beta_{su}(t)$ denotes the time-varying azimuth angle between the BS and the UR, δ_s denotes the constant antenna spacing in wavelength at the BS. Further, $\mathbf{h}_{su}^{r}(t)$, denoting the scattered component of channel vector between $\mathbb{S}$ and the UR, each element of which follows quasi-static i.i.d complex Gaussian random variable with zero mean and unit variance, i.e., $\mathbf{h}_{su}^{r} \sim \mathcal{CN}\left(\mathbf{0}_{1\times N_s}, \mathbf{I}_{N_s}\right)$.

Furthermore, the air-ground channel vector between the UR and the single receiving antenna terrestrial node i_j is represented by $\mathbf{h}_{ui_j}(t)$ as

$$\mathbf{h}_{ui_j}(t) = \sqrt{\frac{K_{ui_j}(t)}{K_{ui_j}(t)+1}}\mathbf{h}_{ui_j}^{o}(t) + \sqrt{\frac{1}{K_{ui_j}(t)+1}}\mathbf{h}_{ui_j}^{r}(t), \tag{9}$$

where $K_{ui_j}(t) = \omega_3 \exp\left(\omega_4 \theta_{ui_j}(t)\right)$ denotes the Rician K-factor of the channel between the UR and the node i_j, $\mathbf{h}_{ui_j}^{o}$ denotes the corresponding LoS component, defined as

$$\mathbf{h}_{ui_j}^{o}(t) = \frac{1}{\sqrt{N_u}}\left[1, \cdots, \exp(-\mathrm{j}2\pi\,(N_u-1)\,\delta_u \sin(\beta_{ui_j}(t))\cos(\theta_{ui_j}(t))\right], \tag{10}$$

where $\beta_{ui_j}(t)$ denotes the time-varying azimuth angle between the UR and the ground node i_j, δ_u denotes the constant antenna spacing in wavelength at the UR. and $\mathbf{h}_{ui_j}^{r} \sim \mathcal{CN}\left(\mathbf{0}_{1\times N_u}, \mathbf{I}_{N_u}\right)$. We define the $N_e \times N_u$ matrix $\mathbf{H}_{ue} = [\mathbf{h}_{ue_1}; \cdots; \mathbf{h}_{ue_{N_e}}]$. Since each AE has one antenna for jamming transmission, therefore, the corresponding channel from j-th AE $e_j \in \mathcal{E}$ to the UR at time slot t can be represented as

$$h_{e_j u}(t) = \sqrt{\frac{K_{e_j u}(t)}{K_{e_j u}(t)+1}} + \sqrt{\frac{1}{K_{e_j u}(t)+1}}h_{e_j u}^{r}(t), \tag{11}$$

where $K_{e_i u}(t)$ denotes the corresponding Rician K-factor and $h_{e_i u}^{r} \sim \mathcal{CN}(0, 1)$. Therefore, we can define the $1 \times N_e$ vector $\mathbf{h}_{eu} \triangleq [h_{e_1 u}(t), h_{e_2 u}(t), \cdots, h_{e_{N_e} u}(t)]$. Additionally, the $M \times L$ channel matrix from the AEs to the users can be represented as

$$\mathbf{H}_{ed} = \begin{bmatrix} h_{e_1 d_1} & h_{e_1 d_2} & \cdots & h_{e_1 d_L} \\ \vdots & \ddots & & \\ h_{e_M d_1} & h_{e_M d_2} & & h_{e_M d_L} \end{bmatrix}, \tag{12}$$

where $\mathbf{H}_{ed}$ is subject to i.i.d Rayleigh fading with normalized channel power gains. Next, let the link between the BS and the terrestrial node i_j be the $1 \times N_s$ channel vector $\mathbf{h}_{si_j}$, such that $\mathbf{h}_{si_j} \sim \mathcal{CN}\left(\mathbf{0}_{1\times N_s}, \mathbf{I}_{N_s}\right)$. Further, the self-interference channel from the UR's transmitting antennas to the receiving antenna is characterized as the $1 \times N_u$ vector $\sqrt{\rho_u}\mathbf{h}_{uu}$ wherein $\rho_u \in [0, 1]$ characterizes the effect of imperfect self-interference cancellation such that $\rho_u = 0$ implies zero self-interference and $0 < \rho_u \leq 1$ takes the level of self-interference into account. Further, $\mathbf{h}_{uu} \sim \mathcal{CN}\left(\mathbf{0}_{1\times N_u}, \mathbf{I}_{N_u}\right)$. Likewise, the self-interference channels of AEs as well as the cross-interference channels arising from the other AEs are respectively denoted as $h_{e_p} \sim \mathcal{CN}(0, 1)$ and $h_{e_p e_q} \triangleq h_{e_{pq}} \sim \mathcal{CN}(0, 1)$ with $e_{p(q)} \in \mathcal{E}, \forall p, q\ (p \neq q) \in \{1, \cdots, N_e\}$, and can be represented as the $N_e \times N_e$ channel matrix given by

$$\mathbf{H}_{ee} = \begin{bmatrix} \sqrt{\rho_{e_1}}h_{e_1} & h_{e_{12}} & \cdots & h_{e_{1N_e}} \\ \vdots & \ddots & & \\ h_{e_{N_e 1}} & h_{e_{N_e 2}} & & \sqrt{\rho_{e_{N_e}}}h_{e_{N_e}} \end{bmatrix}, \tag{13}$$

where $0 \leq \rho_{e_p} \leq 1$ demonstrate the self-interference factors of the AE e_p due to FD operation.

B. USER SELECTION

Since at each time slot t, $\mathbb{U}$ forwards the intended confidential message to one scheduled user via performing FD relaying, we let $\zeta_j(t) \in \{0, 1\}$ be a binary variable for user $\mathbb{D}_j$ where $j \in \{1, \cdots, L\}$ at time slot t to indicate user scheduling, i.e., $\zeta_j(t) = 1$ if user $\mathbb{D}_j$ is scheduled at time slot t and $\zeta_j(t) = 0$ otherwise. Therefore, we have the user scheduling constraint given as

$$\sum_{j=1}^{L} \zeta_j(t) \leq 1, \tag{14}$$

In this work, we consider the user selection criterion based upon the best channel condition of the second hop of information relaying. Thus, the scheduled user at time slot t can be obtained as

$$j^{\star} = \arg \max_{\forall \mathbb{D}_j \in \mathcal{D}} \|h_{ud_j}(t)\|, \tag{15}$$

where j^* denotes the index of the selected user $\mathbb{D}_{j^\star}$ and we denote it as the destination node D.

C. MIMO-ENABLED ARTIFICIAL NOISE BEAMFORMING

In order to establish secure end-to-end transmission, we adopt ANI technique with the confidential messages for secure transmission at both the BS and UR according to MIMO-beamforming, as shown in Fig. 2. In this scheme, we assume that the source transmits noise-like signals in addition to information signals, in order to confuse the malicious nodes. As such, we design the transmitted unit-power signal x_s from the BS as

$$x_s = \left[\mathbf{w}_s^{N_s\times 1} \;\; \mathbf{W}_{s,an}^{N_s\times(N_s-1)}\right]\left[\begin{matrix} t_s(\tau) \\ \mathbf{t}_{s,an}^{(N_s-1)\times 1}\end{matrix}\right], \tag{16}$$

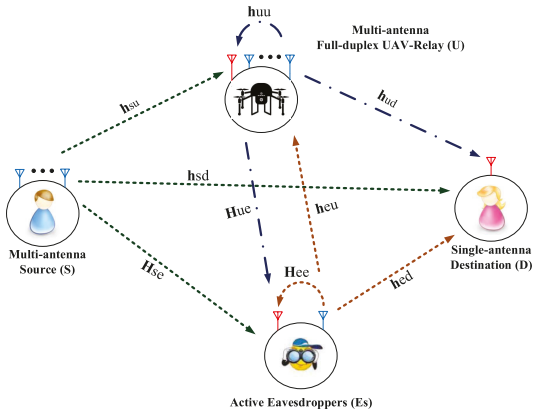

FIGURE 2. Proposed secure transmission at a given time slot towards the selected user D based on MIMO-enabled ANI for FD-operated UAV-assisted relaying and under active eavesdropping.

where $\mathbf{w_s} = \frac{\mathbf{h}_{su}^\dagger}{||\mathbf{h}_{su}||}$ is chosen to maximize the information signal transmission towards the UR, wherein $(\cdot)^\dagger$ indicates the transpose conjugate operator, $\mathbf{W}_{s,an}$ is assumed to be the projection matrix onto the null space of $\mathbf{h}_{su}$, i.e., $\mathbf{h}_{su}\mathbf{W}_{s,an} = \mathbf{0}$, and therefore, it is evident that the columns of the first matrix in the right-hand side of (16), or the so-called beamforming matrix, form orthonormal basis. Note that following Eigen-decomposition of the matrix $\mathbf{H}_{su} = \mathbf{h}_{su}^\dagger\mathbf{h}_{su}$, $\mathbf{w_s}$ can be chosen as the eigenvector corresponding to the maximum eigenvalue, and the remaining eigenvectors can form the matrix $\mathbf{W}_{s,an}$. It should be mentioned that the designed beamforming matrix aims at degrading the quality of the received signal at the unintended devices while improving the quality of the reception at the UR. Further, t_s denotes scalar information signal to be sent securely to the end-user and $\mathbf{t}_{s,an}$ represents the ANI vector at the BS with dimensions $(N_s-1)\times 1$. Letting α_s where $0 < \alpha_s \leq 1$ be the power allocation factor between information signal and ANI at the BS, we have $\mathbb{E}\{|t_s|^2\} = \alpha_s$ and $\mathbb{E}\{\mathbf{t}_{s,an}\mathbf{t}_{s,an}^\dagger\} = \frac{1-\alpha_s}{N_s-1}\mathbf{I}_{N_s-1}$, wherein $\mathbb{E}\{\cdot\}$ indicates the expectation operator.

Likewise, the UR performs ANI to the previously decoded signal to be forwarded to the selected user D. Then, the forwarded information-bearing signal can be represented as

$$\mathbf{x}_u = \left[\mathbf{w}_u^{N_u\times 1} \;\; \mathbf{W}_{u,an}^{N_u\times(N_u-1)}\right]\left[\begin{matrix} t_u(\tau-1) \\ \mathbf{t}_{u,an}^{(N_u-1)\times 1}\end{matrix}\right], \tag{17}$$

where $\mathbf{w_u} = \frac{\mathbf{h}_{ud}^\dagger}{||\mathbf{h}_{ud}||}$ is chosen to maximize the information signal transmission towards the scheduled user D, $\mathbf{W}_{u,an}$ is chosen such that $\mathbf{h}_{ud}\mathbf{W}_{u,an} = \mathbf{0}$. Further, $t_s(\tau-1)$ denotes the previously decoded information signal, $\mathbf{t}_{u,an}$ represents the ANI vector at the BS with dimensions $(N_u-1)\times 1$. Letting α_u where $0 < \alpha_u \leq 1$ be the power allocation factor between information signal and ANI at the UR, we have $\mathbb{E}\{|t_u|^2\} = \alpha_u$ and $\mathbb{E}\{\mathbf{t}_{u,an}\mathbf{t}_{u,an}^\dagger\} = \frac{1-\alpha_u}{N_u-1}\mathbf{I}_{N_u-1}$.

D. TRANSMISSION PROTOCOL AND SIGNALS REPRESENTATION

Let P_s, P_u, and P_{e_j} with $j = \{1,\cdots,N_e\}$ be the transmission powers of the BS, UR, and AEs, respectively. Then, the received signal at the UR at time slot t can be represented as

$$y_u = \sqrt{P_s L_{su}}\mathbf{h}_{su}\mathbf{w}_s t_s + \sqrt{\rho_u P_u}\mathbf{h}_{uu}\mathbf{x}_u + \sum_{j=1}^{N_e}\sqrt{P_{e_j}L_{e_ju}}h_{e_ju}x_{e_j} + n_u, \tag{18}$$

where $n_u \sim \mathcal{CN}(0,\sigma_u^2)$ is the AWGN, $\mathbf{x}_u$ and x_{e_j} are the unit-power signals, i.e., $\mathbb{E}\{\|\mathbf{x}_u\|^2\} = 1$ and $\mathbb{E}\{|x_{e_j}|^2\} = 1$. Note that the first term in the right-hand side of (18) denotes the information-bearing signal, the second term is the residual self-interference at the UR, the third term denotes the disturbance arises from the AEs and their jamming transmission. The SINR can be given as

$$\Gamma_{su} = \frac{\gamma_{su}}{\gamma_{eu}+\gamma_{uu}+1}, \tag{19}$$

where

$$\gamma_{su} = \alpha_s P_s L_{su}\|\mathbf{h}_{su}\|^2/\sigma_u^2,$$
$$\gamma_{eu} = \sum_{j=1}^{N_e} P_{e_j}L_{e_j}|h_{e_ju}|^2/\sigma_u^2,$$
$$\gamma_{uu} = \rho_u P_u\|\mathbf{h}_{uu}\|^2/\sigma_u^2,$$

The received signal at the scheduled user D at time slot t is given by

$$y_d = \sqrt{P_u L_{ud}}\mathbf{h}_{ud}\mathbf{w}_u t_u + \sqrt{P_s L_{sd}}\mathbf{h}_{sd}\mathbf{x}_s + \sum_{j=1}^{N_e}\sqrt{P_{e_j}L_{e_jd}}h_{e_jd}x_{e_j} + n_d, \tag{20}$$

where the antenna noise n_d is modeled as the AWGN, i.e., $n_d \sim \mathcal{CN}(0,\sigma_d^2)$. Note that the first term in (20) is the information-bearing signal transmitted by the UR, the second term is the signal coming from the BS, however assuming the worst-case scenario such that the legitimate users have low-complex receivers which are unable to perform joint processing, the user treats this signal as an interference. In other

words, since the signals coming from the BS and the UR are inherently different, for example, due to having been encoded with different codebooks, so the user is only able to detect one signal. Besides that, it is common in the literature to consider solely the signal coming from the UR, however, in this work we take into account the direct transmission as well. Finally, the third term denotes the disturbance coming from the AEs. Hence, the SINR at D can be given as

$$\Gamma_d = \frac{\gamma_{ud}}{\gamma_{ed} + \gamma_{sd} + 1}, \tag{21}$$

where

$$\begin{aligned}
\gamma_{ud} &= \alpha_u P_u L_{ud} \|\mathbf{h}_{ud}\|^2/\sigma_d^2,\\
\gamma_{sd} &= \frac{P_s L_{sd}}{\sigma_d^2}\left(\alpha_s \|\mathbf{h}_{sd}\mathbf{w}_s\|^2 + \frac{1-\alpha_s}{N_s-1}\mathbf{h}_{sd}\mathbf{W}_{s,an}\mathbf{W}_{s,an}^{\dagger}\mathbf{h}_{sd}^{\dagger}\right),\\
\gamma_{ed} &= \sum_{j=1}^{N_e} P_{e_j} L_{e_j d} |h_{e_j d}|^2/\sigma_d^2.
\end{aligned}$$

It should be mentioned that $\forall \mathbb{E}_j \in \mathcal{E}$ may receive two copies of the information signal from the BS and the UR with some delay as the relay needs to first process the received signal before forwarding. In contrast to [44] and the assumption that we made for the scheduled user, where the direct transmission is treated as an interference, each AE $\mathbb{E}_j$ here is assumed to be able to fully combine these signals and perform a joint processing method such as an ideal Rake receiver [45], [46]. As such, $\mathbb{E}_j$ can appropriately co-phase and merge these two signals via applying MRC and perform more harmful eavesdropping attacks. Besides, we assume that the AEs are non-colluding, i.e., each AE decodes the received signals from the source and the UR without cooperating with other AEs. Consequently, the received signal at j-th AE, denoted by y_{e_j}, can be represented as

$$\begin{aligned}
y_{e_j} &= \sqrt{P_s L_{se_j}}\mathbf{h}_{se_j}(\mathbf{w}_s t_s + \mathbf{W}_{s,an}\mathbf{t}_{s,an})\\
&\quad + \sqrt{P_u L_{ue_j}}\mathbf{h}_{ue_j}(\mathbf{w}_u t_u + \mathbf{W}_{u,an}\mathbf{t}_{u,an})\\
&\quad + \sum_{i=1, i\neq j}^{N_e} \sqrt{P_{e_i} L_{e_i e_j}} h_{e_{ij}} x_{e_i} + \sqrt{\rho_{e_j} P_{e_j}} h_{e_j} x_{e_j} + n_{e_j},
\end{aligned} \tag{22}$$

where x_{e_j} is the unit-power jamming signal transmitted by other AEs, n_{e_j} is the AWGN at j-th AE where $n_{e_j} \sim \mathcal{CN}(0, \sigma_e^2)$. We assume that $\mathbb{E}_j$ applies MRC to effectively decode the received information, and hence, the SINR at D can be given as

$$\Gamma_{e_j} = \Gamma_{se_j} + \Gamma_{sue_j}, \tag{23}$$

where the SINR Γ_{se_j} is given by

$$\Gamma_{se_j} = \frac{\gamma_{se_j}}{\gamma_{se_j,an} + 1}, \tag{24}$$

in which

$$\begin{aligned}
\gamma_{se_j} &= \alpha_s P_s L_{se_j} \|\mathbf{h}_{se_j}\mathbf{w}_s\|^2/\sigma_{e_j}^2,\\
\gamma_{se_j,an} &= \frac{1-\alpha_s}{N_s-1} P_s L_{se_j}\mathbf{h}_{se_j}\mathbf{W}_{s,an}\mathbf{W}_{s,an}^{\dagger}\mathbf{h}_{se_j}^{\dagger}/\sigma_{e_j}^2,
\end{aligned}$$

and Γ_{sue_j} can be obtained as per the rules of DF protocol as

$$\Gamma_{sue_j} = \min\left(\Gamma_{su}, \Gamma_{ue_j}\right), \tag{25}$$

where Γ_{su} is given in (19) and Γ_{ue_j} can be represented as

$$\Gamma_{ue_j} = \frac{\gamma_{ue_j}}{\gamma_{ue_j,an} + \gamma_{e_{\neq j}e_j} + \gamma_{e_j} + 1}, \tag{26}$$

where

$$\begin{aligned}
\gamma_{ue_j,an} &= \frac{1-\alpha_u}{N_u-1} P_u L_{ue_j}\mathbf{h}_{ue_j}\mathbf{W}_{u,an}\mathbf{W}_{u,an}^{\dagger}\mathbf{h}_{ue_j}^{\dagger}/\sigma_{e_j}^2,\\
\gamma_{ue_j} &= \alpha_u P_u L_{ue_j}\|\mathbf{h}_{ue_j}\mathbf{w}_u\|^2/\sigma_{e_j}^2,\\
\gamma_{e_j} &= \rho_{e_j} P_{e_j}|h_{e_j}|^2/\sigma_{e_j}^2,\\
\gamma_{e_{\neq j}e_j} &= \sum_{i=1, i\neq j}^{N_e} P_{e_i} L_{e_i e_j}|h_{e_{ij}}|^2/\sigma_{e_j}^2.
\end{aligned}$$

III. PROBLEM FORMULATION

The achievable instantaneous system secrecy rate of the proposed UAV-based FD relaying scenario is defined, assuming normalized shared bandwidth in bit-per-second-per-Hertz (bit/s/Hz), as [47]

$$R_{sec}(t) = [I_D(t) - I_E(t)]^+, \tag{27}$$

where $[x]^+ \triangleq \max\{x, 0\}$, $I_D(t)$ is given by

$$I_D(t) = \sum_{i=1}^{L} \zeta_i(t)\log_2(1 + \Gamma_d(t)), \tag{28}$$

which represents the capacity of the main channel including both the direct and relaying links from the BS to the scheduled user D at time slot t, and $I_E(t)$ is given by

$$I_E(t) = \max_{\forall e\in\mathcal{E}} \log_2(1 + \Gamma_e), \tag{29}$$

which determines the Shannon capacity of the non-colluding eavesdropping links at time slot t. In this work, we aim at maximizing the ASSR during the mission time T by trajectory design and resource allocation. Thus, the optimization problem can be formulated as

$$\begin{aligned}
\underset{P_s,P_u,\alpha_s,\alpha_u,T,Q_u}{\text{maximize}} \quad & \frac{1}{T}\int_{t_0}^{t_0+T} R_{sec}(t)dt\\
s.t.\ \text{C1}:\ & R_{sec}(t) \geq R_{sec}^{th}, \quad \forall t\\
\text{C2}:\ & \int_{t_0}^{t_0+T} R_{sec}^{(j)}(t)dt \leq \bar{B}_j^{max}\ \forall j \in \mathcal{D}\\
\text{C3}:\ & 0 \leq P_s(t) \leq P_s^{\max}, \quad \forall t\\
\text{C4}:\ & 0 \leq P_u(t) \leq P_u^{\max}, \quad \forall t\\
\text{C5}:\ & 0 \leq \alpha_s(t), \alpha_u(t) \leq 1, \quad \forall t\\
\text{C6}:\ & 0 \leq x_u(t) \leq \mathcal{R}_l, \quad \forall t\\
\text{C7}:\ & 0 \leq y_u(t) \leq \mathcal{R}_w, \quad \forall t\\
\text{C8}:\ & Q_u(t_0) = Q_i, \quad Q_u(t_0+T) = Q_f,\\
\text{C9}:\ & \sqrt{(y_F - y_0)^2 + (x_F - x_0)^2} \leq T v_{max},\\
\text{C10}:\ & 0 < T \leq T_{max},
\end{aligned} \tag{30}$$

where the constraint C1 indicates the minimum instantaneous secrecy rate requirement, otherwise secrecy outage may occur, C2 is to ensure that amount of data securely received by the users does not go beyond users' capacity to make fairness service amongst them, C3 and C4 should be satisfied due to green communications and hardware limitations, C5 indicates the ANI factors limitation, constraints C6 and C7 are posed by the restricted flying region requirement, C8 arises due to the mission requirement in terms of pre-specified start and ending locations, and finally C9 ensures the feasibility of the mission, and C10 is for guaranteeing the mission completion goes not beyond a reasonably feasible maximum allowed time T_{max}. The problem (30) is too complicated to solve due to non-convex complex model of the objective function and constraints C1 and C2. Our approach is to employ some learning-based reinforcement techniques to tackle the problem. In the following section, we detail our RL-based solutions to approximately solve the original problem (30) after providing a brief introduction of RL fundamentals.

IV. REINFORCEMENT LEARNING BASED SOLUTION

A. PRELIMINARIES

Here, we first briefly explain the RL fundamentals (interested readers are encouraged to refer to excellent resources such as [40] for detailed discussions), by which we then reformulate our optimization problem in (30) for trajectory optimization of the proposed UR scenario and then efficiently solving via learning approximate solution approach.

RL problems can be mathematically studied based on the MDP frameworks which basically establish a relationship between interaction-based learning and goal achievement. Consequently, it is worthwhile to first recall some key, though abstract, components of the MDP frameworks in the following, which shall be explicitly semanticated later on. In MDPs there is a decision-maker or the so-called learning agent that continually interacts with the environment over a sequence of discrete time steps $t = 1, 2, 3, \cdots$ via taking an *action*, then receiving some feedback signal from the environment — termed as *reward*, and then being presented in a new situation or *state*. The objective of the agent is to maximize the received rewards over time. In our work, we consider the finite MDP wherein the number of elements in the state-action-reward set, i.e., $\{\mathcal{S}, \mathcal{A}, \mathcal{R}\}$ is finite.

The interaction of the learning agent with the environment is well visualized in Fig. 3. Particularly, given $s \in \mathcal{S}$ be the agent's current state, it takes an action $a \in \mathcal{A}$ and then goes to the next state $s' \in \mathcal{S}$, observes the environment, and receives a numerical reward $r \in \mathcal{R} \subset \mathbb{R}$ following to the action taken. According to the finite MDP framework, for particular values from the reward set $r \in \mathcal{R}$ and state set $s' \in \mathcal{S}$, there is a well defined discrete probability distribution, which depends only on the preceding state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ and represented by

$$p(s', r|s, a) = \Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a, \quad \forall s, s' \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}\}, \tag{31}$$

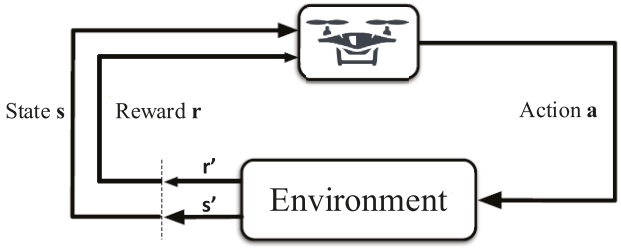

FIGURE 3. The iterative interaction process of learning UR-agent and environment in the RL.

The function p defines the dynamics of the finite MDP, which sums up to 1 according to the rule of total probability. Hence, we have

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1, \quad \forall s \in \mathcal{S}, \ a \in \mathcal{A}. \tag{32}$$

Given $p(s', r|s, a)$, one can obtain transition probabilities of the learning agent as a three argument function $\mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, given by

$$\begin{aligned} \pi(s'|s, a) &= \Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} \\ &= \sum_{r \in \mathcal{R}} \pi(s', r|s, a), \end{aligned} \tag{33}$$

which results in the expected reward function of the state-action-next-state triples, defined below as a three-argument function

$$\begin{aligned} r(s, a, s') &= \mathbb{E}\{R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'\} \\ &= \sum_{r \in \mathcal{R}} r \frac{\pi(s', r|s, a)}{\pi(s'|s, a)}, \end{aligned} \tag{34}$$

RL algorithms, which are employed to solve the above-mentioned finite MDP, are basically instructing the learning agent via estimating the *action-value function* or the so-called *Q-function*. Precisely, Q-function estimates the quality of action taken by the agent in a given state in terms of the expected discounted return Υ_t. This return captures not only the immediate reward but also a scaled version of the future rewards in the long run for all successive steps, which can be mathematically represented as

$$\Upsilon_t = \sum_{n=t+1}^{\mathcal{L}} \gamma^{n-t-1} R_n, \tag{35}$$

where R_{t+k} for $k = 1, 2, \cdots, \mathcal{L}$ indicate the future rewards after time step t, $\gamma \in [0, 1]$ denotes the discount rate, specifying to what degree of importance the future rewards should be taken into account, and $\mathcal{L}$ represents the final time step. Note that, since the expected future rewards depend on the particular action the agent will take in the future, therefore, this Q-function should be defined in regards of the agent's way of acting or the so-called *policy*. Indeed, policy is the core element of RL methods and this decision-making rule is merely a mapping from states to the probability of taking each possible action. Mathematically speaking, if the agent follows policy π at time step t, then it will take the action $a_t = a$ at

state $S_t = s$ according the probability $\pi(a|s)$, which is given by

$$\pi(a|s) = \Pr\{A_t = a|S_t = s\}, \quad \forall a \in \mathcal{A},\ s \in \mathcal{S} \tag{36}$$

In light of this, the value of taking action a under policy π and in state s is defined as

$$\begin{aligned} Q_\pi(s,a) &= \mathbb{E}_\pi\left\{\Upsilon_t \middle| S_t = s, A_t = a\right\} \\ &= \mathbb{E}_\pi\left\{\sum_{n=t+1}^{T} \gamma^{n-t-1} R_n \middle| S_t = s, A_t = a\right\}, \end{aligned} \tag{37}$$

where the Q-function $Q_\pi(s,a)$ indicates the expected return starting from the state s, taking the action a and following policy π thereafter, in which Υ_t is defined in (35). Likewise, the value function of a state s under a policy π is defined as the expected return when starting in state s and then following policy π afterwards, which can be given by

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi\{\Upsilon_t \big| S_t = s\} \\ &= \mathbb{E}_\pi\left\{\sum_{n=t+1}^{T} \gamma^{n-t-1} R_n \middle| S_t = s,\ \forall s \in \mathcal{S}\right\} \end{aligned} \tag{38}$$

These two functions can be related via

$$v_\pi(s) = \sum_a \pi(a|s) Q_\pi(s,a), \tag{39}$$

Solving an RL problem for the finite MDP is roughly equivalent to finding an optimal policy,[1] which can be expressed as

$$\pi^\star(a|s) = \begin{cases} 1, & \text{if } a = \arg\max\limits_{a \in \mathcal{A}} Q^\star(s,a) \\ 0, & o.w. \end{cases}$$

We note that the optimal policy also shares the same optimal Q-function, i.e., $Q^\star(s,a)$, defined precisely as

$$Q^\star(s,a) = \max_\pi Q_\pi(s,a), \quad \forall s \in \mathcal{S},\ \forall a \in \mathcal{A} \tag{40}$$

The above function provides the optimal expected long-term return as a value that is locally and immediately available for each state–action pair. Additionally, $Q^\star(s,a)$ is required to satisfy the *Bellman optimality equation*, given by

$$\begin{aligned} Q^\star(s,a) &= \mathbb{E}\{R_{t+1} + \gamma \max_{a'} Q^\star(S_{t+1}, a') | S_t = s, A_t = a\} \\ &= \sum_{s',r} p(s', r|s, a)\left[r + \gamma \max_{a'} Q^\star(s', a')\right]. \end{aligned} \tag{41}$$

It is worth pointing out that the Bellman optimality equation is non-linear and explicitly solving it is, in practice, too hard, since at the outset we need to accurately aware of the dynamics of the environment − corresponding to have p available, and secondly we need to have sufficient computational resources for solving the equations amongst other required assumptions. Alternatively, we can use some iterative methods to approximately solve it and hence estimate the optimal Q-function in an efficient time. In light of this, TD-based reinforcement algorithms are model-free methods that recursively approximate the Q-function. Furthermore, there are two main types of TD learning methods: On-policy and Off-policy. While the former attempts to evaluate and improve the policy that is used to make decisions, the latter evaluates or improves a policy known as behavior policy, which may or may not correlate with that used to generate the data, namely the *estimation policy*.

Now, we have the necessary tools to dive into our problem in terms of reformulating the UR's trajectory design problem such that we solve it via the finite MDP-based RL algorithms. Towards that end, we reformulate (30) as a model-free RL problem. Specifically, we divide the original problem into three sub-problems: user scheduling, trajectory optimization, and jointly power and ANI allocation sub-problems. In other words, a three-stage decision-making process is considered to cope with the original problem. First, the UR takes one of the possible actions to obtain its trajectory, then, selects one of the users as per protocol detailed in II-B, i.e., mainly the closest user to the UR, for conducting relaying service. Upon UR's changing position and scheduling the ground user, by optimizing the available resources $\prec P_s, P_u, \alpha_s, \alpha_u \succ$, the ISSR is improved, which in turn contributes to the reward received by the learning agent (i.e., UR) following the action taken in the previous stage. Therefore, to approximately reformulate the original problem to be solved efficiently, we need to precisely specify the RL models in terms of state set, action set, rewards, and the algorithm, all of which are detailed below.

B. STATE SET

Since we are interested in UR's trajectory design, therefore, we can consider each state representing the position of the UR in 3D space. However, seeing that UR's position can generally be modeled as a continuous function of time, i.e., $Q_u(t)$ where $t_0 \le t \le t_0 + T$, hence, this leads to having an infinite state set. Nonetheless, we hold attention to the restricted number of possible states as per the finite MDP framework. Therefore, the considered rectangular region, where the fixed-altitude UR aims to learn the optimal trajectory via RL is partitioned into N_w by N_l small tiles. Consequently, the region $[0, \mathcal{R}_w]$ by $[0, \mathcal{R}_l]$ in Fig. 1 is converted into a finite grid-world of $N_w \times N_l$ tiles which the x-y coordinates of the center of each tile represents one state. As a result, the state set $\mathcal{S}$ can be represented as

$$\mathcal{S} = \{S_n = (x_n, y_n) \big| n = 1, 2, 3, \cdots, N\}, \tag{42}$$

where $N = N_w \times N_l$, x_n and y_n are given respectively by

$$x_n = \frac{\mathcal{R}_l}{2N_l} + \frac{(n-1)\mathcal{R}_l}{N_l}, \quad \text{for } n = 1, 2, \cdots, N_l \tag{43}$$

$$y_n = \frac{\mathcal{R}_w}{2N_w} + \frac{(n-1)\mathcal{R}_w}{N_w}, \quad \text{for } n = 1, 2, \cdots, N_w \tag{44}$$

[1] A policy π_1 is defined to be better than or equal to a policy π_2 if its expected return is greater than or equal to that of π_2 for all states. In other words, $\pi_1 \ge \pi_2 \iff v_{\pi_1}(s) \ge v_{\pi_2}(s), \forall s \in \mathcal{S}$. We also note that there the optimal policy is not necessarily unique, however, there exists always at least one policy that outperforms all other policies in terms of the expected return.

It should be mentioned that here we also assume that each tile, representing the position in a x-y coordinate, might be occupied by a user, an AE, or an obstacle. Thus, the corresponding occupied states can be represented as sets $\mathcal{S}_d \subset \mathcal{S}$, $\mathcal{S}_e \subset \mathcal{S}$, and $\mathcal{S}_o \subset \mathcal{S}$, respectively, where it is assumed that $\mathcal{S}_d \cap \mathcal{S}_e \cap \mathcal{S}_o = \emptyset$. Further, according to the definitions given above, we can define the initial state s_{init} and the termination state s_{flag} of the considered MDP problem as

$$s_{init} = (x_1, y_1), \quad s_{flag} = (x_{N_l}, y_{N_w}), \tag{45}$$

Moreover, based on the constraint C8 in (30), we have $q_i = q_u(0) = s_{init}$ and $q_f = q_u(N_{sp}^{max}) = s_{flag}$, where N_{sp}^{max} is a positive integer corresponding to the mission completion time. Moreover, with a slight change in the notation, the UR's discrete position at time step t can be considered as $q_u(t) = s \in \mathcal{S}$.

C. ACTION SET

As illustrated in Fig. 1, the available action set for the UR is assumed to be

$$\mathcal{A} = \{\text{N, S, W, E, NE, SE, NW, SW}\}, \tag{46}$$

where N refers to flying one tile towards the north ($-y$ direction), S refers to flying one tile towards the south ($+y$ direction), E refers to flying forward for one tile in the direction of $+x$, W refers to flying backward for one tile in the direction of $-x$. Analogously, NE, SE, NW, SW indicate flying one tile with higher speed but equal time duration, when compared to other aforementioned directions, towards north-east, south-east, north-west, and south-west, respectively. Therefore, the cardinality of the action set is 8. Note that, in practice the UR is capable of selecting any direction, however, the optimal continuous trajectory can be approximately considered when the number of states in our problem goes to infinity.

It should be mentioned that here we adopt ε-*greedy* strategy for action selection in order to balance between exploration and exploitation of the environment. As such, action $a^\star$ is selected according to

$$a^\star = \begin{cases} \underset{a \in \mathcal{A}}{\text{argmax}}\, Q(s,a), & \text{if } \text{rand}(\cdot) \geq \varepsilon \\ \text{random action selection}, & o.w. \end{cases} \tag{47}$$

where $\text{rand}(\cdot) \in [0,1]$, and ε represents the probability of exploration, wherein the agent has the chance to improve its current knowledge about each action, which of course, enables the agent to make more informed decisions in the future. Further, $1-\varepsilon$ denotes the exploitation rate, which refers to choosing the greedy action to obtain the most reward by exploiting the UR-agent's current action-value estimates. In general, we want the UR-agent to start off the learning of environment with fairly randomized policy and later gradually move towards a deterministic one, which implies ε should be decaying. Note that in this work we consider ε as a decreasing function of episode number episode as

$$\varepsilon^{new} = \varepsilon^{old} \epsilon^{\left\lfloor \frac{\text{episode}}{k} \right\rfloor}, \tag{48}$$

where $\epsilon \in (0,1)$, k is a constant positive integer value such that makes ε parameter decay every kth episodes, and $\lfloor \cdot \rfloor$ represents floor bracket operator. This dynamic choice of epsilon parameter with proper constants results in having more chance of environment exploration at the beginning of learning process and mainly following the learned policy at the last episodes.

D. REWARD FORMULATION

In order to enable the UR to be successful in the quest for an optimal trajectory to maximize the ASSR during the flight mission, we devise the reward function in such a way that the constraints imposed by the environment are also satisfied in the RL. It is worth pointing out that when the UR takes an action a at time step t such that $t = 1, 2, \cdots, N_{sp}^{max}$, and then transits from current state s to next state s' receiving the reward r' at the successive time step, the UR-agent assigns a score for the taken action to indicate how important that action was in rendering the future reward. Therefore, the reward function of the UR-agent is defined as

$$r = \zeta_1 \mathcal{F}_1 \hat{R}_{sec} - \zeta_2 \mathcal{F}_2 + \zeta_3 \mathcal{F}_3, \tag{49}$$

where $\hat{R}_{sec}$ represents the improved ISSR as an immediate reward, $\mathcal{F}_1$ is the indication function for taking into account QoS requirements in terms of both the communication secrecy outage and user service fairness, $\mathcal{F}_2$ is also an indication function that encourages the UR to complete the mission at the final desired destination as soon as possible (decreasing N_{sp}^{max} which corresponds to minimising the mission time T) in order to improve the overall ASSR performance. Finally, $\mathcal{F}_3$ is a function which penalizes the UR-agent to avoid collision with an obstacle, to make it fly inside the restricted region, and discourage it from getting stuck in a loop which may result in a mission failure. It is worth pointing out that the reward function parameters $(\zeta_1, \zeta_2, \zeta_3)$ should be selected in such a way to balance between positive rewards (revenue) and negative rewards (cost). Now, we delve into mathematically explaining the functions making the instantaneous reward r in more detail.

1) DEFINITION OF FUNCTION $\hat{R}_{sec}$

Following user scheduling according to (15), for the sake of maximizing the ASSR in (30), it is important to optimize the ISSR via proper resources allocation, which in turn, improves the system sum secrecy rate and accordingly ASSR performance. We define the function $\hat{R}_{sec}$ for a given time step t as

$$\begin{aligned} \hat{R}_{sec} = \underset{P_s, P_u, \alpha_s, \alpha_u}{\text{maximize}}\ & R_{sec}(P_s, P_u, \alpha_s, \alpha_u) \\ s.t.\ & \widehat{\text{C1}}: \ 0 \leq P_s \leq P_s^{\max} \\ & \widehat{\text{C2}}: \ 0 \leq P_u \leq P_u^{\max} \\ & \widehat{\text{C3}}: \ 0 \leq \alpha_s, \alpha_u \leq 1 \end{aligned} \tag{50}$$

where $\widehat{C1}$ and $\widehat{C2}$ are the maximum transmission power constraints posed by hardware limitations and regulatory standards. $\widehat{C3}$ indicates the ANI power allocation constraints. The above optimization problem, having non-linear objective function with convex constraints, can be readily solved via known optimization toolbox such as *fmincon* in MATLAB.

2) DEFINITION OF FUNCTION $\mathcal{F}_1$

According to the minimum QoS requirement of the mission, we assume there exists a minimum required secrecy rate R_{sec}^{th} constraint to be satisfied, or the secure communication undergoes an outage. To that aim, we formulate this constraint as a penalty function to penalize the UR for taking those actions during the learning process that lead to any QoS failure. As a result, to avoid such circumstances, we define the function $\mathcal{F}_1$ as

$$\mathcal{F}_1 = \begin{cases} 1, & \hat{R}_{sec}(t) \geq R_{sec}^{th} \text{ and } \sum_{\tau=1}^{t} \hat{R}_{sec}^{(j^\star)}(\tau) \leq \bar{R}_{sec}^{max} \\ 0, & \text{o.w.} \end{cases} \tag{51}$$

where $\hat{R}_{sec}$ is given by (50), R_{sec}^{th} represents the minimum instantaneous secrecy rate requirement at each user, $\bar{R}_{sec}^{max}$ indicates the maximum sum secrecy rate threshold, below which the selected user's cumulative secrecy rates up to the given time step t should be, in order to ensure fairness in providing service amongst the users.

3) DEFINITION OF FUNCTION $\mathcal{F}_2$

Since the UR is required to complete the mission in the pre-specified final location denoted as a flag in Fig. 1, we need to have a termination state. However, the UR-agent may not, depending on the environment, complete the mission due to getting stuck in some states. To avoid this, we penalize the UR by the function $\mathcal{F}_2$, defined precisely by

$$\mathcal{F}_2 = \left(1 + \frac{t}{N_{sp}^{max}}\right) \frac{\|q_u(t) - q_f\|}{\|q_f - q_i\|}, \tag{52}$$

It is worth mentioning that penalty function $\mathcal{F}_2$ is designed as the multiplication of two terms: I) $\left(1 + \frac{t}{N_{sp}^{max}}\right)$ is a penalty corresponding to the number of time steps taken so far, and in general, we target at a reasonable mission completion time of the UR (i.e., as fast as it can with fewer steps) to reduce mission time duration and in some sense mechanical energy consumption, II) $\frac{\|q_u(t)-q_f\|}{\|q_f-q_i\|}$ indicates the normalized distance between the UR's current location and the desired final location to motivate the UR to find its way towards the termination state. We emphasize that penalty function $\mathcal{F}_2$ via multiplication of both terms ensures that the UR does not get stuck in some specific states, which may have higher secrecy rates, for an unreasonable period of time, and thus, avoids mission incompletion.

4) DEFINITION OF FUNCTION $\mathcal{F}_3$

Apart from the previous functions contributing the UR's reward, we need another function to impose the environmental and mission requirement constraints. To that aim, we define $\mathcal{F}_3$ as

$$\mathcal{F}_3 = \begin{cases} -f_p, & \text{if } q_u(t) \in \mathcal{S}_o, \\ -f_p, & \text{if } q_u(t) \notin \mathcal{S}, \\ -f_p, & \text{if } q_u(t) = q_u(t-\tau)\ \forall \tau \in \{1, 2, \cdots, t-1\}, \\ +f_r, & \text{if } q_u(t) = s_{flag}, \\ 0, & \text{o.w.} \end{cases} \tag{53}$$

where f_p and f_r are the absolute values of some immediate penalty and reward (both of them will be quantified in the simulations) subject to the conditions with which the UR-agent has encountered, respectively. The first penalty term in (53) ensures that the UR-agent flies at a safety distance of obstacles to avoid possible collisions, the second penalty is to motivate the UR not to go beyond the operating region of interest, the third penalty term is also to further discourage the UR-agent to avoid getting stuck in some specific states making an infinite loop, and the forth term is a reward for reaching the termination state and completing the mission.

E. TD-BASED MODEL-FREE RL ALGORITHMS

TD-based techniques refer to a class of model-free reinforcement learning algorithms which can learn from raw experience without knowing a model of the environment's dynamic, and also update estimates based in part on other learned estimates, i.e., learns by bootstrapping from the current estimate of the value function without waiting for an outcome. In this work, we consider *SARSA* as an On-policy TD learning algorithm and *Q-learning* as an Off-policy TD learning, as well as some generalized versions based on these two, i.e., Expected SARSA, Double Q-learning, and SARSA(λ), for the UR-agent's trajectory optimization and then we compare their performances in the numerical section. We note that the main difference between these algorithms lies in the Q-value update rule which are detailed below.

1) SARSA: ON-POLICY TD LEARNING-BASED ALGORITHM

SARSA is an On-policy reinforcement learning TD method with the Q-value update rule expressed by

$$Q_{\text{sa}}^{new}(s,a) = Q_{\text{sa}}^{old}(s,a) + \alpha \left[r + \gamma Q_{\text{sa}}^{old}(s',a') - Q_{\text{sa}}^{old}(s,a) \right], \tag{54}$$

where α and γ denote the learning rate and the discount factor, respectively. Note that (54) demonstrates how learning is conducted from one state-action pair to another and the Q-value is updated. This update rule is calculated after every transition from a non-terminal state. It has been proven that SARSA converges with unit probability to an optimal policy provided that all state–action pairs are visited an infinite number of times and the policy converges in the limit to the greedy policy [40]. SARSA-based intelligent trajectory design for the proposed flying FD-operated MIMO-UR system is, therefore, given in Algorithm 1.

Algorithm 1 SARSA-Based Trajectory Design for the Proposed FD-Operated MIMO-UR Scenario

Inputs:
learning parameters (α, γ),
reward function factors $(\zeta_1, \zeta_2, \zeta_3)$
termination parameters $(N_{sp}^{max}, N_{ep}^{max})$
state sets $\mathcal{S}, \mathcal{S}_d, \mathcal{S}_e, \mathcal{S}_0$, and action set $\mathcal{A}$
Outputs:
$\pi_{sa}^{\star} = \arg\max_{a\in\mathcal{A}} \ Q(s,a) \ \ \forall s \in \mathcal{S}$
Initialize:
$\varepsilon \leftarrow 1$; $Q = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
for episode = 1 to N_{ep}^{max} **do**
  $s \leftarrow s_{init}$; sp $\leftarrow 0$
  update ε using (48)
  choose a based on (47)
  while $s \neq s_{flag}$ and sp $< N_{sp}^{max}$ **do**
    take a, then observe s' and calculate r using (49)
    select a' using (47)
    $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$
    $s \leftarrow s'$; $a \leftarrow a'$; sp $\leftarrow$ sp $+ 1$
  end while
end for

Algorithm 2 Q-Learning Based Trajectory Design for the Proposed FD-Operated MIMO-UR Scenario

Inputs:
learning parameters (α, γ),
reward function factors $(\zeta_1, \zeta_2, \zeta_3)$
termination parameters $(N_{sp}^{max}, N_{ep}^{max})$
state sets $\mathcal{S}, \mathcal{S}_d, \mathcal{S}_e, \mathcal{S}_0$, and action set $\mathcal{A}$
Outputs:
$\pi_{ql}^{\star} = \arg\max_{a\in\mathcal{A}} \ Q(s,a) \ \ \forall s \in \mathcal{S}$
Initialize:
$\varepsilon \leftarrow 1$; $Q = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
for episode = 1 to N_{ep}^{max} **do**
  $s \leftarrow s_{init}$; sp $\leftarrow 0$
  update ε using (48)
  while $s \neq s_{flag}$ and sp $< N_{sp}^{max}$ **do**
    choose a based on (47),
    take a, then observe s' and calculate r using (49)
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a\in\mathcal{A}} Q(s',a) - Q(s,a)]$
    $s \leftarrow s'$; sp $\leftarrow$ sp $+ 1$
  end while
end for

2) Q-LEARNING: OFF-POLICY TD LEARNING-BASED ALGORITHM

Q-learning is an Off-policy reinforcement TD learning method with the update rule defined by

$$Q_{\mathrm{ql}}^{new}(s,a) = Q_{\mathrm{ql}}^{old}(s,a) + \alpha\left[r + \gamma \max_{a\in\mathcal{A}} Q_{\mathrm{ql}}^{old}(s',a) - Q_{\mathrm{ql}}^{old}(s,a)\right], \tag{55}$$

where the learned action-value function directly approximates Q-value, regardless of what policy being followed by the UR-agent. Q-learning based UR trajectory design for the proposed scenario is given in Algorithm 2.

3) EXPECTED SARSA

Since the use of next action a' introduces additional variance into the update rule for On-policy SARSA method, it may slow the convergence. For this reason, a modified version of SARSA, i.e., Expected SARSA, has been proposed in [40] and systematically investigated in [48]. Expected SARSA algorithm, instead of using the next action a', indeed, employs an expectation (weighted sum) over all available actions in state s' considering the probability of each action under the current policy, with the update rule as

$$Q_{\mathrm{es}}^{new} = Q_{\mathrm{es}}^{old}(s,a) + \alpha\left[r + \gamma V_{s'} - Q_{\mathrm{es}}^{old}(s,a)\right], \tag{56}$$

where the state-value function can be defined as

$$V_{s'} \triangleq \sum_{a\in\mathcal{A}} \pi(a|s') Q_{\mathrm{es}}^{old}(s',a) \tag{57}$$

in which $\pi(s'|a)$ is given by

$$\pi(a|s') = \begin{cases} \dfrac{\varepsilon}{|\mathcal{A}|}, & \text{if } a \text{ is non-greedy} \\ \dfrac{1-\varepsilon}{|\mathcal{A}_g|} + \dfrac{\varepsilon}{|\mathcal{A}|}, & \text{if } a \text{ is greedy} \end{cases} \tag{58}$$

wherein $|\mathcal{A}_g|$ shows the number of greedy actions, and ε is given in (47). This may offer substantial advantages over SARSA learning by reducing the variance and accordingly speeding up the convergence of the learning algorithm. On the other hand, Expected SARSA is quite similar to Q-learning for the case when the estimation policy is greedy, and therefore, it can be viewed as an On-policy version of Q-learning. We note that Expected SARSA requires more calculations than SARSA but lacks the high variance due to random selection of the next action a'. As a result, it moves deterministically towards the same direction that SARSA moves in expectation, and this leads to relatively better performance for the Expected SARSA than the normal SARSA with the same amount of experience as shall be seen in the numerical section. Expected SARSA based UR trajectory design for the proposed scenario is given in Algorithm 3.

4) DOUBLE Q-LEARNING

Using max operation in the single-estimator Q-learning algorithm may result in poor performance due to a large overestimation issue, particularly in some stochastic environments. To remedy this issue, Double Q-learning has been proposed in [49], which employs two Q-functions such that they both get randomly updated based on the value from the other Q-function for the next state s', and this partly relieves the situation in terms of overestimation issue, however, this approach may result in underestimation of the maximum

Algorithm 3 Expected SARSA Based Trajectory Design for the Proposed FD-Operated MIMO-UR Scenario

Inputs:
learning parameters (α, γ),
reward function factors $(\zeta_1, \zeta_2, \zeta_3)$
termination parameters $(N_{sp}^{max}, N_{ep}^{max})$
state sets $\mathcal{S}, \mathcal{S}_d, \mathcal{S}_e, \mathcal{S}_0$, and action set $\mathcal{A}$
Outputs:
$\pi_{es}^{\star} = \arg\max_{a\in\mathcal{A}} \ Q(s,a) \ \ \forall s \in \mathcal{S}$
Initialize:
$\varepsilon \leftarrow 1$; $Q = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
for episode $= 1$ to N_{ep}^{max} **do**
  $s \leftarrow s_{init}$; sp $\leftarrow 0$
  update ε using (48)
  while $s \neq s_{flag}$ and sp $< N_{sp}^{max}$ **do**
    choose a based on (47)
    take a, then observe s' and calculate r using (49)
    calculate $V_{s'}$ according to (57)
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma\ V_{s'} - Q(s,a)]$
    $s \leftarrow s'$; sp $\leftarrow$ sp $+ 1$
  end while
end for

expected value. Double Q-learning is regarded as an unbiased estimate of action-value function since the update of two Q-functions occurs in the same problem but learning is accomplished from dissimilar experience sample sets. Here, we proposed a Double Q-learning based RL algorithm for UR-agent's trajectory optimization in Algorithm 4, wherein for action selection, the sum of both the state-value functions for each action is considered to capture the effects of both Q-functions. Although this requires higher computational storage, and hence, Double Q-learning is generally less data-efficient than normal Q-learning, it significantly speeds up the convergence, as we will see later on in the numerical section.

5) SARSA(λ)

So far, all the abovementioned TD learning algorithms, i.e., Q-learning, SARSA, Expected SARSA, and Double Q-learning, only consider one step at most for updating the Q-table values corresponding to the operation state. However, before getting a given state, every step taken resulting to that might be important to consider with different level of degrees. SARSA(λ) algorithm [40] is a generalized multi-step version of SARSA which not only updates the Q-value of the latest step, but also, efficiently rewards all the related steps using the so-called *eligibility trace*. Eligibility trace is indeed a matrix such as $E_{|\mathcal{S}|\times|\mathcal{A}|}$, which is initialized with zeros prior to each episode, and saves each step in the path experience whose state-action value gets incremented by one. However, after each step all the elements of the eligibility trace will be decayed proportionally to the the bootstrapping factor λ. This ensures that all the action values from the beginning of

Algorithm 4 Double Q-Learning Based Trajectory Design for the Proposed FD-Operated MIMO-UR Scenario

Inputs:
learning parameters (α, γ),
reward function factors $(\zeta_1, \zeta_2, \zeta_3)$
termination parameters $(N_{sp}^{max}, N_{ep}^{max})$
state sets $\mathcal{S}, \mathcal{S}_d, \mathcal{S}_e, \mathcal{S}_0$, and action set $\mathcal{A}$
Outputs:
$\pi_{dq}^{\star} = \arg\max_{a\in\mathcal{A}} \ Q(s,a) \ \ \forall s \in \mathcal{S}$
Initialize:
$\varepsilon \leftarrow 1$; $Q_1 = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$; $Q_2 = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
for episode $= 1$ to N_{ep}^{max} **do**
  $s \leftarrow s_{init}$; sp $\leftarrow 0$
  update ε using (48)
  while $s \neq s_{flag}$ and sp $< N_{sp}^{max}$ **do**
    calculate $Q \leftarrow Q_1 + Q_2$
    choose a using (47) based on π derived from Q
    take a, then observe s' and calculate r using (49)
    if rand$(\cdot) \geq 0.5$ **then**
      let $a^{\star} = \arg\max_{a\in\mathcal{A}} Q_1(s', a)$
      $Q_1(s,a) \leftarrow Q_1(s,a) + \alpha[r + \gamma Q_2(s', a^{\star}) - Q_1(s,a)]$
    else
      let $b^{\star} = \arg\max_{a\in\mathcal{A}} Q_2(s', a)$
      $Q_2(s,a) \leftarrow Q_2(s,a) + \alpha[r + \gamma Q_1(s', b^{\star}) - Q_2(s,a)]$
    end if
    $s \leftarrow s'$; sp $\leftarrow$ sp $+ 1$
  end while
end for

the episode up to the last step taken are updated with different degrees following the recency fading. The SARSA(λ) based trajectory design for the proposed UR-based relaying scenario is given in Algorithm 5. We will see how well this algorithm might perform compared to the others in the numerical section.

F. COMPLEXITY DISCUSSIONS

In this paper, we use the grid-world for the exploration of the UR to find its optimal trajectory. Therefore, the state space topology has linear upper action bound, i.e., the number of possible actions in each state capped with $|\mathcal{A}| = 8$ at most. Further, we have finite set of states with cardinality of $|\mathcal{S}| = N_l N_w \triangleq N$. In Algorithms 1 and 2, the UR-agent learns by visiting all the states, updating the corresponding Q-values during each episode[2] Assuming that all the actions are known to the UR, and the state space is fully observable in that the UR is capable of determining it's current state, therefore, the learning agent can reach the goal state and terminate after at most $\mathcal{O}(|\mathcal{S}| \sum_{s\in\mathcal{S}} |\mathcal{A}(s)|)$ steps. Further, since the considered grid-world has the special property of 1-step

[2]In this work, we define an episode as all the states that come in between the initial state and when the UR-agent reaches the termination state, or the number of steps taken goes beyond the maximum time limit N_{ep}^{max} during the course of the agent-environment learning process.

Algorithm 5 SARSA(λ) Based Trajectory Design for the Proposed FD-Operated MIMO-UR Scenario

Inputs:
learning parameters (α, γ),
reward function factors $(\zeta_1, \zeta_2, \zeta_3)$
termination parameters $(N_{sp}^{max}, N_{ep}^{max})$
state sets $\mathcal{S}, \mathcal{S}_d, \mathcal{S}_e, \mathcal{S}_0$, and action set $\mathcal{A}$
Outputs:
$\pi_{sl}^{\star} = \arg\max_{a\in\mathcal{A}} \ Q(s,a) \ \ \forall s \in \mathcal{S}$
Initialize:
$\varepsilon \leftarrow 1$; $Q = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
for ep $= 1$ to N_{ep}^{max} **do**
  reset the eligibility trace $E = \mathbf{0}_{|\mathcal{S}|\times|\mathcal{A}|}$
  $s \leftarrow s_{init}$; sp $\leftarrow 0$
  update ε using (48)
  while $s \neq s_{flag}$ and sp $< N_{sp}^{max}$ **do**
    choose a based on (47)
    take a, then observe s' and calculate r using (49)
    $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    $E(s, a) \leftarrow E(s, a) + 1$
    for $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$ **do**
      $Q(s, a) \leftarrow Q(s, a) + \alpha\delta E(s, a)$
      $E(s, a) \leftarrow \gamma\lambda E(s, a)$
    end for
    $s \leftarrow s'$; $a \leftarrow a'$; sp $\leftarrow$ sp $+ 1$
  end while
end for

invertible due to having no duplicate actions [40], the worst-case complexity for both the proposed algorithms becomes $\mathcal{O}(8N^2)$. This implies that the worst-case complexity of proposed learning-based algorithms, though perform undirected exploration, are in polynomial order with respect to the number of states N. The only major difference between the two basic algorithms is that the On-policy SARSA learns action values relative to the policy it follows and hence may impose high sample complexity, while Off-policy Q-learning does it relative to the greedy policy in each state visiting, and hence, might be comparably slower. Further, the generalized version of the above algorithms which have been mainly proposed in the literature to speed up the convergence at the cost of higher complexity such as Expected SARSA or SARSA(λ), or higher storage resources but with the same computational complexity for Double Q-learning.

V. NUMERICAL RESULTS

In this section, we demonstrate the performance of the proposed algorithms to find the optimal trajectory for the considered MIMO-beamforming based secure UAV-assisted flying relay system. Simulation settings, mainly adopted from the literature, all are given in Table 2, unless *otherwise stated*. It is worth noting that we use both Python 3.7 and Matlab R2020a to implement the algorithms and conduct the simulations. Further, all the learning experiments have been conducted

TABLE 2. Simulation parameters.

Notation	Simulation value	description
(L, M)	$(10, 5)$	number of users and AEs
$(\mathcal{R}_w, \mathcal{R}_l)$	$(100m, 200m)$	size of rectangular operation region
H_u	$80m$	UR's operating altitude
Q_i	$[x_1, y_1, H_u]^T$	UR's start location
Q_f	$[x_{N_l}, y_{N_w}, H_u]^T$	UR's final location
q_s	$[x_1, y_1, 0]^T$	location of the BS
(P_s^{max}, P_u^{max})	(10mW, 5mW)	BS and UR's maximum transmit powers
$P_{j\in\mathcal{E}} = P_e$	1mW	AEs' jamming power
(ρ_u, ρ_e)	$(10^{-9}, 10^{-9})$	residual interference factors
(η_L, η_N)	$(2, 5)$	LoS and non-LoS path-loss exponents
(δ_s, δ_u)	$(1, 1)$	antennas constant spacing, in wavelength
(N_s, N_u)	$(32, 4)$	number of transmitting antennas
$\sigma_u^2 = \sigma_d^2 = \sigma_e^2$	-40dBm	background noise power
(N_w, N_l)	$(10, 16)$	grid-world dimensions (number of states)
$(\omega_1, \omega_2, \omega_3, \omega_4)$	$(11.95, 0.14, 3.1623, 0.0256)$	environmental parameters
$(R_{sec}^{th}, R_{sec}^{max})$	(2 bps/Hz, 20 bps/Hz)	secrecy QoS requirement
(f_p, f_r)	$(R_{sec}^{th}, 2R_{sec}^{th})$	absolute values of penalty and reward in $\mathcal{F}_3$
α	0.05	learning rate (step size)
γ	0.95	discounting factor
$(\zeta_1, \zeta_2, \zeta_3)$	$(1, 0.1, 1)$	reward parameters
(ϵ, k)	$(0.9, 50)$	ε-greedy constants
$(N_{ep}^{max}, N_{sp}^{max})$	$(3000, 60)$	algorithms' termination parameters

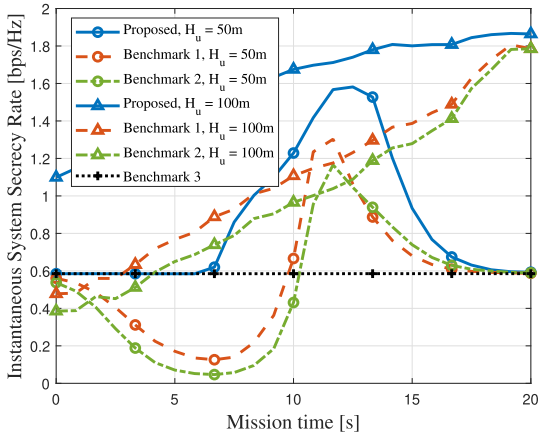

FIGURE 4. Simple illustration of the achievable secrecy rate performance of the proposed FD-operated UR-based scenario and impacts of resource allocation.

using Python 3.7 on i5-8265U CPU @ 1.6 GHz with 16 GB of RAM system.

First, we supply Fig. 4 to demonstrate how the proposed FD-operated MIMO-UR with ANI scenario, denoted as *Proposed*, performs in terms of the ISSR and the impact of resource allocation according to (50). In this figure, the ISSR performance of the ANI-based UR system with fixed resource allocation, i.e., $P_s(t) = P_s^{max}, P_u(t) = P_u^{max}, \alpha_s(t) = 0.5, \alpha_u(t) = 0.5 \ \forall t$, is represented by *Benchmark 1*. Further, *Benchmark 2* illustrates the ISSR performance of the UR system without ANI operation (equivalent to $\alpha_s(t) = 1, \alpha_u(t) = 1 \ \forall t$) and with fixed transmit powers $P_s(t) = P_s^{max}, P_u(t) = P_u^{max} \ \forall t$. The ISSR performance of the direct transmission protocol from the BS with ANI beamforming to destination and optimized communication resources, represented by *Benchmark 3*, also taken into account for comparison. Note that the curves in Fig. 4 are plotted for different altitudes of UAV to demonstrate the effect of this key system parameter. In Fig. 4, we consider a simple scenario with fixed-line trajectory of the UAV, wherein there are one

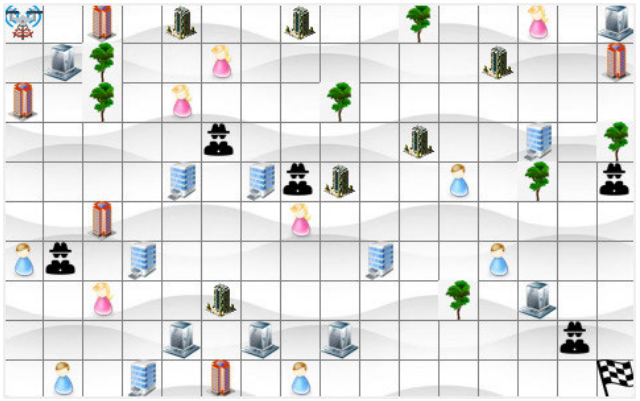

FIGURE 5. Considered environment for the learning purpose of the UR-agent.

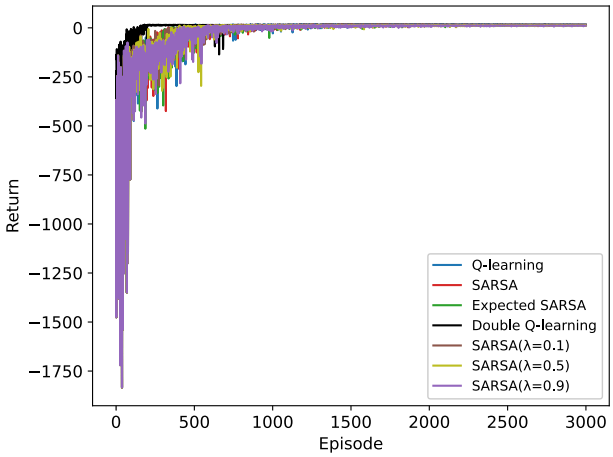

FIGURE 6. Expected discounted return of rewards vs. Episode.

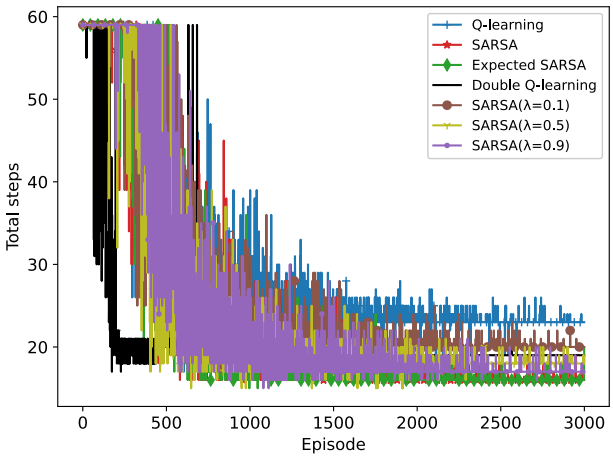

FIGURE 7. Total steps vs. Episode.

BS located at $q_s = [0, 0]^T$, one user at $q_d = [q, 0]^T$, and one AE at $q_e = [\frac{3q}{4}, 0]^T$ with $q = 100$m. Plus, the UAV flies with constant speed at a fixed altitude H_u from just above the BS towards just above the user. Therefore, at mission time

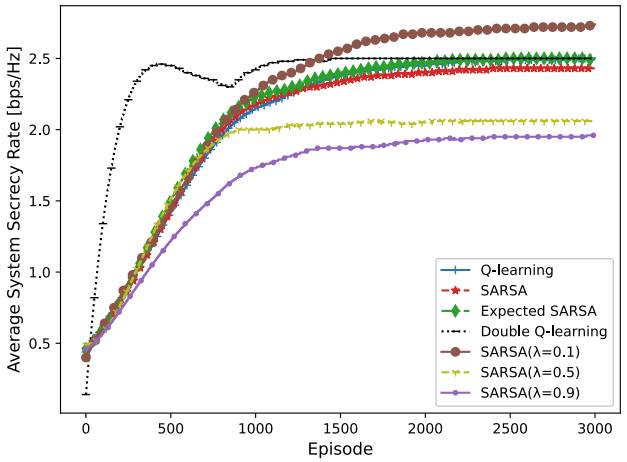

FIGURE 8. Average system secrecy rate (smoothened) vs. Episode.

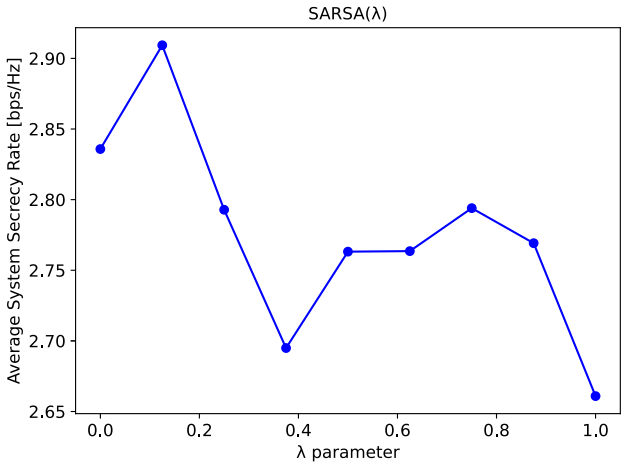

FIGURE 9. Average system secrecy rate of SARSA(λ) against different λ values. Reward factors are set as $(\zeta_1, \zeta_2, \zeta_3) = (1, 0.1, 1)$.

$t = 0$s, the UR is located at the same x-y coordinate of the BS but with altitude H_u, and at $t = 20$s, the UR reaches its final location assumed to have a similar projected x-y coordinate of the user's one. It is crystal clear from the curves that our proposed scheme well outperforms the other benchmarks. Further, we see that for the lower altitude of the UR, i.e., $H_u = 50$m, there is an optimal location for the UR to offer the best ISSR for all these UR-based scenarios, and this location is roughly closer to the destination user than the BS. Further, this figure also illustrates that when the low-altitude UR is too far from the destination user, Benchmarks 1 and 2 bring comparably deteriorated ISSR performance than the traditional direct transmission in Benchmark 3 without exploiting a UR. The justification behind this can be explained as for the lower altitude of UR and without proper resource allocation, the secrecy capacity of the relaying link may be dramatically impacted due to overall larger attenuation. However, for the reasonably higher UR's altitude, i.e., $H_u = 100$m, the advantages of having likely LoS channel due to higher altitude plays a significant role so much so that it may result in a

(a) SARSA based path-planning using Algorithm 1.

(b) Q-learning based path-planning using Algorithm 2.

(c) Expected SARSA based path-planning using Algorithm 3.

(d) Double Q-learning based path-planning using Algorithm 4.

(e) SARSA($\lambda = 0.1$) based path-planning using Algorithm 5.

(f) SARSA($\lambda = 0.5$) based path-planning using Algorithm 5.

FIGURE 10. Model-free Reinforcement learning based path-planning for the proposed FD-operated UR-based scenario obtained via optimal policy of different On-policy and Off-policy TD algorithms such as Q-learning, SARSA, Expected SARSA, Double Q-learning, and SARSA(λ) with λ = 0.1, 0.5.

higher ISSR performance than the properly resource allocated scheme at the lower altitude. We also observe that for the higher altitude of $H_u = 100$m, the ISSR performance of all the UR-assisted scenarios gets improved as the UR gets closer to the destination user, so the better channel condition due to UR's placement has a stronger effect than the proper communication resource allocation. However, it is worth mentioning that there is an inherent trade-off between the LoS channel and the larger path loss attenuation due to the higher altitude of the UR, and this should be taken into account for system design. Note that in this work we consider a fixed altitude of the UR for the remaining simulations, albeit

this can be readily extended to a general case in our future work.

Now, we turn our attention to employ the model-free RL-based algorithms mentioned in the previous section to design the UR's path planning in order to maximize the ASSR performance. The considered environment is shown in Fig. 5, wherein obstacles such as tall trees and buildings denote the prohibited flying regions, and also the randomly located users (visualized by either boy or girl icons) and AEs are considered quasi-stationary such that their possible movements do not result in a change in their occupied states during the UR's flight duration. Having implemented the proposed TD-based RL algorithms detailed in the previous section, we obtain the following results.

Figs. 6 and 7 are provided to demonstrate the accumulated discounted rewards (return) and mission duration (total steps) performance of all the considered RL-based algorithms, i.e., Q-learning, SARSA, Expected SARSA, Double Q-learning, and SARSA(λ) with ε-greedy action selection strategy, and in regards to the episode number. As can be observed from Fig. 6, initially the return fluctuates dramatically as the UR explores the environment with a higher probability and takes actions randomly. This results in mainly getting negative rewards due to, for example, collisions with obstacles, revisiting already visited states in a given episode, or flying outside the limited region. Nevertheless, after sufficient time when the UR-agent is trained well via the received feedback from the environment, which further leads to having a decent knowledge of the topology of the environment, it tends to utilize from its experience, and consequently, the fluctuation in the accumulated discounted reward function gets negligible, indicating that the Q-function updating is settled. This also implies that the maximum achievable ASSR is attained. Further, we observe from Fig. 7 that the number of steps that corresponds to the mission time is decreased as the learning agent interacts with the environment in each episode. Further, it is evident that the UR intends to complete the mission as fast as possible, while accomplishing the required objective. Indeed, there is a trade-off between the mission time and the achievable ASSR of the system. The more the mission time, the higher the sum system secrecy rate could be due to better positioning and data relaying which improves the objective function of the ASSR, but the larger the denominator of the objective function, which of course, degrades the ASSR. Therefore, the UR intends to complete the mission as fast as it can, while being smart. That's why the path plannings obtained via the algorithms generally look like the way that the UR is instructed to find the best, but not necessarily the shortest, path from the initial location to the final pre-specified location. While (Expected) SARSA could achieve the least total steps of 16 according to Fig. 7, Q-learning has performed the worst with regards to total steps taken and obtained the highest mission duration of 23 time steps, according to their final routes derived from the optimal policies.

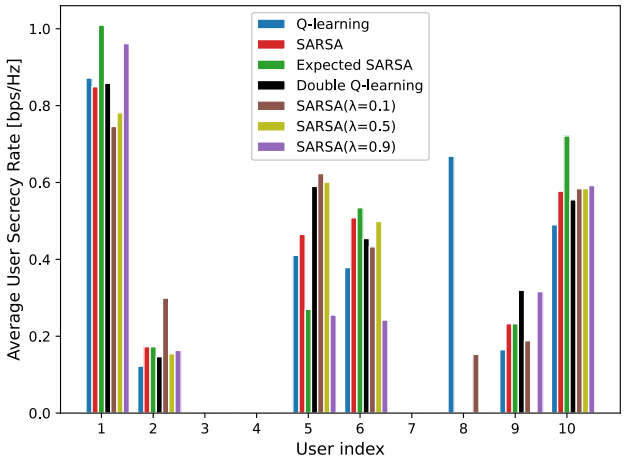

FIGURE 11. Average user secrecy rate versus user index.

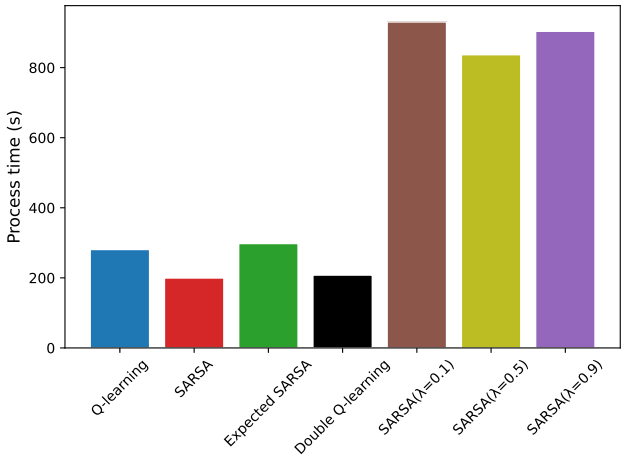

FIGURE 12. Processing time for different model-free TD-based RL algorithms considered for ASSR optimization. Reward factors are set as $(\zeta_1, \zeta_2, \zeta_3) = (1, 0.1, 1)$.

Fig. 8 is plotted to demonstrate the convergence of the proposed RL-based algorithms in terms of the ASSR. Note that in the figure, the ASSR values are smoothed via a Savitzky–Golay filter [50] to better illustrate the overall trends. It is evident that all the algorithms get converged and demonstrate quite identical increasing trend to that of discounted cumulative reward given in Fig. 6. We also note that SARSA($\lambda = 0.1$) achieves the highest value, i.e., 2.73 bps/Hz, amongst all the algorithms with the equal number of training episodes. The convergence speed of Double Q-learning is comparatively faster than the others, particularly, normal Q-learning, despite the fact that they achieve quite identical optimal ASSR values. We also observe that higher values of λ degrades the performance of the algorithm in the considered environment. Due to the latter result, we also empirically investigate the effect of λ factor in the performance of SARSA(λ) algorithm in Fig. 9 and observe that $\lambda = 0.1$ brings the best ASSR, so it can be considered the best choice of bootstrapping factor in SARSA(λ) based algorithm for the given environment.

Fig. 10 depicts the trajectories learned by the UR-agent using the proposed TD-based algorithms. Note that final

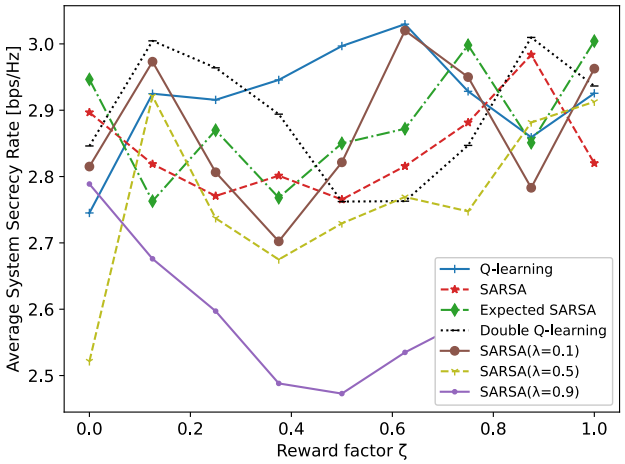

FIGURE 13. ASSR of different model-free RL-based learning algorithms versus Reward factor are set as $(\zeta_1, \zeta_2, \zeta_3) = (1, \zeta, 1)$.

routes are derived from the optimal policy of all the aforementioned algorithms. We note that all the algorithms find slightly different trajectories compared to each other and this difference is on the grounds that they employ different updating approaches, and owing to the fact that SARSA based algorithms are more conservative than Q-learning ones when it comes to the actions explored.

Fig. 11 depicts the average user secrecy rate against user index for different proposed RL-based algorithms. We see that all the proposed algorithms satisfy the minimum and sum secrecy rate requirements of the system. Despite that some users are not scheduled according to Fig. 11, the considered secrecy rate requirements of the scheduled users are well satisfied, indicating the effectiveness of the proposed trajectory design and resource allocations algorithms. Fig. 12 is supplied to show the processing time taken for the algorithms to complete using our system. As can be observed from the figure, all one-step algorithms, particularly, SARSA perform better than the multi-step ones in terms of the lowest process time. The impacts of reward function parameter ζ is explored in Fig. 13. we note, intuitively, that the proper choice of ζ results in a balance between system sum secrecy rate and the total mission completion time, which further leads to the best ASSR performance. Further investigation regarding tuning the learning parameters and reward factors is required, which we leave as interesting future work.

VI. CONCLUSION

In this paper, we considered an FD-operated UR-assisted secure communication system to serve multiple ground users in the presence of randomly located AEs. We proposed a secure relaying scheme, wherein both the BS and UR adopt MIMO-enabled ANI-based beamforming to combat AEs. Our problem of interest was to maximize the ASSR of the considered scenario. To achieve this objective, we invoked some model-free TD-based RL algorithms, i.e., Q-learning, SARSA, Expected SARSA, Double Q-learning, and SARSA(λ) for trajectory optimization. The proposed algorithms were subject to some QoS requirements in terms of minimum instantaneous secrecy rate, user's maximum sum secrecy rate, and also without the need for system identification. Simulation results revealed that all of the proposed algorithms were capable of finding an optimal trajectory of the UR while improving the ASSR, avoiding collision with environmental obstacles, and completing the mission as fast as possible. As a future research direction, one can extend this work to investigate more practical scenarios when the state and action spaces are continuous and/or have very large dimensions suffering from the curse of dimensionality issue. Then, the promising DRL techniques such as DQN may be explored to perform the functional optimization to efficiently tackle the computationally intensive learning process of tabular methods investigated in this work.

REFERENCES

[1] Y. Zhao, J. Zhao, W. Zhai, S. Sun, D. Niyato, and K.-Y. Lam, "A survey of 6G wireless communications: Emerging technologies," 2020, *arXiv:2004.08549*. [Online]. Available: http://arxiv.org/abs/2004.08549

[2] S. Ali *et al.*, "6G white paper on machine learning in wireless communication networks," 2020, *arXiv:2004.13875*. [Online]. Available: http://arxiv.org/abs/2004.13875

[3] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[4] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[5] O. Bouachir, M. Aloqaily, I. A. Ridhawi, O. Alfandi, and H. B. Salameh, "UAV-assisted vehicular communication for densely crowded environments," in *Proc. NOMS IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2020, pp. 1–4.

[6] M. Aloqaily, O. Bouachir, A. Boukerche, and I. A. Ridhawi, "Design guidelines for blockchain-assisted 5G-UAV networks," 2020, *arXiv:2007.15286*. [Online]. Available: https://arxiv.org/abs/2007.15286

[7] X. Liu, Z. Li, N. Zhao, W. Meng, G. Gui, Y. Chen, and F. Adachi, "Transceiver design and multihop D2D for UAV IoT coverage in disasters," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1803–1815, Apr. 2019.

[8] G. Zhang, H. Yan, Y. Zeng, M. Cui, and Y. Liu, "Trajectory optimization and power allocation for multi-hop UAV relaying communications," *IEEE Access*, vol. 6, pp. 48566–48576, 2018.

[9] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in UAV-enabled wireless-powered mobile-edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1927–1941, Sep. 2018.

[10] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-Relaying-Assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, May 2019.

[11] M. M. Azari, F. Rosas, and S. Pollin, "Cellular connectivity for UAVs: Network modeling, performance analysis, and design guidelines," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3366–3381, Jul. 2019.

[12] M. M. Azari, A. H. Arani, and F. Rosas, "Mobile cellular-connected UAVs: Reinforcement learning for sky limits," 2020, *arXiv:2009.09815*. [Online]. Available: http://arxiv.org/abs/2009.09815

[13] O. Bouachir, M. Aloqaily, F. Garcia, N. Larrieu, and T. Gayraud, "Testbed of QoS ad-hoc network designed for cooperative multi-drone tasks," in *Proc. 17th ACM Int. Symp. Mobility Manage. Wireless Access MobiWac*, 2019, pp. 89–95.

[14] A. Yener and S. Ulukus, "Wireless physical-layer security: Lessons learned from information theory," *Proc. IEEE*, vol. 103, no. 10, pp. 1814–1825, Oct. 2015.

[15] H.-M. Wang, Y. Zhang, X. Zhang, and Z. Li, "Secrecy and covert communications against UAV surveillance via multi-hop networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 389–401, Jan. 2020.

[16] H.-M. Wang, X. Zhang, and J.-C. Jiang, "UAV-involved wireless physical-layer secure communications: Overview and research directions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 32–39, Oct. 2019.

[17] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," *Proc. IEEE*, vol. 104, no. 9, pp. 1727–1765, Sep. 2016.
[18] X. Sun, D. W. K. Ng, Z. Ding, Y. Xu, and Z. Zhong, "Physical layer security in UAV systems: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 40–47, Oct. 2019.
[19] N. Wang, P. Wang, A. Alipour-Fanid, L. Jiao, and K. Zeng, "Physical-layer security of 5G wireless networks for IoT: Challenges and opportunities," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8169–8181, Oct. 2019.
[20] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.
[21] J. M. Hamamreh, H. M. Furqan, and H. Arslan, "Classifications and applications of physical layer security techniques for confidentiality: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1773–1828, 2nd Quart., 2019.
[22] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb. 2019.
[23] M. T. Mamaghani and Y. Hong, "On the performance of low-altitude UAV-enabled secure AF relaying with cooperative jamming and SWIPT," *IEEE Access*, vol. 7, pp. 153060–153073, 2019.
[24] M. T. Mamaghani and Y. Hong, "Improving PHY-security of UAV-enabled transmission with wireless energy harvesting: Robust trajectory design and communications resource allocation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8586–8600, Aug. 2020.
[25] M. T. Mamaghani and Y. Hong, "Joint trajectory and power allocation design for secure artificial noise aided UAV communications," 2020, *arXiv:2011.10245*. [Online]. Available: http://arxiv.org/abs/2011.10245
[26] Y. Cai, Z. Wei, R. Li, D. W. K. Ng, and J. Yuan, "Joint trajectory and resource allocation design for energy-efficient secure UAV communication systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4536–4553, Jul. 2020.
[27] L. Xiao, Y. Xu, D. Yang, and Y. Zeng, "Secrecy energy efficiency maximization for UAV-enabled mobile relaying," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 180–193, Mar. 2020.
[28] Q. Wang, Z. Chen, H. Li, and S. Li, "Joint power and trajectory design for physical-layer secrecy in the UAV-aided mobile relaying system," *IEEE Access*, vol. 6, pp. 62849–62855, 2018.
[29] Q. Yuan, Y. Hu, C. Wang, and Y. Li, "Joint 3D beamforming and trajectory design for UAV-enabled mobile relaying system," *IEEE Access*, vol. 7, pp. 26488–26496, 2019.
[30] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.
[31] S. Yin, Y. Zhao, L. Li, and F. R. Yu, "UAV-assisted cooperative communications with power-splitting information and power transfer," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 1044–1057, Dec. 2019.
[32] Q. Song, F.-C. Zheng, Y. Zeng, and J. Zhang, "Joint beamforming and power allocation for UAV-enabled full-duplex relay," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1657–1671, Feb. 2019.
[33] M. Hua, Y. Wang, Z. Zhang, C. Li, Y. Huang, and L. Yang, "Outage probability minimization for low-altitude UAV-enabled full-duplex mobile relaying systems," *China Commun.*, vol. 15, no. 5, pp. 9–24, May 2018.
[34] T. Nuradha, K. T. Hemachandra, T. Samarasinghe, and S. Atapattu, "Physical-layer security for untrusted UAV-assisted full-duplex wireless networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2019, pp. 1–6.
[35] M. T. Mamaghani, A. Mohammadi, P. L. Yeoh, and A. Kuhestani, "Secure two-way communication via a wireless powered untrusted relay and friendly jammer," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.
[36] M. Tatar Mamaghani, A. Kuhestani, and K.-K. Wong, "Secure two-way transmission via wireless-powered untrusted relay and external jammer," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8451–8465, Sep. 2018.
[37] M. T. Mamaghani and R. Abbas, "Security and reliability performance analysis for two-way wireless energy harvesting based untrusted relaying with cooperative jamming," *IET Commun.*, vol. 13, no. 4, pp. 449–459, Mar. 2019.
[38] C. Liu, J. Lee, and T. Q. S. Quek, "Safeguarding UAV communications against full-duplex active eavesdropper," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 2919–2931, Jun. 2019.
[39] D. Liu, C. Sun, C. Yang, and L. Hanzo, "Optimizing wireless systems using unsupervised and reinforced-unsupervised deep learning," *IEEE Netw.*, vol. 34, no. 4, pp. 270–277, Jul. 2020.
[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[41] B. Khamidehi and E. S. Sousa, "Reinforcement learning-based trajectory design for the aerial base stations," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–6.
[42] J. Cui, Z. Ding, Y. Deng, A. Nallanathan, and L. Hanzo, "Adaptive UAV-trajectory optimization under quality of service constraints: A model-free solution," *IEEE Access*, vol. 8, pp. 112253–112265, 2020, doi: 10.1109/ACCESS.2020.3001752.
[43] Y. Li, R. Zhang, J. Zhang, S. Gao, and L. Yang, "Cooperative jamming for secure UAV communications with partial eavesdropper information," *IEEE Access*, vol. 7, pp. 94593–94603, 2019.
[44] C. Liu, N. Yang, R. Malaney, and J. Yuan, "Artificial-noise-aided transmission in multi-antenna relay wiretap channels with spatially random eavesdroppers," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7444–7456, Nov. 2016.
[45] B. C. Nguyen, X. N. Tran, D. T. Tran, and L. T. Dung, "Full-duplex amplify-and-forward relay system with direct link: Performance analysis and optimization," *Phys. Commun.*, vol. 37, Dec. 2019, Art. no. 100888.
[46] E. E. B. Olivo, D. P. M. Osorio, H. Alves, J. C. S. S. Filho, and M. Latva-Aho, "Cognitive full-duplex decode-and-forward relaying networks with usable direct link and transmit-power constraints," *IEEE Access*, vol. 6, pp. 24983–24995, 2018.
[47] P. K. Gopala, L. Lai, and H. El Gamal, "On the secrecy capacity of fading channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4687–4698, Oct. 2008.
[48] H. van Seijen, H. van Hasselt, S. Whiteson, and M. Wiering, "A theoretical and empirical analysis of expected sarsa," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn.*, Mar. 2009, pp. 177–184.
[49] H. V. Hasselt, "Double Q-learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 2613–2621.
[50] R. Schafer, "What is a savitzky-golay filter? [lecture notes]," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, Jul. 2011.

MILAD TATAR MAMAGHANI (Graduate Student Member, IEEE) was born in Tabriz, Iran, in May 1994. He received the dual B.Sc. degrees in electrical engineering fields - telecommunications and control systems - from the Amirkabir University of Technology, Tehran, Iran, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Systems Engineering, Monash University, Melbourne, VIC, Australia.

He is the author of several papers published in prestigious journals/conferences, and has served as a volunteer reviewer for various reputable publication venues. His research interests include beyond 5G wireless communications and networking, physical-layer security, UAV communications, optimization, and machine learning. He is a member of the IEEE Communications Society and the IEEE Signal Processing Society.

YI HONG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering and telecommunications from The University of New South Wales (UNSW), Sydney, NSW, Australia. She is currently a Senior Lecturer with the Department of Electrical and Computer Systems Engineering, Monash University, Melbourne, VIC, Australia. Her research interests include communication theory, coding, and information theory with applications to telecommunication engineering. She was a Technical Program Committee Member for many IEEE leading conferences. She received the NICTA-ACoRN Earlier Career Researcher Award at the Australian Communication Theory Workshop, Adelaide, SA, Australia, 2007. She was the General Co-Chair of IEEE Information Theory Workshop 2014, Hobart, the Technical Program Committee Chair of Australian Communications Theory Workshop 2011, Melbourne, and the Publicity Chair of the IEEE Information Theory Workshop 2009, Sicily. She served on the Australian Research Council College of Experts from 2018 to 2020. She was an Associate Editor of IEEE Wireless Communication Letters and *Transactions on Emerging Telecommunications Technologies (ETT)*.