RIS-Aided Ground-Aerial NOMA Communications: A Distributionally Robust DRL Approach

Jingjing Zhao, *Member, IEEE*, Lanchenhui Yu, *Student Member, IEEE*, Kaiquan Cai, *Member, IEEE*, Yanbo Zhu, *Member, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*— A reconfigurable intelligent surface (RIS) aided air-to-ground uplink non-orthogonal transmission framework is investigated for next generation multiple access. Occupying the same spectrum resource, unmanned aerial vehicle (UAV) users and ground users (GUs) are connected to terrestrial cellular networks via the uplink non-orthogonal multiple access (NOMA) protocol. As the flight safety is important for employing UAVs in civil airspace, the collision avoidance mechanism has to be considered during the flight. Therefore, a joint optimization problem of the UAV trajectory design, RIS configuration, and uploading power control is formulated for maximizing the network sum rate, while ensuring the UAV's fight safety and satisfying the minimum data rate requirements of both the UAV and GU. The resultant problem is a sequential decision making one across multiple coherent time slots. Besides, the unknown locations of obstacles bring uncertainties into the decision making process. To tackle this challenging problem, a sample-efficient deep reinforcement learning (DRL) algorithm is proposed to optimize the UAV trajectory, RIS configuration, and power control simultaneously. Moreover, considering the ambiguous uncertainties in the environment, a distributionally robust DRL algorithm is further proposed to provide the worst-case performance guarantee. Numerical results demonstrate that the two proposed DRL algorithms outperform the conventional ones in terms of learning efficiency and robustness. It is also shown that the network sum rate is significantly improved by the proposed RIS-NOMA scheme compared to the conventional RIS-orthogonal multiple access (OMA) scheme and the case where no RIS is deployed.**

***Index Terms*— Air-to-ground communications, next generation multiple access, non-orthogonal multiple access, reconfigurable intelligent surface, distributionally robust deep reinforcement learning.**

Manuscript received August 18, 2021; revised November 14, 2021; accepted December 17, 2021. Date of publication January 14, 2022; date of current version March 17, 2022. This work was supported in part by the Funds of the National Natural Science Foundation of China under Grant 61822102 and Grant U2033215, in part by the U.S. National Science Foundation under Grant CNS-2128368 and Grant CNS-2107216, in part by Toyota, and in part by Amazon. *(Corresponding author: Kaiquan Cai.)*

Jingjing Zhao is with the Research Institute for Frontier Science, Beihang University, Beijing 100191, China, and also with the National Key Laboratory of CNS/ATM, Beijing 100191, China (e-mail: jingjingzhao@buaa.edu.cn).

Lanchenhui Yu and Kaiquan Cai are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China, and also with the National Key Laboratory of CNS/ATM, Beijing 100191, China (e-mail: yulanchenhui@buaa.edu.cn; caikq@buaa.edu.cn).

Yanbo Zhu is with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China, and also with Aviation Data Communication Corporation, Beijing 100191, China (e-mail: zyb@adcc.com.cn).

Zhu Han is with the Electrical and Computer Engineering Department, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: hanzhu22@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3143230.

Digital Object Identifier 10.1109/JSAC.2022.3143230

I. Introduction

In THE past several years, the use of unmanned aerial vehicles (UAVs) as flying communication platforms to boost the capacity and coverage of current wireless networks has attracted fast-growing interests [1]–[3]. In contrast to terrestrial wireless communications, UAV-aided networks possess many appealing advantages including high mobility, flexible deployment, low cost, and line-of-sight (LoS) dominated air-to-ground (A2G) links [4], [5]. Therefore, UAVs are expected to bring in promising gains to numerous use cases in next generation wireless networks. Particularly, UAVs acting as the aerial users and uploading data to the ground network is a vital application for use cases such as search and rescue missions, traffic monitoring, and remote location sensing [6]. A cost-effective approach to achieve high-quality A2G communications is to utilize already existing and accessible technologies like the ground cellular network, which brings in the concept of *cellular-connected UAV communications* [7]. Cellular-connected UAVs are anticipated to realize significant performance enhancement over the existing A2G communications based on unlicensed bands, in terms of reliability, throughput, and coverage [8].

Despite the evident merits of cellular-connected UAVs, one of the critical issues that has to be resolved is the limited spectrum resources available in cellular networks. Meanwhile, to leverage the spectrum resource more efficiently, power-domain non-orthogonal multiple access (NOMA)[1] has been envisioned to be a promising technique for its potential to enhance spectrum efficiency and massive connectivity by allowing simultaneous transmission of multiple users in the same resource block [9], [10]. More specifically, the fundamental concept of NOMA is to facilitate the access of multiple users in a new dimension-power domain, by employing superposition coding (SC) and successive interference cancellation (SIC) at the transmitter and receiver, respectively [11], [12].

In addition to the limited spectrum resources, another challenging issue for the efficient facilitation of A2G communications is the unstable A2G data link, which is caused by the existence of potential obstacles during flight especially in low-altitude dense urban airspace. As a remedy, reconfigurable intelligent surfaces (RISs) have been recently proposed to enable a promising new paradigm to achieve smart and reconfigurable wireless propagation environment [13]. RIS is

[1]In the rest of this paper, we use "NOMA" to refer to "power-domain NOMA" for simplicity.

a thin surface inlaid with numerous sub-wavelength elements, each of which is able to induce a controllable amplitude and phase-shift change to the incident signal independently via simple programmable PIN diodes [14]. By deploying RISs in wireless system and intelligently configuring their reflection coefficients, the communication quality can be improved by reconfigured wireless channels at an extremely low cost of power and hardware [15].

A. State-of-the-Art

1) UAV Communications With NOMA: To reap the benefits of NOMA in terms of spectrum efficiency, integration of NOMA into UAV-based wireless networks has attracted some research contributions recently [2], [16]–[20]. In [2], the authors proposed a novel framework for UAV networks with massive access capability supported by NOMA, where the joint UAV trajectory design and power control problem was comprehensively studied. The authors of [16] studied the max-min rate optimization problem under total power, total bandwidth, UAV altitude, and antenna beamwidth constraints in a downlink NOMA UAV network. The joint UAV placement design, admission control, and power control optimization problem was studied in [17] for the NOMA-based UAV downlink system to maximize the number of connected users with satisfied quality-of-service (QoS) requirements. The authors of [18] jointly optimized the resource allocation, the NOMA decoding order, and the deployment location of the UAVs to maximize the system sum rate. The authors of [19] studied the UAV-supported cluster-based NOMA system, where a synergetic scheme for UAV trajectory design and subslot allocation was proposed to maximize the uplink average sum rate. In [20], considering the cellular-connected UAV, the authors aimed to minimize the UAV mission completion time by jointly optimizing the UAV trajectory as well as the UAV and ground based station association order.

2) RIS-Aided UAV Communications: The potential benefits of RISs motivate researchers to investigate the RIS-aided UAV communications [21]–[27]. In [21], the authors investigated the average rate optimization problem by jointly optimizing the UAV trajectory and the RIS phase shifts. In [22], an RIS-enhanced multi-UAV NOMA network was considered, where the three-dimensional placement of UAV, the reflection-coefficient matrix of the RIS, and the NOMA decoding orders among users were jointly optimized for maximizing the network sum rate. The UAV trajectory, RIS configuration, and power allocation were jointly designed in [23] for the minimization of energy consumption. The authors of [24] exploited both the significant beamforming gain brought by the RIS and the high mobility of UAV for improving system sum rate. In [25], the joint optimization of the UAV trajectory, RIS configuration, terahertz sub-bands allocation, and power control was investigated for the maximization of the minimum average sum rate. The authors of [26] addressed the coverage and link performance problems of the aerial-terrestrial communication system and designed an adaptive RIS-assisted transmission protocol. Moreover, in [27], the authors investigated the scenario where the UAV and RIS delivered short ultra-reliable and low-latency instruction packets between Internet-of-Things devices on the ground, and studied the joint beamforming and UAV deployment problem.

B. Motivation and Contributions

Although the aforementioned works have studied the benefits of applying NOMA and RIS in UAV communications, whether the NOMA-RIS scheme is still able to provide performance gain in cellular-connected UAV uplink communications remains to be further discussed in the open literature. Moreover, the previous works mainly ignore the flight safety constraint on the UAV trajectory design, which should be imposed in practical UAV-based communication systems. The main challenges for solving the above issues lie in the following three aspects: *First*, the introduction of NOMA protocol brings in more complicated interference environment and channel condition-based decoding order design [28]. This leads to a highly coupled UAV movement, RIS configuration and uplink power control problem, rendering the optimal scheme hard to obtain. *Second*, as the reflection coefficients are shared by both the UAV and ground user (GU), the optimal shaping of reflected signals is not just to get aligned with the direct signals. Thus, the RIS configuration becomes more complicated due to the existence of co-channel interference. *Third*, due to the unknown locations of obstacles, resilient UAV trajectory, RIS configuration, and power control decisions should be made under an uncertain environment. Besides, as the uncertainty can not be accurately modelled, how to utilize efficient mathematical methods to improve the robustness of decision making process in face of ambiguous uncertainties is another challenge.

To address the above issues, in this paper, we study the RIS aided air-to-ground uplink NOMA cellular network where the direct links between the UAV/GU to the ground base station (GBS) suffer from deep shadowing. To be more specific, the UAV and GU simultaneously upload data to a GBS via the NOMA protocol with the assistance of RIS to provide concatenated virtual LoS links. The proposed framework introduces the new paradigm of flexibility on efficient spectrum sharing between the UAV and GU by taking advantage of the UAV's high mobility, reconfigurable wireless environment as well as power-domain multi-user access. Our main contributions can be summarized as follows:

- We propose a novel RIS aided air-to-ground communication framework, where the NOMA protocol is employed for facilitating flexible multiple access scheme. Given the proposed framework, we formulate the sum rate maximization problem by jointly optimizing the UAV trajectory, RIS configuration, and uplink power control, while guaranteeing the flight safety and the minimum data rate requirements of both the UAV and GU.
- We propose a distributionally robust DRL algorithm based on the soft actor-critic framework to jointly optimize the UAV trajectory, RIS configuration, and power control under uncertainties brought by the unknown locations of obstacles. The ambiguity set is constructed endogenously to capture the uncertainty by integrating

the partial distribution information, thereby guaranteeing the worst-case performance under uncertain environment.
- We prove that the proposed DRL algorithm improves the learning efficiency and robustness compared to conventional DRL algorithms both theoretically and numerically. Simulation results reveal that the designed RIS aided air-to-ground uplink non-orthogonal transmission framework can achieve distinct sum rate improvement over the traditional RIS-OMA case and the case where no RIS is deployed.

C. Organization and Notation

The rest of this paper is organized as follows. In Section II, we introduce the model of the RIS-aided air-to-ground uplink NOMA communication system, and formulate the sum rate maximization problem. In Section III, a sample-efficient DRL algorithm is proposed for solving the formulated problem. In Section IV, a novel distributionally robust DRL algorithm is further proposed to enhance the robustness of the DRL algorithm with respect to (w.r.t.) uncertainties. Numerical results are presented in Section V to verify the effectiveness of the proposed algorithms compared to other benchmarks. Finally, conclusions are drawn in Section VI.

Notation: Scalars, vectors and matrices are denoted by italic letters, bold-face lower-case, and bold-face upper-case, respectively. $\mathbb{C}^{N\times 1}$ denotes the set of $N\times 1$ complex-valued vectors. For a complex-valued vector $\mathbf{a}$, $\|\mathbf{a}\|$ denotes its Euclidean norm, $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with the elements of vector $\mathbf{a}$ on the main diagonal, and $\mathbf{a}^H$ denotes its conjugate transpose. Δ_X denotes the set of probability distributions over a finite set X. $\langle \mathbf{a}, \mathbf{b}\rangle$ denotes the Frobenius inner product of vectors $\mathbf{a}$ and $\mathbf{b}$.

II. System Model and Problem Formulation

A. System Model

Consider an A2G communication system where a fixed-wing[2] UAV acts as the aerial user (AU) to upload data to the GBS, as shown in Fig. 1. Due to limited spectrum resources, the UAV is enabled to reuse the cellular spectrum by employing NOMA with GUs. To obtain the fundamental insight on the system performance, we consider the simple model where only one GU is served by the GBS.[3] Assume that the UAV, GU and GBS are all equipped with a single omni-directional antenna.[4] Due to the complicated and dynamic wireless environment including potential obstacles, the direct links between the UAV/GU and the GBS may be blocked.

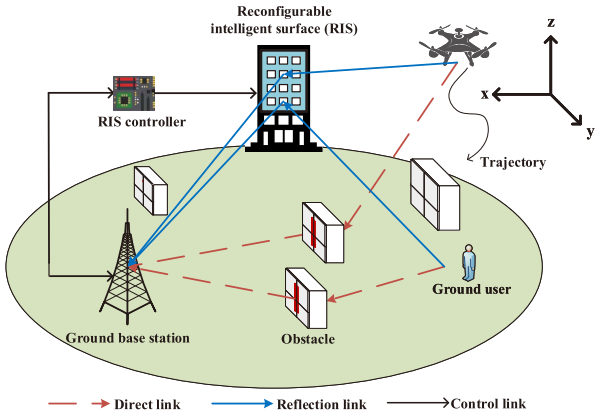


Fig. 1. System model.

Therefore, an RIS having K reflecting elements is deployed upon highrises to provide high-quality reflection links.

To proceed further, we assume that the considered network operates on a discrete-time basis where the total flight duration T is partitioned into M equal non-overlapped time slots. The communication parameters are assumed to remain constant during each time slot m, i.e., $m \in \mathcal{M} = \{1, \ldots, M\}$, with duration δ_m. Without loss of generality, a 3D Cartesian coordinate system is considered. The locations of the GBS, the GU, and the center of the RIS[5] are fixed at (x_b, y_b, z_b), $(x_{gu}, y_{gu}, 0)$, and (x_s, y_s, z_s), respectively. Assume that the UAV flies at constant height of z_u with constant speed V. The trajectory of the UAV is denoted as $\mathbf{q}[m] = (x[m], y[m], z_u), m \in \mathcal{M}$. Since the UAV-RIS-GU cascaded links suffer from substantial path loss, a large number of RIS elements are required for achieving favorable reflected communications. However, the massive number of RIS elements result in excessive reflection coefficients design complexity [31]. To solve this problem, as in [32], [33], the K RIS elements are partitioned into N sub-surfaces, denoted by the set $\mathcal{N} = \{1, \ldots, N\}$, each consisting of $\overline{K} = K/N$ (assumed to be an integer) adjacent elements that share the same reflection coefficients for reducing the implementation complexity. In this work, we consider the narrowband transmission, where the RIS reflection coefficients are assumed to be approximately constant over the entire signal bandwidth. Specifically, denote the reflection coefficients of the n-th sub-surface at the m-th time slot by $\theta_n[m] = \beta_n[m]e^{j\phi_n[m]}$, where $\phi_n[m] \in [0, 2\pi)$ and $\beta_n[m] \in [0, 1]$ represent the phase shift and amplitude reflection coefficient, respectively. Then, the diagonal reflection coefficient matrix can be denoted by $\mathbf{\Theta}[m] = \mathrm{diag}\left(\boldsymbol{\theta}[m] \otimes \mathbf{1}_{\overline{K}\times 1}\right) \in \mathbb{C}^{K\times K}$, where $\boldsymbol{\theta}[m] = [\theta_1[m], \ldots, \theta_n[m], \ldots, \theta_N[m]]^T$. To maximize the signal power reflected by the RIS and reduce hardware cost, we set $\beta_n[m] = 1, \forall n \in \mathcal{N}, m \in \mathcal{M}$, and consider the practical discrete phase-shift values [31], i.e., $\phi_n[m] \in \{0, \Delta\phi, \ldots, (L-1)\Delta\phi\}, \forall n \in \mathcal{N}, m \in \mathcal{M}$, where

[2]In this paper, we consider the fixed-wing UAV with the advantage of a considerable long flight duration (i.e., up to several hours) for accomplishing given tasks, while for the rotary-wing UAV, the corresponding flight duration is quite limited (i.e., 20-30 minutes) [1].

[3]The considered scenario can be extended to the multi-GUs and multi-UAVs case by deploying hybrid NOMA and OMA scheme [29]. Specifically, different UAV-GU pairs can occupy orthogonal sub-carriers, where the efficient user clustering problem needs to be solved. This is out of the scope of this treatise, which will be left for our future work.

[4]The proposed algorithm is also applicable to the case with directional beamforming at the GBS/GU/UAV by considering their specific antenna radiation patterns.

[5]In practice, the location of the RIS can be either optimized or selected according to the geographical environment [30].

$\Delta\phi = 2\pi/L$ and L represents the number of discrete phase-shift levels.

Due to the limited spectrum resources in cellular networks, uplink NOMA communication is considered to enable spectrum sharing between the UAV and GU. The received signal at the GBS consists of four parts: the UAV-GBS direct link, UAV-RIS-GBS reflection link, GU-GBS direct link and GU-RIS-GBS reflection link. Let $h_{x,b} \in \mathbb{C}$, $\mathbf{h}_{s,b}^H \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{x,s} \in \mathbb{C}^{N\times 1}$ represent the channels from the GU/UAV to the GBS, that from the RIS to the GBS, and that from the GU/UAV to the RIS, respectively, where $x \in \{gu, u\}$. The UAV-GBS and GU-GBS links are modelled as Rayleigh fading channels due to the blocked LoS links and potential extensive scattering. The UAV-RIS, GU-RIS and RIS-GBS links are modelled as Rician fading channels due to the existence of LoS components.[6] Let $h_{gu}[m] = h_{gu,b}[m] + \mathbf{h}_{s,b}^H[m]\mathbf{\Theta}[m]\mathbf{h}_{gu,s}[m]$ and $h_u[m] = h_{u,b}[m] + \mathbf{h}_{s,b}^H[m]\mathbf{\Theta}[m]\mathbf{h}_{u,s}[m]$ denote the effective channels from the GU and UAV to the GBS, respectively. It is assumed that the Doppler effect caused by the UAV's high mobility can be compensated at the receiver [36]. Therefore, the received signal at the GBS at the m-th time slot can be represented by

$$y_b[m] = \underbrace{h_{gu}[m]\sqrt{p_{gu}[m]}x_{gu}[m]}_{\text{GU's signal}} + \underbrace{h_u[m]\sqrt{p_u[m]}x_u[m]}_{\text{UAV's signal}} + \underbrace{n_b[m]}_{\text{noise signal}}, \quad \forall m \in \mathcal{M}, \tag{1}$$

where $p_{gu}[m]$ and $p_u[m]$ denote the transmit powers of the GU and UAV, respectively, $x_{gu}[m]$ and $x_u[m]$ are the transmitted signals of the GU and the UAV, respectively, and $n_b[m] \sim \mathcal{CN}(0, \sigma_b^2)$ is the additive white Gaussian noise (AWGN).

For uplink NOMA, the signals of the users having better channel conditions are usually detected first and then subtracted from the received signal, while other signals can be detected suffering from less interference. In the proposed model, the effective channels for the UAV and GU may vary w.r.t. the UAV trajectory, $\mathbf{q}[m]$, and the RIS reflection-coefficient matrix, $\mathbf{\Theta}[m]$. Hence, the uplink NOMA detection order in this paper can not be previously determined based on the effective channels. However, in order to impose no negative effect on the GU due to the spectrum sharing, we enable the GBS to first detect the UAV's signal by treating the GU's signal as noise. Then, the GBS detects the GU's signal by subtracting the remodulated UAV's signal from the received composite signal via the successive interference cancellation (SIC). By doing so, despite the UAV uses the same spectral resource as the GU, the signal of the GU can still be detected in a interference-free manner as in the system without the UAV, thus guaranteeing the performance of the GU and improving the spectral efficiency. To ensure that SIC can be successfully carried out at the GBS, the following constraint has to be satisfied [37]:

$$|h_u[m]|^2 p_u[m] \geq |h_{gu}[m]|^2 p_{gu}[m], \quad \forall m \in \mathcal{M}. \tag{2}$$

Remark 1: In contrast to conventional uplink NOMA systems, where the users' channels are passively determined by the wireless environment, the proposed transmission framework is capable of beneficially modifying the users' channels via both the RIS phase-shift configuration and the UAV trajectory design. As a result, the SIC constraint (2) *can be satisfied in a more flexible manner as compared to the conventional system. This provides a promising air-to-ground non-orthogonal transmission strategy for NGMA.*

For the given UAV-GU decoding order, the received signal-to-interference-plus-noise ratio (SINR) of the UAV's signal at the GBS at time slot m is given by

$$\gamma_u[m] = \frac{|h_u[m]|^2 p_u[m]}{|h_{gu}[m]|^2 p_{gu}[m] + \sigma_b^2}, \quad \forall m \in \mathcal{M}. \tag{3}$$

After subtracting the UAV's signal from the received composite signal via SIC, the signal-to-noise ratio (SNR) of the GU at the GBS at time slot m can be expressed as follows

$$\gamma_{gu}[m] = \frac{|h_{gu}[m]|^2 p_{gu}[m]}{\sigma_b^2}, \quad \forall m \in \mathcal{M}. \tag{4}$$

Accordingly, the achievable communication rate of the UAV and GU at the m-th time slot are given by $R_u[m] = \log_2(1 + \gamma_u[m])$ and $R_{gu}[m] = \log_2(1 + \gamma_{gu}[m])$, respectively. Then, the sum rate at time slot m is given by

$$\begin{aligned} R[m] &= \log_2(1 + \gamma_u[m]) + \log_2(1 + \gamma_{gu}[m]) \\ &= \log_2\left(1 + \frac{|h_u[m]|^2 p_u[m] + |h_{gu}[m]|^2 p_{gu}[m]}{\sigma_b^2}\right), \\ &\quad \forall m \in \mathcal{M}. \end{aligned} \tag{5}$$

B. Collision Avoidance Model

Considering the dynamic urban environment where unexpected surrounding obstacles[7] in low altitude airspace may threaten the UAV's flight safety, we need to take into account of the collision avoidance mechanism to facilitate a safe flight operation. The UAV needs to detect surroundings in carrying out missions to perceive the surrounding information (i.e., locations of obstacles) via onboard sensors. Let R_s denote the sensing range of onboard sensors. Then, we can define the perceptual range of the UAV as a circular region centered at the UAV, and R_s is the radius in the 3D space. Due to the limited sensing scope, the collision avoidance mechanism needs to be carried out in an online manner, which leads to stringent requirements on the decision speed.

To facilitate the collision avoidance mechanism, We define a forbidden zone around the obstacle as shown in Fig. 2. The UAV is not allowed to fly over this zone to keep a safe distance to threats. Denote any obstacle may appear during the UAV's

[6]We assume that the perfect channel state information (CSI) is known to the GBS via communications with the RIS controller. Similar approaches as presented in [34], [35] can be deployed in our work for channel estimation with acceptable complexity and overhead, which is out of the scope of this treatise.

[7]Here, unexpected obstacles denote objects that are not characterized in the geography map, such as other UAVs and helikite. For simplicity, we assume the obstacles are static in this treatise. Collision avoidance with moving objects will be considered in our future work.

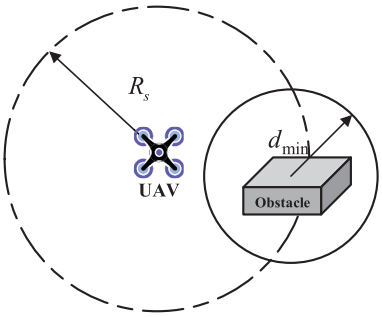


Fig. 2. UAV collision avoidance model.

flight by $o_i \in \mathcal{O}$, where $\mathcal{O}$ represents the set of all potential obstacles. To achieve the flight safety, the following constraint needs to be satisfied:

$$\|\mathbf{q}[m] - \mathbf{q}_{o_i}\| \geq d_{\min}, \quad \forall o_i \in \mathcal{O},\ m \in \mathcal{M}, \tag{6}$$

where $\mathbf{q}_{o_i}$ represents the location of obstacle o_i and $d_{\min}$ denotes the minimum separation distance. We make the assumption that $d_{\min} < R_s$.

C. Problem Formulation

Let $\mathbf{Q} \triangleq \{\mathbf{q}[m], m \in \mathcal{M}\}$, $\mathbf{\Theta} \triangleq \{\mathbf{\Theta}[m], m \in \mathcal{M}\}$, and $\mathbf{P} \triangleq \{p_u[m], p_{gu}[m], m \in \mathcal{M}\}$. Our objective in this work is to maximize the network sum rate over the total flight time T by jointly optimizing the trajectory of the UAV, the reflection-coefficient matrix of the RIS, and the power control of both the UAV and GU, subject to the constraints on the UAV flight safety and the instantaneous rate requirements of both the UAV and GU. The optimization problem is formulated as follows:

$$\max_{\mathbf{Q},\mathbf{\Theta},\mathbf{P}} \sum_{m=1}^{M} R[m], \tag{7a}$$

$$\text{s.t. } R_u[m] \geq R_u^{thr}, \quad R_{gu}[m] \geq R_{gu}^{thr}, \ \forall m \in \mathcal{M}, \tag{7b}$$

$$0 \leq p_u[m] \leq P_u^{\max}, \quad 0 \leq p_{gu}[m] \leq P_{gu}^{\max}, \quad \forall m \in \mathcal{M}, \tag{7c}$$

$$\phi_n[m] \in \{0, \Delta\phi, \ldots, (L-1)\Delta\phi\}, \quad \forall m \in \mathcal{M}, \ n \in \mathcal{N}, \tag{7d}$$

$$(2), (6), \tag{7e}$$

where (7b) represents the minimum data rate requirements for the UAV and the GU, (7c) is the maximum allowed transmit power of the UAV and the GU, and (7d) is the discrete phase-shift constraint of RIS elements. From the sum rate expression in (5), one can observe that the UAV and GU can just transmit in full power in each time slot for maximizing the sum rate. However, due to the instantaneous communication constraint (7b) and the SIC constraint (2), such a full power transmission is generally not the optimal solution. Therefore, the power control has to be jointly optimized with the UAV trajectory and the RIS reflection-cofficent matrix.

The main challenges of solving problem (7) lie in the following two aspects. First, the involved variables are highly coupled and the objective function is not concave w.r.t. the optimization variables. Second, the locations of unexpected obstacles are unknown, which causes substantial uncertainties for the UAV trajectory design. Such uncertainties make it hard for the conventional convex optimization-based approaches to solve this problem. To tackle the above challenges, we adopt RL algorithm, which is well known for its capability to address sequential decision making problems under uncertainties, to solve the joint problem (7). RL works with offline training and online deployment, and thus the online computational time can be distinctly saved. Specifically, due to the high complexity of the formulated problem, we opt to apply soft actor-critic (SAC) algorithm based on the maximum entropy RL framework to provide sample-efficient learning in Section III. Moreover, to offer robustness w.r.t. uncertainties during learning process, a novel distributionally robust DRL algorithm is proposed in Section IV.

III. Sample-Efficient DRL for UAV Trajectory Design, RIS Configuration and Power Control

In this section, we first formulate the joint UAV trajectory design, RIS configuration, and power control problem as a single-agent Markov Decision Process (MDP). Then, an off-policy actor-critic DRL algorithm with high sample efficiency is proposed to maximize the expected long-term reward of the considered network.

A. MDP Formulation

Problem (7) can be designed as a sequential decision making process on the time span, i.e., decision at a single time step is determined based on the current situation. In this sense, MDP that aims at finding the best policy, i.e., a mapping function from the current situation to the best decision, is suitable for solving this problem. We define a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ to model the MDP, where $\mathcal{S}$ is the set of environment states, $\mathcal{A}$ is the set of actions available to the agent, $\mathcal{P}$ is state transition probability matrix, $\mathcal{R}$ is a real-valued reward function for the agent taking an action based on present state, and γ is the discount factor. The agent takes action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at each time step m with the policy. A policy π is a distribution over actions given states, i.e., $\pi(a|s) = \mathbb{P}[A_m = a \mid S_m = s]$, $\pi(a|s) \in [0, 1]$. After taking the action a, the agent will move to the next state s' and receive the reward R. The agent's objective is to find the optimal policy π to maximize the state-value function. The state-value function is defined as the expected accumulated discounted reward, for which the Bellman expectation equation can be expressed as

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi \left[R_{m+1} + \gamma R_{m+2} + \gamma^2 R_{m+3} + \ldots \mid S_m = s \right] \\ &= \mathbb{E}_\pi \left[R_{m+1} + \gamma v_\pi(S_{m+1}) \mid S_m = s \right] \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s') \right), \end{aligned} \tag{8}$$

where the discount factor $\gamma \in [0, 1]$ indicates the present value of future rewards. γ close to 0 leads to "myopic" evaluation, while γ close to 1 leads to "far-sighted" evaluation. Furthermore, $\mathcal{R}_s^a$ is the reward function with $\mathcal{R}_s^a = \mathbb{E}[R_{m+1} \mid S_m = s, A_m = a]$, and $\mathcal{P}_{ss'}^a$

is the state transition probability matrix with $\mathcal{P}^a_{ss'} = \mathbb{P}\left[S_{m+1} = s' \mid S_m = s, A_m = a\right]$. In the formulated MDP, we consider the central controller as the agent to explore the unknown environment. The state, action and reward are defined in the following.

1) State: The environment state at time step m, i.e., S_m, includes three parts: i) the location of the UAV, i.e., $Q[m]$; ii) the distance from the UAV to the center of obstacles, i.e., $\mathcal{D}[m] = \{d_{u,o_i}[m], \forall o_i \in \mathcal{O}\}$; and iii) the sum rate of the UAV and GU from time step 1 to $m-1$, i.e., $R_{\text{sum}}[m-1] = \sum_{m'=1}^{m-1}(R_u[m'] + R_{gu}[m'])$.

$$S_m = \{Q[m], \mathcal{D}[m], R_{\text{sum}}[m-1]\}. \tag{9}$$

2) Action: The action space of the formulated MDP includes the UAV's maneuver direction, the phase shift of each RIS sub-surface, as well as the power control for both the UAV and GU. Considering the discrete phase-shift value settings of RIS elements, the action space is a hybrid of discrete and continuous spaces, which makes the proposed MDP problem non-trivial to solve. To tackle this challenge, we need to discretize the UAV's maneuver direction and the transmit power of both the UAV and GU. Specifically, the discrete action space contains three parts: i) the maneuver direction of UAV with $(-1,0),(0,1),(1,0),(0,-1)$ representing left, forward, right, and backward, respectively; ii) the phase shift of each RIS sub-surface, i.e., $\phi_n[m] \in \{0, \Delta\phi, \ldots, (L-1)\Delta\phi\}, \forall n \in \mathcal{N}$; and iii) the power control for the UAV and GU, i.e., $p_u[m] \in \{\tilde{p}_1^u, \ldots, \tilde{p}_X^u\}, p_{gu}[m] \in \{\tilde{p}_1^{gu}, \ldots, \tilde{p}_X^{gu}\}$, where X is the number of power control levels.

3) Reward: As shown in (7), the objective of the joint UAV trajectory design, RIS configuration and power control problem is to maximize the sum rate over time span T with given constraints. The reward that guides the learning should be consistent with the objective. In response to the sum rate maximization objective, we simply include the instantaneous sum rate of UAV and GU, i.e., $C[m] = R_u[m] + R_{gu}[m]$, in the reward at each time step. In response to constraints (7b)-(7e), we set a penalty if any of these constraints are not satisfied and terminate the episode. As such, we define the reward as

$$R_t = \begin{cases} -W, & \text{if } S_m = NS, \\ C[m], & \text{otherwise}, \end{cases} \tag{10}$$

where NS denotes the negative state when any of the constraints in (7b)-(7e) is unsatisfied. W is a positive constant which is set large enough to avoid the dissatisfaction of any of these constraints.

In the proposed MDP model, due to the uncertain locations of obstacles, the state transition probability matrix $\mathcal{P}^a_{ss'}$ is unknown to the agent. Therefore, approaches like dynamic programming (DP), which are based on a known MDP, is not suitable for solving our problem. Nevertheless, RL is promising since it enables the agent to control its action without the prior knowledge on the environment. In the proposed model, the locations of obstacles in the wireless environment are randomly generated by the simulator.

B. Sample-Efficient DRL Algorithm Design

It is known that the widespread adoption of model-free DRL frameworks (e.g., deep Q-learning, DDPG, etc.) in practice has remained slow primarily due to the poor sample efficiency and brittle convergence as stated in [38]: "a dominant concern in RL." To overcome this issue, a novel off-policy actor-critic DRL algorithm, called soft actor-critic (SAC), which is based on the maximum entropy framework, was firstly proposed by Haarnoja *et al.* [39] to realize sample-efficient training. Compared to prior state-of-the-art RL algorithms, SAC has the following advantages: 1) the policy is encouraged to explore more widely during the training process; 2) the policy can capture multiple modes of near-optimal trajectories; 3) the learning speed is improved for complicated tasks. However, one challenge of directly applying the SAC algorithm proposed in [39] into our work is that our work is based on discrete action settings, while [39] focused on continuous action space. Therefore, in order to reap the benefits of the SAC algorithm, some relevant modifications should be taken into consideration for constructing the discrete SAC algorithm. In the remaining context of this section, we will discuss in details about the main principles of the discrete SAC algorithm for solving our formulated problem.

The objective of the conventional RL framework is to maximize the long-term return starting from the initial state. Let τ_π denote the state-action trajectory distribution following the policy π, then the objective is denoted by

$$\max_{\pi} \sum_{m=1}^{M} \mathbb{E}_{(S_m, A_m)\sim\tau_\pi} \gamma^{m-1} \mathcal{R}_{S_m}^{A_m}. \tag{11}$$

In the maximum entropy framework, an entropy term is included in the objective to favor exploration with the loss of upcoming avenues. Specifically, the objective is expressed as follows:

$$\max_{\pi} F(\pi), \tag{12}$$

where

$$\begin{aligned} &F(\pi) \\ &= \sum_{m=1}^{M} \mathbb{E}_{(S_m, A_m)\sim\tau_\pi} \left[\gamma^{m-1} \mathcal{R}_{S_m}^{A_m} + \alpha \mathcal{H}\left(\pi\left(A_m | S_m\right)\right)\right] \\ &= \sum_{m=1}^{M} \mathbb{E}_{(S_m, A_m)\sim\tau_\pi} \left[\gamma^{m-1} \mathcal{R}_{S_m}^{A_m} - \alpha \log\left(\pi\left(A_m | S_m\right)\right)\right]. \end{aligned} \tag{13}$$

The new objective function (13) takes into account of the term $\alpha\mathcal{H}\left(\pi\left(\cdot|S_m\right)\right)$, where $\mathcal{H}\left(\pi\left(\cdot|S_m\right)\right) = -\mathbb{E}_{(S_m, A_m)\sim\tau_\pi} \log\left(\pi\left(A_m|S_m\right)\right)$ is the entropy of the policy distribution, which denotes the stochasticity of policy π, and the temperature parameter α denotes the weight of the entropy. Note that (13) is the same as (11) when α is set to 0. The optimal setting of temperature α is closely related to different tasks as well as the reward magnitude during training. In order to generate a flexible tuning of the entropy weight, a transformation of the objective function in (13) by treating the average

entropy as a constraint can be made as follows [40]:

$$\max_{\pi} \sum_{m=1}^{M} \mathbb{E}_{(S_m,A_m)\sim\tau_\pi}\left[\gamma^{m-1}\mathcal{R}_{S_m}^{A_m}\right], \tag{14a}$$

$$\text{s.t. } \mathbb{E}_{(S_m,A_m)\sim\tau_\pi}\left[-\log\left(\pi(A_m|S_m)\right)\right] \geq \mathcal{H}_{min}, \quad \forall m, \tag{14b}$$

where $\mathcal{H}_{min}$ is the minimum entropy constraint at each time step. By applying the recursive expression of $\mathbb{E}_{(S_m,A_m)\sim\tau_\pi}\left[\gamma^{m-1}\mathcal{R}_{S_m}^{A_m}\right]$ and the strong duality property, the optimal dual variable α_m^* at each time step is given by

$$\alpha_m^* = \arg\min_{\alpha_m} \mathbb{E}_{A_m\sim\pi_m^*}\big[-\alpha_m \log(\pi_m^*(A_m|S_m;\alpha_m)) - \alpha_m\mathcal{H}_{min}\big], \tag{15}$$

where $\pi_m^*(A_m|S_m;\alpha_m)$ denotes the optimal policy corresponding to temperature α_m. The dual gradient descent [41] is a promising solution for problem (15), where the objective is defined by

$$\mathcal{L}(\alpha) = \mathbb{E}_{A_m\sim\pi_m}\big[-\alpha\log\left(\pi_m\left(A_m|S_m\right)\right) - \alpha\mathcal{H}_{min}\big]. \tag{16}$$

Remark 2: One can observe that the optimal temperature depends on the optimal policy at each time step. Meanwhile, the optimal policy is also influenced by the temperature setting, which means that the policy and temperature update should be carried out iteratively.

The basic structure of SAC is based on the policy iteration algorithm which consists of policy evaluation and policy improvement. For policy evaluation, the goal is to evaluate the action values (i.e., Q-values) for a given policy π based on the Bellman expectation equation, which is given by

$$Q_\pi(s,a) = \mathcal{R}_s^a + \gamma\sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s'). \tag{17}$$

Different from the conventional state-value function, by taking the entropy into consideration, the *soft* state-value function for the maximum entropy framework is given by

$$v_\pi\left(s\right) = \mathbb{E}_{a\sim\pi}\left[Q_\pi(s,a) - \alpha\log\left(\pi(a|s)\right)\right]. \tag{18}$$

For continuous state space, the policy evaluation is not supported by tabular settings, and thus neural networks can be applied for practical approximation. The soft Q-network parameter ω is trained to minimize the following soft Bellman residual:

$$\mathcal{L}_Q\left(\omega\right) = \mathbb{E}_{(S_m,A_m)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\omega(S_m,A_m) - \hat{Q}(S_m,A_m)\right)^2\right], \tag{19}$$

where

$$\begin{aligned}\hat{Q}(S_m,A_m) &= \mathcal{R}_{S_m}^{A_m} + \gamma V_{\hat{\omega}}\left(S_{m+1}\right)\\ &= \mathcal{R}_{S_m}^{A_m} + \gamma\mathbb{E}_{A_{m+1}\sim\pi}\Big[Q_{\hat{\omega}}\left(S_{m+1},A_{m+1}\right)\\ &\quad - \alpha\log\left(\pi\left(A_{m+1}|S_{m+1}\right)\right)\Big].\end{aligned} \tag{20}$$

Here, $\mathcal{D}$ is the replay buffer storing transitions $\left(S_m, A_m, \mathcal{R}_{S_m}^{A_m}, S_{m+1}\right)$ following previous policies. $\hat{\omega}$ is the parameter for a target Q-network and duplicated from ω periodically. The terms $\mathcal{R}_{S_m}^{A_m}$ and S_{m+1} in (20) are fetched from replay buffer given S_m and A_m.

Remark 3: For continuous action space, the calculation of $V_{\hat{\omega}}\left(S_{m+1}\right)$ relies on the monte-carlo samples of actions following a policy distribution (such as the Gaussian distribution). For discrete action settings, however, the expectation can be directly derived employing discrete action probabilities, which makes the target value more tractable. Specifically, for discrete action settings in our work, the calculation of $\hat{Q}(S_m,A_m)$ in (20) can be rewritten as (21), where (21) is shown at the bottom of the next page.

For the policy improvement step, the aim is to improve the policy w.r.t. up-to-date Q-values obtained in the policy evaluation step. In [39], the continuous action space was considered. To make the policy tractable, i.e., following a type of distribution, the authors proposed to restrict the updated policy to some set of policies. It was proved that Q-values of the new policy increase when the policy is updated towards the exponential of current Q-values. Therefore, the policy is updated according to the following principle:

$$\pi_{\text{new}} = \arg\min_{\pi'\in\Pi} D_{\text{KL}}\left(\pi'\left(\cdot|S_m\right)\middle\|\frac{\exp\left(\frac{1}{\alpha}Q_{\pi_{\text{old}}}\left(S_m,\cdot\right)\right)}{X_{\pi_{\text{old}}}\left(S_m\right)}\right), \tag{22}$$

where D_{KL} is to calculate the Kullback-Leibler (KL) divergence which is to measure the similarity of two distributions. Here, the KL divergence is applied to project the improved policy $\frac{\exp\left(Q_{\pi_{\text{old}}}(S_m,\cdot)\right)}{\alpha X_{\pi_{\text{old}}}(S_m)}$ into the desired policy set Π. $X_{\pi_{\text{old}}}\left(S_m\right)$ is the normalization parameter which is dependant on the state S_m. Since $X_{\pi_{\text{old}}}\left(S_m\right)$ does not contribute to the gradient w.r.t. the new policy, the derivation of $X_{\pi_{\text{old}}}\left(S_m\right)$ can be ignored.

Aiming for the objective in (22), the loss function for policy network is defined in (23), as shown at the bottom of the next page. Since $X_{\pi_\vartheta}\left(S_m\right)$ depends only on the state, the loss function can be reduced to (multiplied by α) (24). For discrete action space, the expectation over actions in (24) can be calculated based on action probabilities. Therefore, (24) can be rewritten as (25), where (24) and (25) are shown at the bottom of the next page, respectively.

Based on the loss functions for the Q- and policy networks analyzed above, the weights ω and ϑ can be updated with stochastic gradients. The detailed workflow and pseudocode of SAC are shown in Fig. 3 and Algorithm 1, respectively. The algorithm contains two main parts: 1) the interaction with environment following the current policy; 2) the update of neural networks using stochastic gradients with data generated by previous policies, which is the core idea of off-policy RL. Compared to on-policy learning, off-policy learning has been proven to be more efficient by learning from the experience following policies other than the target one [42]. The pseudocode reflects the main idea of SAC by taking the entropy term into consideration in the loss function of critic and actor networks. From the aspect of the neural network structure, one trick used in the SAC algorithm is to adopt two Q-networks for critic and target networks, respectively. When optimizing the loss functions (19) and (24), the minimum of

the Q-functions is utilized. In this way, the training rate can speed up, especially for harder tasks. Since SAC follows the typical policy iteration framework, the SAC algorithm can be proved to converge to the optimal policy in the same way as given in [42] for the analysis of policy iteration, which is omitted here given the page limitation.

IV. Distributionally Robust SAC for UAV Trajectory Design, RIS Configuration and Power Control

As stated in Section III, the SAC algorithm enables sample-efficient learning via off-policy maximum entropy framework with a stochastic policy. However, due to the finite samples of data in practice, the inexact computation of policy state-values, i.e., *estimation errors*, may cause catastrophic policy outcome. Especially for our problem where the flight safety in environment with unexpected obstacles is taken into consideration, it is essential to train a policy with enhanced worst-case performance and support safe exploration. Therefore, to lower the risk in face of uncertain estimations, a novel distributionally robust DRL scheme is first proposed, followed by the distributionally robust SAC (DRSAC) algorithm design.

A. Distributionally Robust DRL

Recall that the policy iteration algorithm is an iterative process that alternates between the policy evaluation and the policy improvement, which can be expressed as follows:

$$\begin{cases} \pi_{i+1} \leftarrow \mathcal{G}(v_{i+1}), \\ v_{i+1} \leftarrow \mathcal{T}^{\pi_{i+1}} v_i, \end{cases} \tag{26}$$

where π_i and v_i denote the updated policy and state-value function at the i-th iteration, respectively. $\mathcal{T}^{\pi_{i+1}}$ denotes the Bellman operator applied for policy evaluation, which can be expressed as follows:

$$\mathcal{T}^{\pi} v(s) = \mathbb{E}_{a\sim\pi(\cdot|s)} \left[\mathcal{R}_s^a + \gamma \mathbb{E}_{s'\sim p(s'|s,a)} v(s')\right], \tag{27}$$

and $\mathcal{G}(v_i)$ is the greedy policy improvement approach, which is given by

$$\mathcal{G}(v) = \arg\max_{\pi} \mathcal{T}^{\pi} v. \tag{28}$$

The objective of a RL framework is given by

$$\max_{\pi} \sum_{m=1}^{M} \mathbb{E}_{(S_m,A_m)\sim\tau_\pi} \gamma^{m-1} \mathcal{R}_{S_m}^{A_m} = \max_{\pi} G(\pi). \tag{29}$$

Since the policy improvement step follows $\pi_{i+1} \leftarrow \mathcal{G}(v_i)$, estimation error ϵ for state value v_i will be reflected on the policy π_{i+1}. Let $\tilde{\epsilon} \in \mathbb{R}^{\mathcal{S}}$ denote the error sequence for the policy, then we can define the RL objective considering the robustness w.r.t. estimation error as follows:

$$\max_{\pi} \min_{\tilde{\epsilon}} G\left(\pi_{\tilde{\epsilon}}\right). \tag{30}$$

In [43], the authors introduced the KL-divergence to quantize the error on policy. To be more specific, given a policy π and an error sequence $\tilde{\epsilon} \in \mathbb{R}^{\mathcal{S}}$, the uncertainty set of policies is given by [43, Definition 2]

$$\mathcal{U}_{\tilde{\epsilon}}(\pi) = \left\{\pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}} | D_{\mathrm{KL}}\left(\pi'(\cdot|s) \,\|\pi(\cdot|s)\right) \leq \tilde{\epsilon}(s), \forall s \in \mathcal{S}\right\}, \tag{31}$$

where $\Delta_{\mathcal{A}}^{\mathcal{S}}$ denotes the set of probability distributions over a finite set $\mathcal{A}$ for all $s \in \mathcal{S}$. Then, the robust objective can be rewritten as

$$\begin{aligned} &\max_{\pi} \min_{\pi_{\tilde{\epsilon}} \in \mathcal{U}_{\tilde{\epsilon}}(\pi)} G\left(\pi_{\tilde{\epsilon}}\right) \\ &\quad = \max_{\pi} \min_{\pi_{\tilde{\epsilon}} \in \mathcal{U}_{\tilde{\epsilon}}(\pi)} \mathbb{E}_{(S_m,A_m)\sim\tau_{\pi_{\tilde{\epsilon}}}} \gamma^{m-1} \mathcal{R}_{S_m}^{A_m}, \end{aligned} \tag{32}$$

which follows the typical distributionally robust optimization (DRO) format [44], [45]. To solve the DRO problem under a RL framework, we need to again refer to the policy iteration

$$\hat{Q}(S_m, A_m) = \mathcal{R}_{S_m}^{A_m} + \gamma \sum_{A_{m+1}\in\mathcal{A}} \pi\left(A_{m+1}|S_{m+1}\right) \Big[Q_{\hat{\omega}}\left(S_{m+1}, A_{m+1}\right) - \alpha \log\left(\pi\left(A_{m+1}|S_{m+1}\right)\right)\Big]. \tag{21}$$

$$\begin{aligned} \mathcal{L}_\pi(\vartheta) &= \mathbb{E}_{S_m\in\mathcal{D}} \left[D_{\mathrm{KL}}\left(\pi_\vartheta(\cdot|S_m) \middle\| \frac{\exp\left(\frac{1}{\alpha} Q_\omega(S_m, \cdot)\right)}{X_{\pi_\vartheta}(S_m)}\right)\right] \\ &= \mathbb{E}_{S_m\in\mathcal{D}} \left[\int_{A_m} \pi_\vartheta(A_m|S_m) \times \left(\log\left(\pi_\vartheta(A_m|S_m)\right) - \frac{1}{\alpha} Q_\omega(S_m, A_m) + \log X_{\pi_\vartheta}(S_m)\right)\right] \\ &= \mathbb{E}_{S_m\in\mathcal{D}} \mathbb{E}_{A_m\sim\pi_\vartheta} \left(\log\left(\pi_\vartheta(A_m|S_m)\right) - \frac{1}{\alpha} Q_\omega(S_m, A_m) + \log X_{\pi_\vartheta}(S_m)\right). \end{aligned} \tag{23}$$

$$\mathcal{L}_\pi(\vartheta) = \mathbb{E}_{S_m\in\mathcal{D}} \mathbb{E}_{A_m\sim\pi_\vartheta} \left(\alpha \log\left(\pi_\vartheta(A_m|S_m)\right) - Q_\omega(S_m, A_m)\right). \tag{24}$$

$$\mathcal{L}_\pi(\vartheta) = \mathbb{E}_{S_m\in\mathcal{D}} \sum_{A_m\in\mathcal{A}} \pi_\vartheta(A_m|S_m) \Big(\alpha \log\left(\pi_\vartheta(A_m|S_m)\right) - Q_\omega(S_m, A_m)\Big). \tag{25}$$

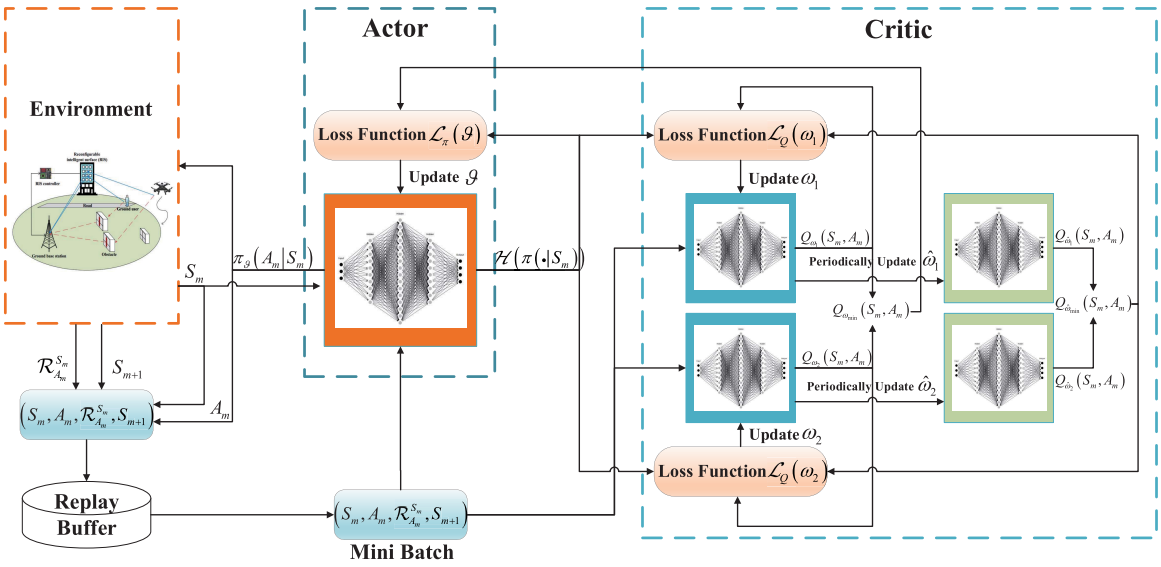


Fig. 3. Workflow of the proposed SAC framework.

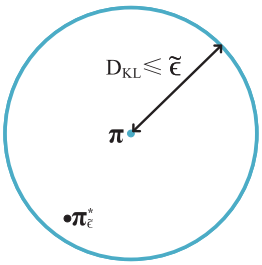


Fig. 4. Uncertainty set of policies $\mathcal{U}_{\tilde{\epsilon}}(\pi)$.

process in (26). For the inner minimization problem in (32), an *adversarial Bellman operator* $\mathcal{T}^{\pi_{\tilde{\epsilon}}^*}$ is defined as follows:

$$\mathcal{T}^{\pi_{\tilde{\epsilon}}^*} v(s) = \min_{\tilde{\pi} \in \mathcal{U}_{\tilde{\epsilon}}(\pi)} \mathcal{T}^{\tilde{\pi}} v(s). \tag{33}$$

By applying $\mathcal{T}^{\pi_{\tilde{\epsilon}}^*}$ for policy evaluation, the minimal state values achieved by policies in the uncertainty set $\mathcal{U}_{\tilde{\epsilon}}(\pi)$ can be obtained, as shown in Fig. 4. Therefore, the policy evaluation with $\mathcal{T}^{\pi_{\tilde{\epsilon}}^*}$ can be named as *distributionally robust policy evaluation*. To derive the computation scheme of adversarial Bellman operator, apply Lagrangian duality to (31) and (33). Then, the problem can be rewritten in (34), as shown at the bottom of the next page, where $\lambda(s)$ is the Lagrange multiplier.

The inner maximization problem can be expressed as

$$\begin{aligned}
&\max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left(-\mathcal{T}^{\tilde{\pi}} v(s) - \lambda(s) D_{\mathrm{KL}} \left(\tilde{\pi}(\cdot|s) \,\|\pi(\cdot|s) \right) \right) \\
&= \max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \lambda(s) \left(-\frac{1}{\lambda(s)} \mathcal{T}^{\tilde{\pi}} v(s) - D_{\mathrm{KL}} \left(\tilde{\pi}(\cdot|s) \,\|\pi(\cdot|s) \right) \right) \\
&= \max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \lambda(s) \left(\left\langle -\frac{Q_v(s,\cdot)}{\lambda(s)}, \tilde{\pi}(\cdot|s) \right\rangle - D_{\mathrm{KL}} \left(\tilde{\pi}(\cdot|s) \,\|\pi(\cdot|s) \right) \right) \\
&= \lambda(s) \Omega^* \left(-\frac{Q_v(s,\cdot)}{\lambda(s)} \right),
\end{aligned} \tag{35}$$

where $\Omega^* \left(-\frac{Q_v(s,\cdot)}{\lambda(s)} \right)$ is the Fenchel duality of $\Omega(\tilde{\pi}(\cdot|s)) = D_{\mathrm{KL}}(\tilde{\pi}(\cdot|s) \| \pi(\cdot|s))$. The Fenchel duality of D_{KL} is given by

$$\Omega^* \left(-\frac{Q_v(s,\cdot)}{\lambda(s)} \right) = \log \mathbb{E}_{a \in \tilde{\pi}} \exp \left(-\frac{Q_v(s,a)}{\lambda(s)} \right), \tag{36}$$

and the solution for problem (35) is

$$\pi_{\tilde{\epsilon}}^* \propto \exp \left(-\frac{Q_v(s,a)}{\lambda^*(s)} \right) \pi(a|s). \tag{37}$$

For outer minimization problem, the optimal solution $\lambda^*(s)$ is given by

$$\lambda^*(s) = \arg\min_{\lambda(s)>0} \left(\lambda(s) \Omega^* \left(-\frac{Q_v(s,\cdot)}{\lambda(s)} \right) + \lambda(s) \tilde{\epsilon}(s) \right), \tag{38}$$

which is a typical convex optimization problem. The construction of policy error $\tilde{\epsilon}(s)$ is in the form $\tilde{\epsilon}(s) = Cn(s)^{-\eta}$ with constants $C > 0$ and $\eta > 0$. $n(s)$ denotes the visitation count of state s. This construction implies that the estimation error should decrease with the amount of collected experience.

B. Distributionally Robust Soft Actor-Critic

The difference of SAC from the regular RL framework is the introduction of per-state entropy bonus. As stated in Section III, the policy improvement for discrete action setting is based on the following principle:

$$\pi(a|s) \propto \exp \left(\frac{1}{\alpha} Q_v(s,a) \right), \tag{39}$$

where α is the entropy temperature. Substituting (39) into (37), we can obtain the adversarial policy in SAC as follows:

$$\pi_{\tilde{\epsilon}}^* \propto \exp \left[\left(\frac{1}{\alpha} - \frac{1}{\lambda^*(s)} \right) Q_v(s,a) \right], \tag{40}$$

where $\lambda^*(s)$ is given in (38).

Algorithm 1 Soft Actor-Critic Algorithm

Initialize environment;
Initialize $\omega_i (i = 1, 2)$ for critic networks, ϑ for actor network;
Initialize target networks $\hat{\omega}_i \leftarrow \omega_i, i = 1, 2$;
Initialize entropy level $\mathcal{H}_{\min}$, replay buffer $\mathcal{D} = \emptyset$, step length for gradient descent of the critic network, actor network and temperature parameter, i.e., λ_Q, λ_π, λ_α;
for *each iteration* **do**
  for *each environment step* **do**
    Execute action based on current policy $A_m \sim \pi_\vartheta$;
    Observe reward $\mathcal{R}_{S_m}^{A_m}$ and next state S_{m+1};
    Store transition $(S_m, A_m, \mathcal{R}_{S_m}^{A_m}, S_{m+1})$ in $\mathcal{D}$;
  for *each gradient step* **do**
    Sample a random minibatch of transitions $(S_m, A_m, \mathcal{R}_{S_m}^{A_m}, S_{m+1})$ from $\mathcal{D}$;
    Update critic networks by minimizing loss function $\mathcal{L}_Q(\omega)$ with stochastic gradients: $\omega_i \leftarrow \omega_i - \lambda_Q \hat{\nabla}_{\omega_i} \mathcal{L}_Q(\omega_i), \forall i \in \{1, 2\}$;
    Update the actor network by minimizing loss function $\mathcal{L}_\pi(\vartheta)$: $\vartheta \leftarrow \vartheta - \lambda_\pi \hat{\nabla}_\vartheta \mathcal{L}_\pi(\vartheta)$;
    Update temperature parameter by minimizing $\mathcal{L}(\alpha)$: $\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha \mathcal{L}(\alpha)$;
    Update target network parameters periodically: $\hat{\omega}_i \leftarrow \omega_i, \forall i \in \{1, 2\}$
return optimal policy π^*

Based on the adversarial policy expression given in (40) for policy evaluation and the policy improvement principle based on (39), the detailed algorithm for DRSAC is given in Algorithm 2. The main difference between DRSAC and SAC is in the policy evaluation step. Instead of updating the critic network towards the true action-value function for the current policy π, the adversarial policy $\pi_{\tilde{\epsilon}}^*$ is adopted for action-value estimation (as shown in the pseudocode for the update of $\hat{Q}(S_m, A_m)$) to provide the lower-bound performance guarantee.

Next we will analyze the properties of the proposed DRSAC algorithm in terms of convergence and optimality. Starting from the construction of policy error $\tilde{\epsilon}(s)$ in the form $\tilde{\epsilon}(s) = Cn(s)^{-\eta}$ with $C > 0$ and $\eta > 0$, one can observe that the error gets smaller with accumulated experience. To make the explanation more intuitive, as can be observed in Fig. 4, the radius of the uncertainty set shrinks with the learning process. Therefore, with $m \to +\infty$, the adversarial policy $\pi_{\tilde{\epsilon}}^*$ converges to the policy π. In other words, the algorithm performs conservatively in a short-term and acts optimistically in a long run. The convergence and optimality can then be guaranteed similarly to SAC.

Algorithm 2 Distributionally Robust Soft Actor-Critic Algorithm

Initialize environment;
Initialize critic network, actor network, replay buffer $\mathcal{D} = \emptyset$;
Set $\mathcal{H}_{\min}$, C, η, $n(s) = 0, \forall s \in \mathcal{S}$;
for *each iteration* **do**
  for *each environment step* **do**
    Execute action based on the current policy;
    Store transition $(S_m, A_m, \mathcal{R}_{S_m}^{A_m}, S_{m+1})$ in $\mathcal{D}$;
  for *each gradient step* **do**
    Sample a random minibatch of transitions $(S_m, A_m, \mathcal{R}_{S_m}^{A_m}, S_{m+1})$ from $\mathcal{D}$;
    $\tilde{\epsilon}(s) \leftarrow Cn(s)^{-\eta}$;
    Solve convex optimization problem (38) to obtain $\lambda^*(s)$;
    $\pi_{\tilde{\epsilon}}^*(a|s) \leftarrow \frac{\exp\left[\left(\frac{1}{\alpha} - \frac{1}{\lambda^*(s)}\right) Q_v(s,a)\right]}{\sum_a \exp\left[\left(\frac{1}{\alpha} - \frac{1}{\lambda^*(s)}\right) Q_v(s,a)\right]}$;
    $\hat{Q}(S_m, A_m) \leftarrow \mathcal{R}_{S_m}^{A_m} + \gamma \mathbb{E}_{A_{m+1} \sim \pi_{\tilde{\epsilon}}^*} \left[Q_{\hat{\omega}}(S_{m+1}, A_{m+1}) - \alpha \log\left(\pi(A_{m+1}|S_{m+1})\right)\right]$;
    Update actor, critic networks and α as in Algorithm 1.
return optimal policy π^*

V. Numerical Results

In this section, numerical results are provided to validate the effectiveness of the proposed RIS-aided air-to-ground uplink NOMA communication framework. We consider a system where the initial location of the UAV is $(0, 0, 60)$ meters, while the locations of GU, GBS and RIS are $(-100, -100, 0)$ meters, $(300, -50, 40)$ meters, and $(200, 80, 60)$ meters, respectively. The speed of the UAV is set to 20 m/s. The GU-RIS, UAV-RIS, and RIS-GBS links are modelled as Rician fading channels. Therefore, $\mathbf{h}_{gu,s}[m]$ can be expressed as

$$\mathbf{h}_{gu,s}[m] = \sqrt{\frac{\kappa}{1+\kappa}} \mathbf{h}_{gu,s}^{\text{LoS}}[m] + \sqrt{\frac{1}{1+\kappa}} \mathbf{h}_{gu,s}^{\text{NLoS}}[m], \quad (41)$$

where κ is the Rician factor, $\mathbf{h}_{gu,s}^{\text{LoS}}[m]$ is the LoS component, and $\mathbf{h}_{gu,s}^{\text{NLoS}}[m]$ is the NLoS component. $\mathbf{h}_{gu,s}^{\text{LoS}}[m]$ is given by

$$\mathbf{h}_{gu,s}^{\text{LoS}}[m] = \sqrt{\beta_0 \left(d_{gu,s}^{(1)}[m]\right)^{-\alpha_1}} \left[e^{-j\frac{2\pi}{\lambda} d_{gu,s}^{(1)}[m]}, \ldots, e^{-j\frac{2\pi}{\lambda} d_{gu,s}^{(k)}[m]}, \ldots, e^{-j\frac{2\pi}{\lambda} d_{gu,s}^{(K)}[m]} \right]^T, \quad (42)$$

where β_0 is the path loss at the reference distance $d_0 = 1$ m, α_1 is the corresponding path loss exponent, $d_{gu,s}^{(k)}$ is

$$\begin{aligned} \mathcal{T}^{\pi_{\tilde{\epsilon}}^*} v(s) &= \max_{\lambda(s)>0} \min_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left(\mathcal{T}^{\tilde{\pi}} v(s) + \lambda(s) D_{\text{KL}}\left(\tilde{\pi}(\cdot|s) \,\|\pi(\cdot|s)\right) - \lambda(s)\tilde{\epsilon}(s) \right) \\ &= \min_{\lambda(s)>0} \max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left(-\mathcal{T}^{\tilde{\pi}} v(s) - \lambda(s) D_{\text{KL}}\left(\tilde{\pi}(\cdot|s) \,\|\pi(\cdot|s)\right) + \lambda(s)\tilde{\epsilon}(s) \right), \end{aligned} \quad (34)$$

TABLE I
SYSTEM PARAMETERS

α_1	Path loss parameter for LoS transmissions	2.1
α_2	Path loss exponent for NLoS transmissions	3.5
κ	Rician factor	10
β_0	Path loss at 1 m	-20 dB
σ_b^2	Noise power	-80 dBm
δ_t	Duration of each time slot	1 s
M	Number of time slots	10
L	Number of discrete phase-shift levels	2
K	Number of reflecting elements in each RIS sub-surface	10
$d_{\min}$	Minimum separation distance	20 m

TABLE II
DRL HYPERPARAMETERS

Number of training episodes	200000
Replay memory size	5000
Mini-batch size	64
Gradient descent step length λ_Q, λ_π, λ_α	0.00001
Constant C	1
Constant η	0.5
Optimizer	Adam
Activation function	ReLU

the distance from the GU to the k-th RIS element, and λ is the carrier wavelength. Due to the fact that the distance between the RIS and UAV is much bigger than that among RIS elements, we use the 1-st RIS element as the reference point for path loss calculation. $\mathbf{h}_{gu,s}^{\text{NLoS}}[m]$ is given by $\mathbf{h}_{gu,s}^{\text{NLoS}}[m] = \sqrt{\beta_0 \left(d_{gu,s}^{(1)}[m]\right)^{-\alpha_2}} \hat{\mathbf{h}}_{gu,s}$, where $\hat{\mathbf{h}}_{gu,s} \in \mathbb{C}^{N\times 1}$ is the small-scale fading component where elements are independently drawn from the circularly symmetric complex Gaussian (CSCG) distribution with unit variance. $\mathbf{h}_{u,s}$ and $\mathbf{h}_{s,b}^H$ can be generated similarly as (41).

The UAV-GBS and GU-GBS links are modelled as Rayleigh fading channels. Then, $h_{gu,b}[m]$ and $h_{u,b}[m]$ are given by

$$h_{u,b}[m] = \sqrt{\beta_0 \left(d_{u,b}[m]\right)^{-\alpha_2}} \hat{h}_{u,b}[m], \tag{43}$$

and

$$h_{gu,b}[m] = \sqrt{\beta_0 \left(d_{gu,b}[m]\right)^{-\alpha_2}} \hat{h}_{gu,b}[m], \tag{44}$$

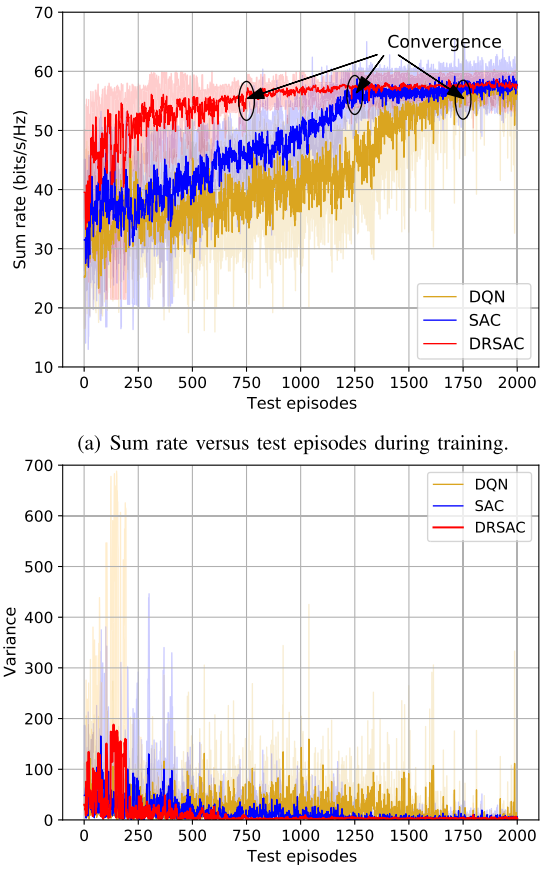
respectively. Here, $d_{u,b}[m]$ and $d_{gu,b}[m]$ represent the distance between the UAV and GBS and that between the GU and GBS, respectively, α_2 denotes the corresponding path loss exponent, and $\hat{h}$ is the small-scale NLoS component. The specific values of adopted system parameters and DRL hyperparameters are summarized in Table I and Table II, respectively.

A. Performance Comparison Among DQN, SAC, and DRSAC

In Fig. 5, we compare the performance of SAC and DRSAC against the benchmark algorithm DQN. For DQN, we consider the following setup:

- **Benchmark 1 (DQN)**: conventional DQN algorithm that does not maximize the entropy. We set the initial exploration probability as 0.9, the minimum exploration probability as 0.05, and the learning rate α as 0.00001.

We train five instances of each algorithm with different random seeds, with each performing one evaluation rollout every 100 episodes. Here different seeds mean running algorithm with various random initialization of the actor and critic networks.



(a) Sum rate versus test episodes during training.

(b) Variance versus test episodes during training.

Fig. 5. Performance comparison among DQN, SAC and DRSAC.

Fig. 5(a) shows the system sum rate with test episodes during training for DQN, SAC, and DRSAC. Note that the test episodes mean evaluation rollouts every 100 training episodes. The solid curves correspond to the mean and the shaded regions to the minimum and maximum sum rate bounds over the five trials. One can observe from the results that both SAC and DRSAC outperform the DQN algorithm with a large margin in terms of learning rate. Specifically, DRSAC and SAC converge at around 750 and 1250 test episodes, respectively, while the DQN algorithm obtains stable outcomes after 1750 test episodes. This is expected since both the SAC and DRSAC have a higher sample-efficiency compared to DQN. In Fig. 5(a), it is also noted that the sum rate lower bound of DRSAC is clearly higher than that of SAC. This empirically confirms the safety guarantee provided by DRSAC. Fig. 5(b) further illustrates the robustness of DRSAC from the aspect of variance obtained over 5 runs with different random seeds. As expected, the DRSAC algorithm greatly reduces the variance thanks to the robustness considered in the learning process. For the merits of DRSAC mentioned above, in the following, we adopt results obtained by the DRSAC algorithm to evaluate the system performance.

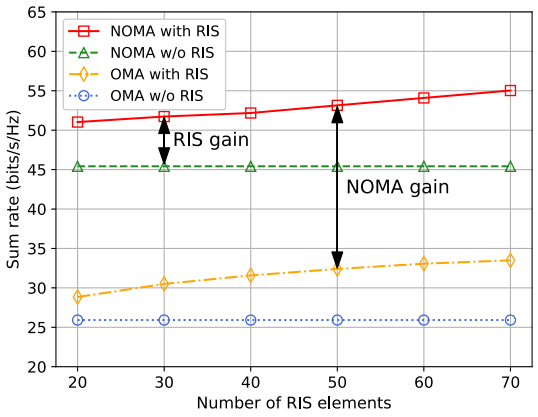
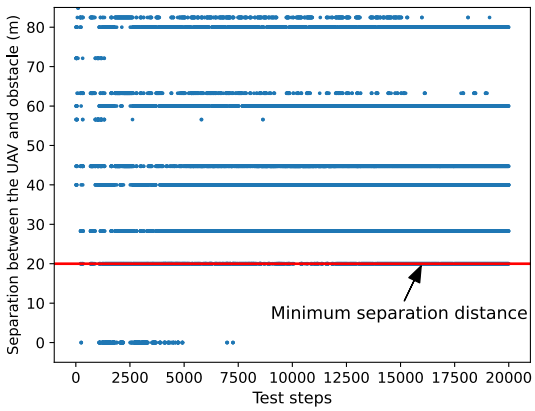
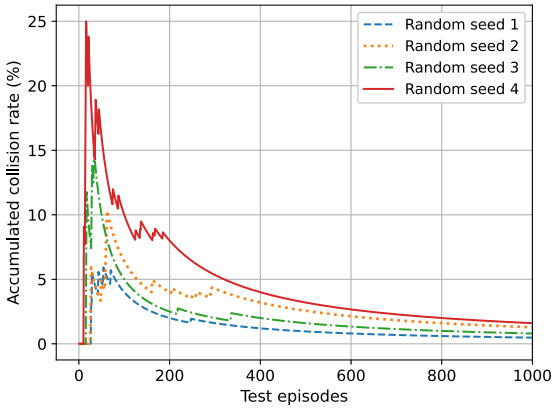


Fig. 6. Illustration on impact of NOMA protocol and RIS deployment.



(a) Separation distance between the UAV and obstacle versus test steps during training.



(b) Accumulated collision rate versus test episodes during training.

Fig. 7. Performance analysis on collision avoidance mechanism.

B. Impact of NOMA Protocol and RIS Deployment

In Fig. 6, we investigate the system performance gain brought by the NOMA protocol and RIS deployment. Specifically, we consider the following three benchmarks.

- **Benchmark 2 (NOMA-w/o-RIS case)**: the case where communication links from the UAV/GU to the GBS only include the direct links;
- **Benchmark 3 (OMA-RIS case)**: the case where time/frequency resources are split equally between the UAV and GU. Therefore, the sum rate for the OMA case over time span T is given by $\sum_{m=1}^{M}\left(\frac{1}{2}\log_2(1+\gamma_u[m])+\frac{1}{2}\log_2(1+\gamma_{gu}[m])\right)$;
- **Benchmark 4 (OMA-w/o-RIS case)**: the same as OMA-RIS case except that no RIS is deployed.

It is worth noting that the proposed DRSAC algorithm can be also applied to the three benchmark schemes. In particular, the sum rate maximization problem for the NOMA-w/o-RIS case can be solved with DRSAC by excluding the RIS phase shift in the action space. For the OMA-RIS case, the optimization problem can be solved assuming that both of the UAV and GU occupy half of the total bandwidth and transmit in full power. For the OMA-w/o-RIS case, the optimization problem is solved by just optimizing the UAV trajectory. We observe from Fig. 6 that the sum rate increases with the number of RIS elements K when the RIS is deployed, which is due to the fact that larger K leads to higher beamforming gain. Regarding the performance of the proposed air-to-ground communication scheme and three benchmarks, the proposed scheme achieves the highest sum rate. This can be explained by the improved spectrum efficiency brought by the NOMA protocol as well as the favorable propagation environment created by the RIS.

C. Impact of Online UAV Collision Avoidance

In Fig. 7, we investigate effectiveness of the proposed DRSAC algorithm for the UAV collision avoidance mechanism. To visualize performance of the proposed algorithm, we consider an obstacle appearing in the airspace with random location, and show the distance between the UAV and obstacle as well as the accumulated collision rate (ACR) during learning process in Fig. 7(a) and Fig. 7(b), respectively. Similar to Section V-A, the evaluation rollout is carried out every 100 episodes, i.e., 1000 time steps.

Fig. 7(a) demonstrates the separation distance between the UAV and obstacle with test steps during the training process. Here, the term "test step" is defined as the accumulated time steps over test episodes. One can observe that the collision happens with high probabilities at the beginning of the learning process. With the increment of training episodes, the DRSAC model learns to keep a safe distance to the obstacle and avoids collision effectively. Note that the separation distance between the UAV and obstacle is a discrete value due to the discrete heading angles and fixed velocity of the UAV. To further prove that the proposed algorithm can work well under different environment setups, we depict Fig. 7(b) to show the ACR with different random seeds. Here we define ACR as

$$\text{ACR} = \frac{\text{Number of episodes with collisions}}{\text{Number of total training episodes}}. \tag{45}$$

We observe that the ACR first increases with the number of training episodes due to the reason that the algorithm is still in the exploration stage. With the increment of number of training episodes, the ACR gets smaller until converges at around 800 test episodes. We also observe that the algorithm can converge to a small ACR value under different seeds, which shows the robustness of the proposed algorithm in terms of

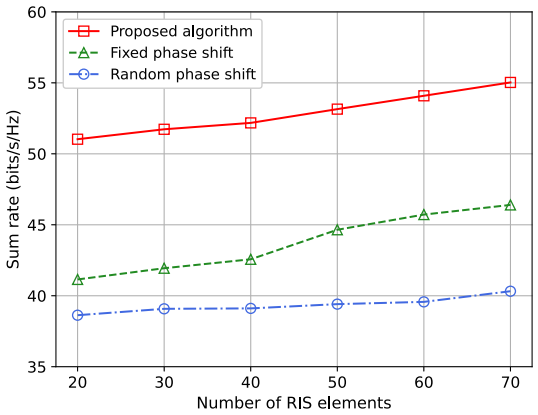


Fig. 8. Illustration on impact of phase shift configuration.

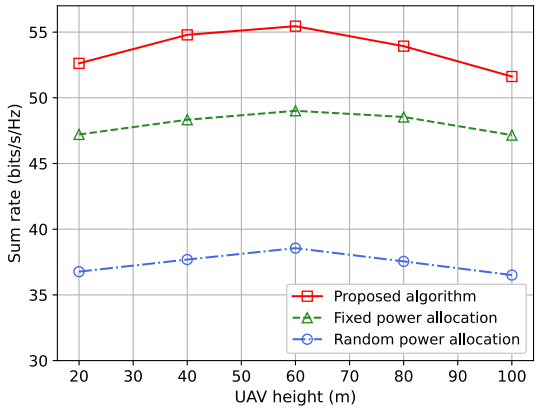


Fig. 9. Illustration on impact of power control.

achieving satisfying performance under various initialization setups.

D. Impact of Phase Shift Configuration

In this subsection, we compare the performance of the proposed algorithm with two benchmark algorithms as follows:

- **Benchmark 5 (Fixed phase shift)**: the proposed DRSAC algorithm that only takes the UAV's movement and power control into consideration. The phase shifts of all RIS sub-surfaces at each time slot, i.e., $\phi_n[m]$, are given as 0;
- **Benchmark 6 (Random phase shift)**: similar to the fixed phase shift scheme except that the phase shift value of each RIS sub-surface at each time slot is generated randomly.

Fig. 8 characterizes the performance of the proposed algorithm against the above two benchmark algorithms. We observe that the proposed algorithm achieves distinct sum rate improvement compared to the benchmarks, which indicates the efficiency of our proposed algorithm to solve the RIS configuration problem. We also find that the fixed phase shift algorithm obtains higher sum rate compared to the random phase shift one. This is due to the fact that the randomness w.r.t. phase shift increases the difficulty for signal alignment, and thus leads to some performance loss.

E. Impact of Power Control

Fig. 9 depicts the sum rate versus height of the UAV, i.e., z_u, obtained by different algorithms with $N = 5$. The two benchmarks, i.e., fixed power control and random power control, are designed as follows:

- **Benchmark 7 (Fixed power control)**: the proposed DRSAC algorithm that only takes the UAV's movement and RIS configuration into consideration. The transmit powers of the UAV and GU are set as 1 W and 0.4 W, respectively;
- **Benchmark 8 (Random power control)**: similar to the fixed power allocation scheme except that the transmit powers of the UAV and GU are generated randomly under the maximum transmit power and minimum SINR constraints.

We observe that the sum rate grows rapidly with a small altitude and decreases afterwards. The inflection point of each curve shows up at around $z_u = 60$ m. This verifies that the UAV-RIS-GBS link is enhanced when the UAV gets closer to the RIS which has the altitude of 60 m. We also observe that the proposed algorithm obtains much better performance compared to the two benchmarks, which implies the effectiveness of the proposed algorithm for power control.

VI. Conclusion

A RIS-aided air-to-ground uplink non-orthogonal transmission framework has been investigated. The UAV trajectory, RIS configuration, and uploading power control were jointly optimized for the maximization of sum rate, subject to the constraints on the UAV flight safety and the minimum data rate requirements of both the UAV and GU. A sample-efficient DRL algorithm was proposed to address the resultant sequential decision making problem. Considering uncertainties brought by the unknown locations of obstacles, a distributionally robust DRL algorithm was further proposed to enhance the robustness of the algorithm. Our numerical results demonstrated that the two proposed DRL algorithms outperformed the conventional ones in terms of learning efficiency and robustness. The results also revealed that the sum rate of air-to-ground communications can be significantly improved by optimizing the UAV trajectory, RIS configuration, and power control. Moreover, this paper considered the single-cell scenario to give fundamental insights on the system performance. The multi-cell scenario, where the RISs provide both the signal enhancement and inter-cell interference mitigation, is expected to be an interesting topic for future work.

References

[1] 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.

[2] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[3] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1986–1999, Sep. 2018.

[4] Z. Han, A. L. Swindlehurst, and K. J. R. Liu, "Optimization of MANET connectivity via smart deployment/movement of unmanned air vehicles," *IEEE Trans. Veh. Technol.*, vol. 58, no. 7, pp. 3533–3546, Sep. 2009.

[5] Z. Xiao, L. Zhu, and X.-G. Xia, "UAV communications with millimeter-wave beamforming: Potentials, scenarios, and challenges," *China Commun.*, vol. 17, no. 9, pp. 147–166, Sep. 2020.

[6] B. Van Der Bergh, A. Chiumento, and S. Pollin, "LTE in the sky: Trading off propagation benefits with interference costs for aerial nodes," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 44–50, May 2016.

[7] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[8] L. Liu, S. Zhang, and R. Zhang, "Exploiting NOMA for multi-beam UAV communication in cellular uplink," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[9] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[10] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[11] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[12] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[13] M. Di Renzo and J. Song, "Reflection probability in wireless networks with metasurface-coated environmental objects: An approach based on random spatial processes," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 99, pp. 1–15, Apr. 2019.

[14] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable ai meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, 2019.

[15] Y. Liu *et al.*, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, 3rd Quart., 2021.

[16] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "UAV-enabled communication using NOMA," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5126–5138, Jul. 2019.

[17] R. Tang, J. Cheng, and Z. Cao, "Joint placement design, admission control, and power allocation for NOMA-based UAV systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 385–388, Mar. 2020.

[18] T. M. Nguyen, W. Ajib, and C. Assi, "A novel cooperative NOMA for designing UAV-assisted wireless backhaul networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2497–2507, Nov. 2018.

[19] Z. Na, Y. Liu, J. Shi, C. Liu, and Z. Gao, "UAV-supported clustered NOMA for 6G-enabled Internet of Things: Trajectory planning and resource allocation," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15041–15048, Oct. 2021.

[20] X. Mu, Y. Liu, L. Guo, and J. Lin, "Non-orthogonal multiple access for air-to-ground communication," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2934–2949, May 2020.

[21] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, Jan. 2020.

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

[23] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.

[24] Z. Wei *et al.*, "Sum-rate maximization for IRS-assisted UAV OFDMA communication systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2530–2550, Apr. 2021.

[25] Y. Pan, K. Wang, C. Pan, H. Zhu, and J. Wang, "UAV-assisted and intelligent reflecting surfaces-supported terahertz communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1256–1260, Jun. 2021.

[26] X. Cao *et al.*, "Reconfigurable intelligent surface-assisted aerial-terrestrial communications via multi-task learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3035–3050, Oct. 2021.

[27] A. Ranjha and G. Kaddoum, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, Mar. 2021.

[28] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Joint deployment and multiple access design for intelligent reflecting surface assisted networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6648–6664, Oct. 2021.

[29] M. Zeng, A. Yadav, O. Dobre, and H. V. Poor, "Energy-efficient joint user-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5119–5131, Feb. 2019.

[30] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 269–273, Jan. 2021.

[31] M. Najafi, V. Jamali, R. Schober, and H. V. Poor, "Physics-based modeling and scalable optimization of large intelligent reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2673–2691, Apr. 2021.

[32] B. Zheng, C. You, and R. Zhang, "Fast channel estimation for IRS-assisted OFDM," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 580–584, Mar. 2021.

[33] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, Apr. 2020.

[34] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[35] E. Shtaiwi, H. Zhang, S. Vishwanath, M. Youssef, A. Abdelhadi, and Z. Han, "Channel estimation approach for RIS assisted MIMO systems," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 452–465, Jun. 2021.

[36] U. Mengali and A. N. D'Andréa, *Synchronization Techniques for Digital Receivers* New York, NY, USA: Springer, 1997.

[37] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[38] Y. Duan, X. Chen, X. Houthooft, R. Schulman, and P. Abbbeel, "Benchmarking deep reinforcement learning for continuous control," in *Proc. Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, Jul. 2016, pp. 1–5.

[39] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 1861–1870.

[40] T. Haarnoja *et al.*, "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[42] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Oct. 2015.

[43] E. Smirnova, E. Dohmatob, and J. Mary, "Distributionally robust reinforcement learning," 2019, *arXiv:1902.08708*.

[44] Y. Chen, B. Ai, Y. Niu, H. Zhang, and Z. Han, "Energy-constrained computation offloading in space-air-ground integrated networks using distributionally robust optimization," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 12113–12125, Nov. 2021.

[45] D. Zhou, M. Sheng, B. Li, J. Li, and Z. Han, "Distributionally robust planning for data delivery in distributed satellite cluster network," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3642–3657, Jul. 2019.

Jingjing Zhao (Member, IEEE) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013, and the Ph.D. degree from the Queen Mary University of London, London, U.K., in 2017. From 2017 to 2018, she was a Post-Doctoral Research Fellow with the Department of Informatics, King's College London, London, U.K. From 2018 to 2020, she was a Researcher at Amazon, London. Currently, she is an Associate Professor with the Research Institute for Frontier Science, Beihang University, Beihang. Her current research interests include non-orthogonal multiple access, reconfigurable intelligent surfaces, aeronautical broadband communications, and machine learning.

Lanchenhui Yu (Student Member, IEEE) received the B.S. and M.S. degrees from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beihang University, China. His research interests include intelligent air navigation and aeronautical broadband communications.

Kaiquan Cai (Member, IEEE) received the B.S. and Ph.D. degrees from Beihang University in 2004 and 2013, respectively. He is currently a Professor with the School of Electronic and Information Engineering, Beihang University, and the Deputy Director of the National Key Laboratory of CNS/ATM. His research interests include intelligent air navigation and networked collaborative air traffic management.

Yanbo Zhu (Member, IEEE) received the B.S. and Ph.D. degrees from Beihang University in 1995 and 2009, respectively. He is currently the Vice President of Aviation Data Communication Corporation, China. He is also a part-time Professor with the School of Electronic and Information Engineering, Beihang University. His research interests include intelligent air navigation, aeronautical datalink communications, and collaborative air traffic management.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University in 1997 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an Research and Development Engineer of JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an Assistant Professor at Boise State University, ID, USA. Currently, he is a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as with the Computer Science Department, University of Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and has been an AAAS Fellow since 2019 and an ACM Distinguished Member since 2019. He received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of communications systems (the Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He has been among the 1% highly cited researchers since 2017 according to Web of Science. He is also the winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."