Simultaneously Transmitting And Reflecting Reconfigurable Intelligent Surface (STAR-RIS) Assisted UAV Communications

Jingjing Zhao, Yanbo Zhu, Xidong Mu, Kaiquan Cai, Yuanwei Liu, *Senior Member, IEEE,* and Lajos Hanzo, *Life Fellow, IEEE*

***Abstract*—A novel air-to-ground communication paradigm is conceived, where an unmanned aerial vehicle (UAV)-mounted base station (BS) equipped with multiple antennas sends information to multiple ground users (GUs) with the aid of a simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS). In contrast to the conventional RIS whose main function is to reflect incident signals, the STAR-RIS is capable of both transmitting and reflecting the impinging signals from either side of the surface, thereby leading to full-space 360 degree coverage. However, the transmissive and reflective capabilities of the STAR-RIS require more complex transmission/reflection coefficient design. Therefore, in this work, a sum-rate maximization problem is formulated for the joint optimization of the UAV's trajectory, the active beamforming at the UAV, and the passive transmission/reflection beamforming at the STAR-RIS. This cutting-edge optimization problem is also subject to the UAV's flight safety, to the maximum flight duration constraint, as well as to the GUs' minimum data rate requirements. Given the unknown locations of obstacles prior to the UAV's flight, we provide an online decision making framework employing reinforcement learning (RL) to simultaneously adjust both the UAV's trajectory as well as the active and passive beamformer. To enhance the system's robustness against the associated uncertainties caused by limited sampling of the environment, a novel "distributionally-robust" RL (DRRL) algorithm is proposed for offering an adequate worst-case performance guarantee. Our numerical results unveil that: 1) the STAR-RIS assisted UAV communications benefit from significant sum-rate gain over the conventional reflecting-only RIS; and 2) the proposed DRRL algorithm achieves both more stable and more robust performance than the state-of-the-art RL algorithms.**

***Index Terms*—Air-to-ground communications, simultaneously transmitting and reflecting reconfigurable intelligent surface, joint beamforming design, collision avoidance, distributionally-robust reinforcement learning.**

I. Introduction

Thanks to the advantages of agile mobility, flexible deployment and low cost, the employment of unmanned aerial vehicles (UAVs) as flying communication platforms has attracted fast-growing interests in the past several years [1]. In contrast to terrestrial wireless communications, the maneuverability enables the dynamic adjustment of UAVs' positions to best suite the communication environment. Therefore, the UAVs are expected to bring in promising gains to the next-generation wireless communications by acting as different types of communication platforms, such as aerial base stations (BSs), mobile relays, and aerial users [2], [3]. Particularly, employing UAVs as aerial BSs is envisioned as a promising solution to offload traffic from the operational wireless networks as well as to provide/recover wireless services for temporary hotspots or emergencies [4]. Compared to terrestrial BSs, the flexible movement of UAVs in the three-dimensional (3D) space can be fully exploited to enhance the coverage area and the communication throughput.

Despite the above merits of the UAVs, one of the challenging issues for the efficient facilitation of air-to-ground (A2G) communications is that the A2G links may become blocked, especially in low-altitude urban airspaces. To overcome this impediment, reconfigurable intelligent surfaces (RISs) come to rescue. An RIS is a two-dimensional (2D) surface inlaid with a large number of low-cost passive elements having controllable electromagnetic responses via onboard positive-intrinsic-negative (PIN) diodes [5], [6]. As a benefit of these programmable characteristics, RISs are capable of intelligently reconfiguring the wireless propagation environment. The resultant signals can either be added constructively at the desired receiver for signal enhancement or destructively at the non-intended receivers for interference mitigation. Therefore, the quality of A2G communication links can be improved with concatenated virtual line-of-sight (LoS) links by deploying RISs in UAV-enabled wireless communication systems and intelligently configuring their reconfigurable coefficients [7]. Moreover, in contrast to conventional MIMO schemes and active relays, RISs passively reflect signals and hence they do not need RF chains, thereby significantly reducing both the hardware cost and the power consumption [8].

A. Prior Works

1) UAV communications: To fully exploit the high mobility of UAVs in the mobile-UAVs enabled wireless networks,

This work was supported by the National Key Research and Development Program of China (Grant No. 2020YFB0505600), National Natural Science Foundation of China (Grant No. U2033215, Grant No. 61822102). L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028).

(Corresponding author: Yanbo Zhu)

J. Zhao, Y. Zhu, and K. Cai are with Beihang University, Beijing, China (email:{jingjingzhao, zhuyanbo, caikq}@buaa.edu.cn).

X. Mu is with Beijing University of Posts and Telecommunications, Beijing, China (email: muxidong@bupt.edu.cn).

Y. Liu is with Queen Mary University of London, London, U.K. (email: yuanwei.liu@qmul.ac.uk).

L. Hanzo is with University of Southampton, Southampton, U.K. (email: lh@ecs.soton.ac.uk).

extensive research efforts have been devoted to studying the trajectory design problems under different UAV communications scenarios [9]–[14]. By proper trajectory design, the communication distance between the UAVs and ground nodes can be shortened and the corresponding A2G data rate can be significantly improved [15]. Currently, the approaches for solving the UAV's trajectory design problem can be categorized into two classes, namely, conventional optimization tools [9]–[12] and machine learning [13], [14]. In particular, Wang et al. [9] proposed an efficient spectrum sharing methord for an aerial UAV and terrestrial device-to-device communications by alternately optimizing the transmit power and UAV's trajectory. Mu et al. [10] studied the UAV mission completion time minimization problem by optimizing the UAV's trajectory, where the successive convex approximation was applied to obtain the locally optimal solution. Considering a practical multiuser multi-input single-output (MISO) UAV communication scenario with no-fly zones, Xu et al. [11] proposed the optimal trajectory and beamforming policy by employing monotonic optimization theory and semidefinite programming relaxation. Furthermore, Wu et al. [12] investigated the throughput maximization problem in a multi-UAV enabled downlink communication system by applying the block coordinate descent and successive convex optimization techniques. As machine learning is regarded as one of the most promising artificial intelligence tools conceived for intelligent adaptive learning in the face of uncertainties [16], the adoption of reinforcement learning (RL) for solving trajectory design problems in UAV communications has also attracted explosive attention these years. Liu et al. [13] proposed an energy-efficient multi-UAV control algorithm based on the deep deterministic policy gradient (DDPG) framework for maximizing the energy efficiency, with the joint consideration of improved communication coverage, fairness, energy consumption and connectivity. Cui et al. [14] developed a distributed multi-agent RL framework, where each UAV acted as a learning agent to automatically select its communication user, power level and specific subchannel for maximizing the long-term rewards.

2) RIS-assisted UAV communications: Motivated by the aforementioned benefits of RISs in UAV-aided wireless networks, some related works have been invested in investigating the performance gain of RISs in various UAV use cases, where the UAV acted as a base station (BS) [7], [17], [18] or as a relay node [19], [20]. Li et al. [7] optimized the joint UAV's trajectory and RISs passive beamforming design with the objective of maximizing the average achievable rate. Given the limited onboard energy of the UAV, Liu et al. [17] studied the joint UAV's trajectory, RIS configuration, and power allocation for minimizing the energy consumption. Furthermore, considering a multi-UAV scenario, Mu et al. [18] investigated the RIS-enhanced UAV-aided non-orthogonal multiple access networks, where the UAVs' deployment, the RIS configuration, and the specific detection order of users was jointly optimized. With respect to the UAV relay scenario, Yang et al. [19] derived the analytical expressions of the outage probability, average bit error rate, and average capacity of the RIS-assisted dual-hop UAV communication systems. Ranjha et al. [20] considered 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 the choice of decoding orders was studied.

B. Motivations and Contributions

Most of the existing literature considered UAV-aided communications relying on conventional reflecting-only RISs, where the incident signals can only be reflected by one side of the surface, resulting in 180° coverage. As such, both the source and destination nodes have to be at the same side of the RIS. However, this geographic constraint makes it difficult to fully reap the benefits of UAVs and RISs, because when the UAV flies to the non-reflective side of the RIS, the corresponding A2G channels cannot benefit from reflection by the RIS. Alternatively, the flying range of the UAV has to be restricted. To overcome this impediment, the concept of simultaneously transmitting and reflecting RISs (STAR-RISs) [21]–[23] is a promising solution. For STAR-RISs, the incident signals can be reflected and transmitted to the half-space at the same side and opposite side of the source node with respect to the surface, respectively, thereby facilitating full-space 360° coverage [22]. In [21] and [22], the main hardware design, physics principles, and communication system design were studied for revealing the superiority of STAR-RISs over conventional reflecting-only RISs. Moreover, the primary prototypes, which resemble STAR-RIS, have already been implemented based on metasurfaces [24]. In [25] and [26], the authors studied the beamforming optimization methods of STAR-RISs in terrestrial communications. Meanwhile, the study on the performance gain brought by STAR-RISs to UAV communications is still quite open, which motivates this work.

Against the above backdrop, our main contributions can be summarized as follows:

- We consider a STAR-RIS assisted UAV communications system, where the multi-antenna UAV acts as the aerial BS and transmits signals to multiple GUs located at both sides of a STAR-RIS. We formulate a sum-rate maximization problem by jointly optimizing the UAV's trajectory, the active beamforming at the UAV, and the passive transmission/reflection beamforming at the STAR-RIS, subject to the constraints on the UAV's flight safety, the maximum flight duration, as well as the GUs' minimum data rate requirements.
- We formulate the sum-rate maximization problem into a Markov Decision Process (MDP) based model. Given the limited sampling of the environment concerning the unknown locations of obstacles, we introduce an ambiguity set for characterizing the uncertain distribution of the agent's policy, which is inspired by the concept of distributionally robust optimization (DRO). On the basis of the proposed ambiguity set, we propose a distributionally robust RL (DRRL) algorithm for solving the formulated problem, which is computationally tractable.
- Our numerical results unveil that the STAR-RIS attains significant sum-rate gains over the conventional reflecting/transmitting-only counterparts for UAV communications. Finally, the proposed DRRL algorithm is

shown to achieve significantly higher learning efficiency and robustness than the state-of-the-art RL algorithms.

C. Organization and Notation

The rest of this paper is organized as follows. In Section II, we introduce the system model of STAR-RIS assisted UAV communications. In Section III, we formulate the sum-rate maximization problem for the joint optimization of the UAV's trajectory, the active beamforming at the UAV, and the passive transmission/reflection beamforming at the STAR-RIS. Then we formulate the problem in the context of an MDP framework. In Section IV, a novel DRRL algorithm is proposed for solving the sum-rate maximization problem formulated in the face of uncertainties. Numerical results are presented in Section V for characterizing the proposed algorithms compared to the relevant benchmarks. Finally, our conclusions are offered 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

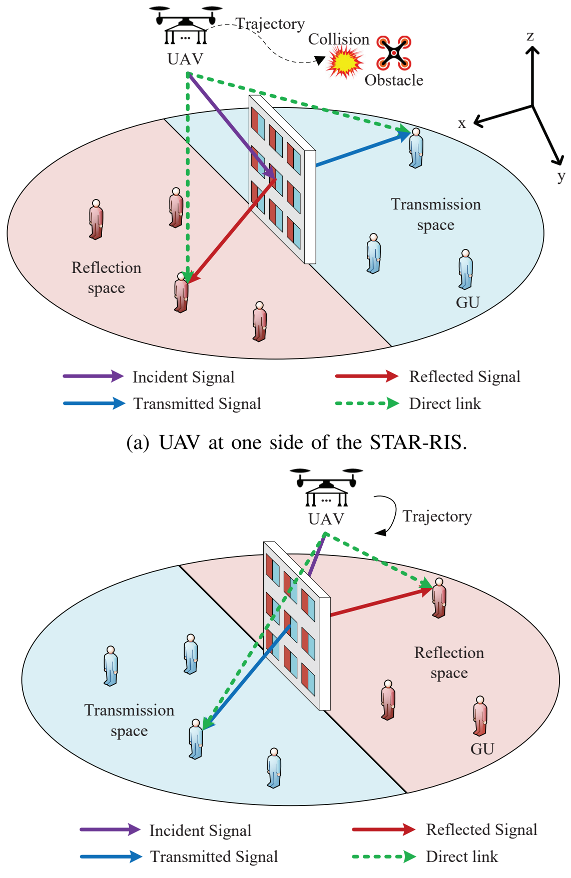
In this section, we first introduce the STAR-RIS assisted UAV communications system where a multi-antenna UAV serves multiple single-antenna GUs via the STAR-RIS, such that the A2G signal propagation can be reconfigured in the full space. The signal model of the A2G transmission is then constructed.

A. Scenario Description

Consider an A2G communication system, where a UAV equipped with K antennas acting as the BS provides wireless service to a total number of J $(K \geq J)$ single-antenna GUs denoted by $\mathcal{J} = \{1, ..., J\}$, as shown in Fig. 1. Due to the complex environment involving potential obstacles, the direct links between the UAV and GUs may not be sufficiently stable or may even be blocked. To alleviate this issue, we propose to deploy a STAR-RIS composed of N sub-wavelength elements, denoted by $\mathcal{N} = \{1, ..., N\}$, upon highrise masts/buildings to provide high-quality transmission/reflection based 360° A2G links. Let T denote the flight duration of the UAV. For tractability, T is divided into L equal non-overlapped time slots, i.e. $T = L\delta_t$. The UAV flies at a fixed altitude z_u with constant speed V and changes its heading angle at each time slot to control its flying trajectory. The trajectory of the UAV can then be denoted by $\mathbf{q}[l] = (x[l], y[l], z_u), l \in \mathcal{L} = \{1, ..., L\}$. In practice, the UAV's trajectory has to satisfy the constraints on the initial and final locations [27], [28], which are mathematically expressed as:

$$\mathbf{q}[1] = \mathbf{q}_i, \mathbf{q}[L] = \mathbf{q}_f, \tag{1}$$

where $\mathbf{q}_i$ and $\mathbf{q}_f$ denote the UAV's initial and final locations, respectively.



(a) UAV at one side of the STAR-RIS.

(b) UAV at the other side of the STAR-RIS.

Fig. 1: Illustration of STAR-RIS assisted full-space UAV communications.

Considering the dynamic urban environment where unexpected obstacles [1] constituted by the urban paraphernalia in low altitude airspace may threaten the UAV's flying safety, we have to use collision avoidance mechanisms to enable safe flight operation. The UAV has to detect its surroundings (i.e. locations of obstacles) via onboard sensors. We assume that the onboard sensors have the sensing range of R_s, which means that the UAV has no information about the obstacles unless they fall within the detection range. Assume that the minimum separation distance between the UAV and obstacles is $d_{\min}$, which satisfies $d_{\min} < R_s$. An exclusion zone is defined around the obstacle with radius $d_{\min}$, and the UAV is not allowed to fly over this zone to keep a safe distance from there. Denote any obstacle that may appear during the UAV's flight as $o_i \in \mathcal{O}$, where $\mathcal{O}$ represents the set of obstacles. To guarantee flight safety, the following constraints have to be satisfied:

$$\|\mathbf{q}[l] - \mathbf{q}_{o_i}\| \geq d_{\min}, \forall o_i \in \mathcal{O}, \tag{2}$$

where $\mathbf{q}_{o_i}$ represents the location of an unexpected obstacle o_i.

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

By supporting surface magnetic currents, both the amplitudes and phase shifts of the transmitted and reflected signals of each STAR-RIS element can be adjusted independently [22]. When a signal impinges from either side of the STAR-RIS, part of the signal is reflected and transmitted to the same side and opposite side of the impinging signal, respectively [23]. To be more specific, as shown in Fig. 1, when the UAV is located in the $x > 0$ region, then the region $x > 0$ is the *reflection space*, while the region $x < 0$ is the *transmission space*. When the UAV moves to the $x < 0$ region, the two spaces are swapped accordingly.

Remark 1. *Due to the high mobility of the UAV, the conventional reflecting-only RIS with opaque substrate cannot guarantee coverage in case the UAV moves to the opposite side of the reflecting surface. With the advent of STAR-RIS, fortunately, the full-space* 360° *A2G propagation environment can be reshaped into a desired form thanks to the omni-surface, which leads to a more flexible UAV trajectory design.*

Since the UAV-RIS-GU cascaded links suffer from substantial path loss, a large number of STAR-RIS elements are required for achieving favorable reflected/transmitted communications. However, the massive number of STAR-RIS elements result in excessive channel state information (CSI) acquisition and transmission/reflection coefficients design complexity [29]. To solve this problem, as in [30], [31], the N STAR-RIS elements are partitioned into M sub-surfaces, denoted by the set $\mathcal{M} = \{1, ..., M\}$, each consisting of $\overline{N} = N/M$ (assumed to be an integer) adjacent elements that share the same transmission/reflection coefficients for reducing the implementation complexity. We denote the transmission and reflection coefficients of the m-th sub-surface at the l-th time slot by $\theta_m^r[l] = \beta_m^r[l] e^{j\phi_m^r[l]}$, and $\theta_m^t[l] = \beta_m^t[l] e^{j\phi_m^t[l]}$, respectively, where $\beta_m^r[l]$, $\beta_m^t[l] \in [0, 1]$ and $\phi_m^r[l], \phi_m^t[l] \in [0, 2\pi)$ denote the transmission and reflection amplitude and phase shift response of the m-th sub-surface, respectively. Then, the diagonal transmission and reflection coefficient matrix of the STAR-RIS can be denoted by $\mathbf{\Theta}^r[l] = \mathrm{diag}\left(\boldsymbol{\theta}^r[l] \otimes \mathbf{1}_{\overline{N}\times 1}\right) \in \mathbb{C}^{N\times N}$, where $\boldsymbol{\theta}^r[l] = [\theta_1^r[l], ..., \theta_m^r[l], ..., \theta_M^r[l]]^T$, and $\mathbf{\Theta}^t[l] = \mathrm{diag}\left(\boldsymbol{\theta}^t[l] \otimes \mathbf{1}_{\overline{N}\times 1}\right) \in \mathbb{C}^{N\times N}$, where $\boldsymbol{\theta}^t[l] = [\theta_1^t[l], ..., \theta_m^t[l], ..., \theta_M^t[l]]^T$, respectively. We assume that the phase-shift coefficients for transmission and reflection can be independently adjusted[2] and are "b-bit controllable", where 2^b possible phase shifts can be defined. For simplicity, the discrete phase-shift values are obtained by uniformly quantizing the interval $[0, 2\pi)$. In other words, the following constraint should be satisfied: $\phi_m^\kappa[l] = \frac{\psi\pi}{2^{b-1}}, \psi \in \left\{0, 1, ..., 2^b - 1\right\}, \forall m \in \mathcal{M}, l \in \mathcal{L}$, where $\kappa \in \{r, t\}$. Complying with the energy conservation law, the sum of the energies of the reflected and transmitted signals should be no higher than the impinging signal. In this work, we assume that no energy is dissipated by STAR-RIS for simplicity, i.e., $\left(\beta_m^r[l]\right)^2 + \left(\beta_m^t[l]\right)^2 = 1, \forall m \in \mathcal{M}, l \in \mathcal{L}$. Note that each STAR-RIS sub-surface can operate in full reflection, full transmission and hybrid mode by appropriately adjusting the energy splitting ratio. Therefore, the conventinal reflecting-only RIS can be regarded as a special case of the STAR-RIS by setting all STAR-RIS elements to the full reflection mode.

B. Signal Model

The locations of the UAV and each GU determine whether the GU receives the reflected signal or the transmitted signal from the STAR-RIS. At each time slot l, we denote the set of GUs located in the transmission and reflection space as $\mathcal{J}_r[l]$ and $\mathcal{J}_t[l]$, respectively, where $|\mathcal{J}_r[l]| + |\mathcal{J}_t[l]| = J$. Let $\mathbf{G}[l] \in \mathbb{C}^{K\times J}$, $\mathbf{H}_{U,R}[l] \in \mathbb{C}^{M\times K}$, $\mathbf{h}_{R,j}^H \in \mathbb{C}^{1\times M}$, $\mathbf{h}_{U,j}^H[l] \in \mathbb{C}^{1\times K}$ represent the beamforming matrix at the UAV, channel matrix of the links from the UAV to STAR-RIS, channel matrix of the links from the STAR-RIS to GU j, and channel matrix from the UAV to GU j, respectively. Note that $\mathbf{H}_{U,R}[l]$ and $\mathbf{h}_{R,j}^H$ are modelled by Rician fading channels given the existence of the LoS component, while $\mathbf{h}_{U,j}^H[l]$ is modelled by a Rayleigh fading channel due to the blocked LoS link and potential extensive scattering. We assume that perfect CSI[3] can be obtained by employing similar techniques to those proposed in [33], [34]. Then, the received signal at user $j \in \mathcal{J}_\kappa[l], \kappa \in \{r, t\}$ at time slot l is given by

$$y_j^\kappa[l] = \mathbf{v}_j^\kappa[l]\mathbf{g}_j[l]x_j[l] + \sum_{j'\neq j}^{J} \mathbf{v}_j^\kappa[l]\mathbf{g}_{j'}[l]x_{j'}[l] + n_j, \forall l \in \mathcal{L}, \tag{3}$$

where we have $\mathbf{v}_j^\kappa[l] = \mathbf{h}_{R,j}^H \mathbf{\Theta}^\kappa[l]\mathbf{H}_{U,R}[l] + \mathbf{h}_{U,j}^H[l]$, which is the concatenated channel from the UAV to GU j, $\mathbf{g}_j[l]$ is the j-th column of $\mathbf{G}[l]$, $x_j[l]$ is the signal transmitted from the UAV to GU j with $\mathbb{E}\left[|x_j[l]|^2\right] = 1$, and $n_j \sim \mathcal{CN}(0, \sigma_j^2)$ is the additive white Gaussian noise (AWGN). Accordingly, the achievable communication rate at GU $j \in \mathcal{J}_\kappa[l]$ at time slot l is expressed as

$$R_j^\kappa[l] = \log_2\left(1 + \frac{\left|\mathbf{v}_j^\kappa[l]\mathbf{g}_j[l]\right|^2}{\sum_{j'\neq j}^{J} \left|\mathbf{v}_j^\kappa[l]\mathbf{g}_{j'}[l]\right|^2 + \sigma_j^2}\right), \forall l \in \mathcal{L}. \tag{4}$$

Therefore, the sum-rate of GUs over L time slots is given by $R_{\text{sum}} = \sum_{l=1}^{L} \sum_{j=1}^{J} R_j^\kappa[l]$.

III. Problem Formulation and Markov Decision Process Model

In this section, we formulate a sum-rate maximization problem for the joint optimization of the UAV's trajectory, the active beamforming at the UAV, and the passive transmission/reflection beamforming at the STAR-RIS, and then model the formulated problem by an MDP framework.

[2] Note that the independent phase-shift model is practically relevant for the semi-passive STAR-RIS, while the coupled phase-shift model is appropriate for the purely passive STAR-RIS [32]. However, the investigation of the coupled phase-shift model is beyond the scope of this work. The results obtained in this work provide an upper bound to the purely passive STAR-RIS associated with the coupled phase-shift model.

[3] Existing channel estimation methods proposed for the conventional reflecting-only counterparts can be applied to the STAR-RIS scenario. Due to space limitations, the detailed discussion is left for our future work.

A. Problem Formulation

As shown in (4), the achievable rate of a GU at each time slot l is determined by the location of the UAV, the active beamforming at the UAV, and the passive transmission/reflection beamforming at the STAR-RIS. To investigate the effect of STAR-RIS on UAV communications with respect to reflected and transmitted signals, our objective is to maximize the sum-rate over the UAV's flight time by jointly optimizing the UAV's trajectory $\mathbf{q}$, the UAV beamformer weights $\mathbf{G}$, and the STAR-RIS beamformer weights $\mathbf{\Theta}^{\kappa}$. In particular, the optimization problem can be formulated as

$$\max_{\mathbf{q},\mathbf{G},\mathbf{\Theta}^{\kappa}} R_{\text{sum}}, \tag{5a}$$

$$\text{s.t.} \;\; T \leq T^{\max}, \tag{5b}$$

$$R_j^{\kappa}[l] \geq R_j^{\min}, \forall j \in \mathcal{J}, \tag{5c}$$

$$\operatorname{Tr}\left(\mathbf{G}^H[l]\mathbf{G}[l]\right) \leq P_t^{\max}, \forall l \in \mathcal{L}, \tag{5d}$$

$$\phi_m^{\kappa}[l] = \frac{\psi\pi}{2^{b-1}}, \psi \in \left\{0, 1, ..., 2^b - 1\right\}, \forall m \in \mathcal{M}, l \in \mathcal{L}, \tag{5e}$$

$$\beta_m^{\kappa}[l] \in [0,1], (\beta_m^r[l])^2 + \left(\beta_m^t[l]\right)^2 = 1, \forall m \in \mathcal{M}, l \in \mathcal{L}, \tag{5f}$$

$$(1), (2). \tag{5g}$$

(5b) is the maximum flight duration constraint, which is limited by the UAV's initial onboard energy as well as by the propulsion power consumption, (5c) is the minimum data rate constraint of GUs, (5d) ensures that the UAV's transmit power should not exceed the maximal power constraint, (5e) and (5f) are the feasible ranges of STAR-RIS phase shift and amplitude coefficients, respectively.

Problem (5) is challenging to solve due to the following reasons. On the one hand, the airspace has unexpected obstacles. Every time the UAV detects an obstacle, the optimal trajectory has to be recalculated. Since the UAV's trajectory as well as the active and passive beamforming are highly coupled, the last two items should be jointly optimized from scratch as well, which is time-consuming and inefficient. On the other hand, under uncertain environments, accurate online decisions are highly dependent on the exhaustive sampling of the environment during offline training. However, due to the limited affordable sampling in practice, how to guarantee the worst-case performance and safe online deployment is another challenging problem. To tackle the above issues, we introduce a DRRL approach, where an ambiguity set is constructed endogenously to capture the learning uncertainty by integrating the partial distribution information on the statistical properties of the parameters in the decision models, to provide resilient decisions in an uncertain environment.

Remark 2. *In contrast to the conventional reflecting-only RIS assisted UAV communication [7], [17], [18], the introduction of STAR-RISs enables the* 360° *coverage, but it also imposes new challenges. On the one hand, there are two types of passive beamforming, namely transmission and reflection beamforming, to be optimized for STAR-RISs, which are generally coupled with each other due to the energy conservation law. On the other hand, as shown in Fig. 1, the transmission and reflection spaces of the STAR-RIS are determined by the location of the UAV, which is dynamically varied during the flight. In other words, the joint UAV trajectory and passive beamforming design problems using the conventional reflecting-only RIS [7], [17] can be regarded as special cases of the STAR-RIS assisted UAV communication, where the transmission function is turned off. Therefore, the formulated problem in* (5) *is more challenging and the existing algorithms proposed for the conventional reflecting-only RIS assisted UAV communication cannot be applied to solve* (5).

B. MDP Model

Before diving into the DRRL algorithm design, we first model the formulated problem by an MDP framework. Problem (5) can be designed as a sequential decision making process, where decisions at a specific time slot are based on the current situation. We define a five-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ for modelling 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 the state transition probability matrix, $\mathcal{R}$ is a real-valued reward function for the agent taking an action based on the 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 slot l with the aid of its policy. Especially, a policy π is a distribution over the actions given the states, which is formulated as $\pi(a|s) = \mathbb{P}[A_l = a \mid S_l = s]$, $\pi(a|s) \in [0,1]$. After taking 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 for maximizing 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

$$v_{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_{\pi}(s') \right), \tag{6}$$

where the discount factor $\gamma \in [0,1]$ indicates the present value of future rewards. A γ value close to 0 leads to "myopic" evaluation, while a γ value close to 1 leads to "far-sighted" evaluation. Furthermore, $\mathcal{R}_s^a$ is the reward function with $\mathcal{R}_s^a = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$, and $\mathcal{P}_{ss'}^a$ is the state transition probability matrix with $\mathcal{P}_{ss'}^a = \mathbb{P}\left[S_{l+1} = s' \mid S_l = s, A_l = a\right]$. In the MDP formulated, we assume that a central controller acting as the agent explores the unknown environment. The state, action, reward and state transition for the formulated MDP are defined in the following.

1) State: The environment state at time slot l is denoted as $S_l = \{\mathbf{q}[l], \mathcal{D}[l], R_{\text{sum}}[l-1], T^{-}[l]\}$, which is composed of four parts:

- $\mathbf{q}[l]$ is the location of the UAV at time slot l;
- $\mathcal{D}[l] = \{d_{u,o_i}[l], \forall o_i \in \mathcal{O}\}$ is the set of distances from the UAV to the center of obstacles that are within detection range at time slot l;
- $R_{\text{sum}}[l-1] = \sum_{l'=1}^{l-1} \sum_{j=1}^{J} R_j^{\kappa}[l']$ is the sum-rate of GUs from time slot 1 to $l-1$;
- $T^{-}[l]$ is the difference between the residual and minimum time required to reach the final destination at time slot l.

2) Action: The action space of the formulated MDP includes the UAV's manoeuver direction, UAV beamformer, and RIS configuration decisions at each time slot. Given the above action space, finding the optimal policy that governs the UAV's trajectory, UAV beamformer, and RIS configuration is a nontrivial challenge for the following reasons. Firstly, considering that only a limited number of discrete transmission/reflection phase shifts can be provided by each STAR-RIS element in practice, the action space of the proposed MDP model is a hybrid of discrete and continuous spaces. Secondly, the high-dimensional action space and the environment uncertainties make the MDP quite a challenge to solve due to the unknown transition probabilities and the curse of dimensionality [35]. To tackle the above two challenging issues, we first discretize the UAV's maneuver directions and the reflection amplitude coefficient of each STAR-RIS sub-surface to H and I levels, respectively, thereby transforming the action space into a discrete set. Furthermore, to maintain a small size of the action space, we bring forward the solution of low-complexity UAV beamformer matrix calculation for the given UAV's location and STAR-RIS configuration, which is detailed as follows.

Since the UAV beamforming weights over different time slots are independent, for ease of explanation, the time slot l is omitted for the UAV's beamformer weights calculation. At a specific time slot, given the UAV's location and STAR-RIS configuration, the UAV's beamforming subproblem is given by

$$\max_{\mathbf{G}} R_{\text{sum}}, \tag{7a}$$

$$\text{s.t.} \;\; R_j^{\kappa} \geq R_j^{\text{min}}, \forall j \in \mathcal{J}, \tag{7b}$$

$$\text{Tr}\left(\mathbf{G}^H \mathbf{G}\right) \leq P_t^{\text{max}}, \forall l \in \mathcal{L}. \tag{7c}$$

Proposition 1. *For the typical digital beamforming problem of* (7)*, zero-forcing (ZF)*[4] *precoding is proposed which is capable of eliminating the multi-user interference and obtain a near-optimal solution at a low complexity [36], [37]. We rewrite (3) in vectorial form as* $\mathbf{y} = \mathbf{VGx} + \mathbf{n}$*, where we have* $\mathbf{y} = [y_1^{\kappa}, ..., y_J^{\kappa}]^T$*,* $\mathbf{x} = [x_1, ..., x_J]^T$*,* $\mathbf{V}$ *is a* $J \times K$ *matrix with the* j*-th row being* $\mathbf{v}_j^{\kappa}$ *defined in Section II.B, and* $\mathbf{n} = [n_1, ..., n_J]^T$ *is the noise vector. The ZF beamformer is given by*

$$\mathbf{G} = \mathbf{V}^H \left(\mathbf{V}\mathbf{V}^H\right)^{-1} \mathbf{P}^{\frac{1}{2}} = \tilde{\mathbf{V}}\mathbf{P}^{\frac{1}{2}}, \tag{8}$$

where $\tilde{\mathbf{V}} = \mathbf{V}^H \left(\mathbf{V}\mathbf{V}^H\right)^{-1}$*, and* $\mathbf{P}$ *is a diagonal matrix with the* j*-th diagonal element being* p_j*. The expression of* p_j *is given by*

$$p_j = \frac{1}{\nu_j} \max\left\{\frac{1}{\mu} - \nu_j \sigma^2, \nu_j p_j^{min}\right\}, \tag{9}$$

where ν_j *is the* j*-th diagonal element of* $\tilde{\mathbf{V}}^H \tilde{\mathbf{V}}$*,* μ *is a normalization factor which is selected for ensuring that* $\sum\limits_{1 \leq j \leq J} \max\{\frac{1}{\mu} - \nu_j \sigma^2, \nu_j p_j^{min}\} = P_t^{max}$*, and* $p_j^{min} = \sigma^2 (2^{R_j^{min}} - 1)$ *is the minimum received power constraint of GU* j*.*

Proof. See Appendix A.

According to (8) and (9), the optimal UAV beamforming matrix can be obtained for a given UAV's location and STAR-RIS configuration. Therefore, for the proposed MDP model, we only have to include the UAV's trajectory and STAR-RIS configuration decisions into the action space and derive the optimal UAV beamformer using (8) and (9) for state-value calculation. Specifically, the action space contains the following four parts:

- The maneuver direction of the UAV, i.e. $f[l] \in \{f_1, ..., f_h, ..., f_H\}$;
- The reflection phase shift coefficient of each sub-surface, i.e. $\phi_m^r[l]$;
- The transmission phase shift coefficient of each sub-surface, i.e. $\phi_m^t[l]$;
- The reflection amplitude coefficient of each sub-surface, i.e. $\beta_m^r[l] \in \{\beta_1, ..., \beta_i, ..., \beta_I\}$.

3) Reward: As shown in (5), the objective of the UAV's trajectory and joint beamforming design is to maximize the sum-rate over the time span T. Since the reward that guides the learning should be consistent with the objective, we simply include the instantaneous sum-rate of all the GUs at each time slot, namely $\overline{R}[l] = \sum_{j=1}^{J} R_j^{\kappa}[l]$, in the reward. In response to the safe flight constraint, we set a penalty of K_1 if the distance between the UAV and any obstacle is less than the minimum separation distance. For the maximum flight duration constraint, a penalty of K_2 is given if $T^{-}[l]$ drops to 0 before arriving at the final destination. When the UAV reaches the final destination ahead of the maximum allowed time, we will grant the UAV an additional reward of K_3. Furthermore, to incorporate some prior knowledge into the reward, we include the dynamics of $T^{-}[l]$ with a weight of K_4 in the reward. As such, the reward function is then defined as

$$R_l = \begin{cases} \overline{R}[l] - K_1, \;\text{if } \|\mathbf{q}[l] - \mathbf{q}_{o_i}\| < d_{\text{min}}, \exists o_i \in \mathcal{O}, \\ \overline{R}[l] - K_2, \;\text{if } T^{-}[l] < 0, \\ \overline{R}[l] + K_3, \;\text{if } \mathbf{q}[l] = \mathbf{q}_f, T^{-}[l] \geq 0, \\ \overline{R}[l] + K_4 \left(T^{-}[l] - T^{-}[l-1]\right), \;\text{otherwise.} \end{cases} \tag{10}$$

Note that parameters K_1, K_2, K_3 and K_4 should be carefully tuned for improving both the convergence rate and the expected accumulated reward.

4) State Transition: The UAV's location $\mathbf{q}[l]$ and distance $\mathcal{D}[l]$ to obstacles are updated based on its maneuver direction at each time slot. The dynamics of $R_{\text{sum}}[l]$ can be expressed as

$$R_{\text{sum}}[l] = R_{\text{sum}}[l-1] + \sum_{j=1}^{J} R_j^{\kappa}[l], \tag{11}$$

where $R_j^{\kappa}[l]$ can be obtained with actions at time slot l based on (4).

The difference $T^{-}[l]$ between the residual and the minimum time required to reach to final destination is updated based on the UAV's movement. In particular, if the UAV moves towards the destination at time slot l, $T^{-}[l]$ remains the same as $T^{-}[l-1]$; otherwise, $T^{-}[l]$ is decreased. Specifically, the update of $T^{-}[l]$ is given by

$$T^{-}[l] = \begin{cases} T^{-}[l-1], \;\text{if } d_{u,f}[l] \leq d_{u,f}[l-1], \\ T^{-}[l-1] - 2\delta_t, \;\;\text{otherwise,} \end{cases} \tag{12}$$

[4] Here we consider ZF as a feasible solution as it is suitable for large-scale antennas thanks to the low-complexity. Other precoding methods can also be utilized and do not affect the insights obtained in this article.

where $d_{u,f}[l] = \|\mathbf{q}[l] - \mathbf{q}_f\|$ denotes the distance between the UAV and final destination at time slot l.

The terminal state of the MDP formulated includes three different types:

- The distance between the UAV and any obstacle is less than the minimum separation distance, i.e. $\|\mathbf{q}[l] - \mathbf{q}_{o_i}\| < d_{\min}, \exists o_i \in \mathcal{O}$;
- The difference between the UAV's residual and minimum time required to reach the final destination is negative, i.e. $T^{-}[l] < 0$;
- The UAV reaches the final destination before the maximum flight duration, i.e. $\mathbf{q}[l] = \mathbf{q}_f, T^{-}[l] \geq 0$.

Due to the unexpected obstacles in the proposed MDP model, it is challenging to determine the perfect state-value information. Therefore, decision-makers have to select the best policy based on partial information of the random events. In order to enhance the system's operational resilience under an uncertain environment, in the next section, we introduce the DRRL algorithm for efficiently capturing the uncertainty and optimizing the worst-case performance. Note that the locations of obstacles are randomly generated by the simulator.

IV. Sum-Rate Maximization Algorithm Design

In this section, we will first introduce a sample-efficient soft actor-critic (SAC) framework, and then develop a distributionally-robust SAC (DRSAC) algorithm to solve the UAV's trajectory and joint beamforming design problem formulated. Its convergence and optimality are also analyzed.

A. SAC Framework

The much awaited application of RL frameworks has remained slow in practice primarily due to the relatively poor sampling efficiency and brittle convergence [38]. To overcome this issue, the SAC framework [39] based on the maximum entropy philosophy was proposed to realize sample-efficient training. Compared to state-of-the-art RL algorithms, SAC was proved to have additional merits, including more efficient exploration, multi-mode near-optimal policies, and improved learning speed, especially for complex tasks [40], [41]. 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. The objective is denoted by

$$\max_{\pi} \sum_{l=1}^{L} \mathbb{E}_{(S_l,A_l)\sim\tau_\pi} \gamma^{l-1} \mathcal{R}_{S_l}^{A_l}. \tag{13}$$

In the maximum entropy framework, an entropy term is included in the objective for encouraging exploration. Specifically, the objective is expressed as follows:

$$\max_{\pi} F(\pi), \tag{14}$$

where

$$\begin{aligned} F(\pi) &= \sum_{l=1}^{L} \mathbb{E}_{(S_l,A_l)\sim\tau_\pi} \left[\gamma^{l-1} \mathcal{R}_{S_l}^{A_l} + \alpha \mathcal{H}\left(\pi\left(A_l|S_l\right)\right)\right] \\ &= \sum_{l=1}^{L} \mathbb{E}_{(S_l,A_l)\sim\tau_\pi} \left[\gamma^{l-1} \mathcal{R}_{S_l}^{A_l} - \alpha \log\left(\pi\left(A_l|S_l\right)\right)\right]. \end{aligned} \tag{15}$$

The new objective function takes into account the entropy of policy distribution, i.e. $\alpha\mathcal{H}\left(\pi\left(\cdot|S_l\right)\right)$. Here, the temperature parameter α denotes the weight of the entropy, and thus characterizes how random the optimal policy π^* is. Note that (15) is the same as (13) when α is set to 0. The optimal setting of the temperature α is closely related to different tasks as well as to the reward magnitude during training. In order to generate a flexible tuning of the entropy weight, the objective function (15) can be transformed by treating the average entropy as a constraint, which can be formulated as follows [40]:

$$\max_{\pi} \sum_{l=1}^{L} \mathbb{E}_{(S_l,A_l)\sim\tau_\pi} \left[\gamma^{l-1} \mathcal{R}_{S_l}^{A_l}\right], \tag{16a}$$

$$\text{s.t. } \mathbb{E}_{(S_l,A_l)\sim\tau_\pi} \left[-\log\left(\pi(A_l|S_l)\right)\right] \geq \mathcal{H}_{\min}, \forall l, \tag{16b}$$

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

$$\alpha_m^* = \arg\min_{\alpha_l} \mathbb{E}_{A_l\sim\pi_l^*} \left[-\alpha_l \log\left(\pi_l^*\left(A_l|S_l;\alpha_l\right)\right) - \alpha_l \mathcal{H}_{\min}\right], \tag{17}$$

where $\pi_l^*\left(A_l|S_l;\alpha_l\right)$ denotes the optimal policy corresponding to the temperature α_l. The dual gradient descent [42] is a promising solution for problem (17), where the objective is defined as

$$\mathcal{L}(\alpha) = \mathbb{E}_{A_l\sim\pi_l} \left[-\alpha \log\left(\pi_l\left(A_l|S_l\right)\right) - \alpha \mathcal{H}_{\min}\right]. \tag{18}$$

One can observe that the optimal temperature depends on the optimal policy at each time slot. 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 the policy evaluation and policy improvement phases. 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. The Bellman expectation equation can be expressed as $Q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$. In contrast to the conventional state-value function, by taking the entropy into consideration, the *soft* state-value function of the maximum entropy framework is given by

$$v_\pi\left(s\right) = \mathbb{E}_{a\sim\pi}\left[Q_\pi(s,a) - \alpha \log\left(\pi(a|s)\right)\right]. \tag{19}$$

Proposition 2. *As the state space of the proposed MDP model is continuous, neural networks can be applied for the practical approximation of the state values. With assuming that the Q-network parameter is denoted by ω, the loss function of the Q-network is given by*

$$\mathcal{L}_Q\left(\omega\right) = \mathbb{E}_{(S_l,A_l)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\omega(S_l,A_l) - \hat{Q}(S_l,A_l)\right)^2\right], \tag{20}$$

where

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

Here, $\mathcal{D}$ is the replay buffer storing the transitions $\left(S_l, A_l, \mathcal{R}_{S_l}^{A_l}, S_{l+1}\right)$ obeying the previous policies, while $\hat{\omega}$ is a parameter of the target Q-network and duplicated from ω periodically. The terms $\mathcal{R}_{S_l}^{A_l}$ and S_{l+1} in (21) are fetched from the replay buffer with given S_l and A_l.

Proof. See Appendix B.

The aim of the policy improvement phase is to improve the policy w.r.t. up-to-date Q-values obtained in the policy evaluation phase. For discrete action settings, following the Boltzmann policy, the improved policy is given by [40]

$$\pi_{\text{new}} = \frac{\exp\left(\frac{1}{\alpha} Q_{\pi_{\text{old}}}\left(S_l, \cdot\right)\right)}{\sum_a \left[\exp\left(\frac{1}{\alpha} Q_{\pi_{\text{old}}}\left(S_l, \cdot\right)\right)\right]}. \tag{22}$$

Proposition 3. *Aiming for the policy improvement principle given in (22), the loss function for the policy network is given by*

$$\mathcal{L}_\pi(\vartheta) =$$
$$\mathbb{E}_{S_l \in \mathcal{D}} \sum_{A_l \in \mathcal{A}} \pi_\vartheta(A_l|S_l)\Big(\alpha \log\left(\pi_\vartheta(A_l|S_l)\right) - Q_\omega\left(S_l, A_l\right)\Big). \tag{23}$$

Proof. See Appendix C.

Given the loss functions of the critic (i.e. Q) and actor (i.e. policy) networks, the weights ω and ϑ can be updated with the aid of their stochastic gradients. The pseudocode of SAC is shown in **Algorithm 1**. In lines 6-10, the agent interacts with the environment following the current policy to collect experience. In lines 11-17, the neural networks, i.e. the critic and actor networks, are updated by the stochastic gradients based on the previously collected experience.

Remark 3. *SAC follows the typical policy iteration framework, while the main difference between SAC and conventional RL just lies in the entropy term considered in the loss functions of the critic and actor networks. Therefore, the convergence and optimality of SAC can be analyzed similarly to that of the policy iteration. It has been proved in [35] that the policy iteration is capable of converging to the optimal policy according to the contraction mapping theorem. Therefore, the convergence and optimality of the proposed SAC algorithm can be guaranteed.*

B. DRSAC Algorithm Design

Although the SAC algorithm improves the learning efficiency by harvesting the maximum entropy framework, the catastrophic policy outcome may be obtained with inaccurate calculation of the state values. This is because the limited sampling of the environment exhibiting uncertainties will lead to estimation errors for the policy evaluation, which is also one of the main concerns for the application of RL in the real world. To enhance the worst-case performance and guarantee safe online implementation in the face of estimation errors, the DRRL algorithm can be a potential approach to enhance the robustness of the system by endogenously constructing the ambiguity set to capture the uncertainties [43].

Algorithm 1 Sum-Rate Maximization Algorithm Based on SAC

```
1: Initialize environment;
2: Initialize $\omega_i (i = 1, 2)$ for critic networks, $\vartheta$ for actor network;
3: Initialize target networks $\hat{\omega}_i \leftarrow \omega_i, i = 1, 2$;
4: Initialize entropy level $\mathcal{H}_{\text{min}}$, replay buffer $\mathcal{D} = \emptyset$, step length for gradient descent of the critic network, actor network and temperature parameter, i.e. $\lambda_Q$, $\lambda_\pi$, and $\lambda_\alpha$, respectively;
5: for each iteration do
6:     for each environment step do
7:         Execute action based on current policy $A_l \sim \pi_\vartheta$;
8:         Observe reward $\mathcal{R}_{S_l}^{A_l}$ and next state $S_{l+1}$;
9:         Store transition $(S_l, A_l, \mathcal{R}_{S_l}^{A_l}, S_{l+1})$ in $\mathcal{D}$;
10:    end for
11:    for each gradient step do
12:        Sample a random minibatch of transitions $(S_l, A_l, \mathcal{R}_{S_l}^{A_l}, S_{l+1})$ from $\mathcal{D}$;
13:        Update critic networks by minimizing loss function $\mathcal{L}_Q(\omega)$ with stochastic gradients: $\omega_i \leftarrow \omega_i - \lambda_Q \hat{\nabla}_{\omega_i} \mathcal{L}_Q(\omega_i), \forall i \in \{1, 2\}$;
14:        Update the actor network by minimizing loss function $\mathcal{L}_\pi(\vartheta)$: $\vartheta \leftarrow \vartheta - \lambda_\pi \hat{\nabla}_\vartheta \mathcal{L}_\pi(\vartheta)$;
15:        Update temperature parameter by minimizing $\mathcal{L}(\alpha)$: $\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha \mathcal{L}(\alpha)$;
16:        Update target network parameters periodically: $\hat{\omega}_i \leftarrow \omega_i, \forall i \in \{1, 2\}$;
17:    end for
18: end for
```

Recall that the policy iteration algorithm is an iterative process that alternates between the policy evaluation and policy improvement. Therefore, the policy iteration process can be formulated as:

$$\pi_{k+1} \leftarrow \mathcal{G}(v_{k+1}), v_{k+1} \leftarrow \mathcal{T}^{\pi_{k+1}} v_k, \tag{24}$$

where π_k and v_k denote the updated policy and state-value function at the k-th iteration, respectively. Furthermore, $\mathcal{T}^{\pi_{k+1}}$ represents the Bellman operator applied for policy evaluation, which can be expressed as $\mathcal{T}^\pi v(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[\mathcal{R}_s^a + \gamma \mathbb{E}_{s' \sim p(s'|s,a)} v(s')\right]$. $\mathcal{G}(v_k)$ is the greedy policy improvement approach, which is given by $\mathcal{G}(v) = \arg\max_\pi \mathcal{T}^\pi v$.

Due to the randomness of obstacles in the MDP model considered, limited sampling of the environment can lead to inaccurate estimation of state values in the policy evaluation step, and thus may result in undesired outcomes. Assume that the state-value estimation error in the k-th iteration is denoted by $\mathbf{e} \in \mathbb{R}^{\mathcal{S}}$. Since the policy improvement phase follows $\pi_{k+1} \leftarrow \mathcal{G}(v_k)$, the estimation error $\mathbf{e}$ will be reflected in the policy π_{k+1}. Let $\tilde{\mathbf{e}} \in \mathbb{R}^{\mathcal{S}}$ denote the error sequence for the policy. The objective of a conventional RL framework is given by

$$\max_\pi \sum_{l=1}^{L} \mathbb{E}_{(S_l, A_l) \sim \tau_\pi} \gamma^{l-1} \mathcal{R}_{S_l}^{A_l} = \max_\pi G(\pi). \tag{25}$$

Taking the policy error $\tilde{\mathbf{e}}$ into consideration, the new robust objective function (OF) is defined as

$$\max_\pi \min_{\tilde{\mathbf{e}}} G\left(\pi_{\tilde{\mathbf{e}}}\right). \tag{26}$$

To quantize the error associated with a specific policy, we apply the Kullback-Leibler (KL) divergence [43] to measure the differences between a pair of probability distributions over the actions in each state. More specifically, given a policy π and an error sequence $\tilde{\mathbf{e}} \in \mathbb{R}^{\mathcal{S}}$, the set of policies within the error range is formulated as

$$\mathcal{U}_{\tilde{\mathbf{e}}}\left(\pi\right) = \left\{\pi' \in \Delta_{\mathcal{A}}^{\mathcal{S}} | D_{\mathrm{KL}}\left(\pi'(\cdot|s)\left\|\pi(\cdot|s)\right.\right) \leq \tilde{\mathbf{e}}(s), \forall s \in \mathcal{S}\right\}, \tag{27}$$

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 OF in (26) can be rewritten as

$$\max_{\pi} \min_{\pi_{\tilde{\mathbf{e}}} \in \mathcal{U}_{\tilde{\mathbf{e}}}(\pi)} G\left(\pi_{\tilde{\mathbf{e}}}\right) = \max_{\pi} \min_{\pi_{\tilde{\mathbf{e}}} \in \mathcal{U}_{\tilde{\mathbf{e}}}(\pi)} \mathbb{E}_{(S_l, A_l) \sim \tau_{\pi_{\tilde{\mathbf{e}}}}} \gamma^{l-1} \mathcal{R}_{S_l}^{A_l}. \tag{28}$$

One can observe that (28) follows the typical DRO format [44], [45]. Due to the difficulty of acquiring perfect information in real world, decision-makers need to find the robust solution with partially known distribution information. DRO theory has been proven to be able to optimize the worst-case expectation cost by constructing an ambiguity set to capture the environment uncertainty distribution [44]. Compared to traditional stochastic programming and robust optimization, DRO has the advantages including non-biased estimation, statistical decision, and low-cost [45]. To solve the DRO problem under a RL framework, we have to refer again to the policy iteration process given in (24).

Definition 1. *The adversarial Bellman operator $\mathcal{T}^{\pi_{\tilde{\mathbf{e}}}^*}$ is defined as*

$$\mathcal{T}^{\pi_{\tilde{\mathbf{e}}}^*} v(s) = \min_{\tilde{\pi} \in \mathcal{U}_{\tilde{\mathbf{e}}}(\pi)} \mathcal{T}^{\tilde{\pi}} v(s). \tag{29}$$

Applying $\mathcal{T}^{\pi_{\tilde{\mathbf{e}}}^*}$ for the associated policy evaluation can provide the lower bound of state values, and thus prevent overly optimistic estimates. Therefore, the policy evaluation associated with $\mathcal{T}^{\pi_{\tilde{\mathbf{e}}}^*}$ can be termed as *distributionally robust policy evaluation.*

Proposition 4. *The policy $\pi_{\tilde{\mathbf{e}}}^*$ for distributionally robust policy evaluation is given by*

$$\pi_{\tilde{\mathbf{e}}}^* \propto \exp\left(-\frac{Q_v\left(s, a\right)}{\lambda^*(s)}\right) \pi(a|s), \tag{30}$$

where

$$\lambda^*(s) = \underset{\lambda(s)>0}{\arg\min} \left(\lambda(s)\Omega^*\left(-\frac{Q_v(s, \cdot)}{\lambda(s)}\right) + \lambda(s)\tilde{\mathbf{e}}(s)\right). \tag{31}$$

Proof. To derive the adversarial Bellman operator, we first apply Lagrangian duality to (27) and (29), and then the problem can be rewritten as (32), where $\lambda(s)$ is the Lagrange multiplier. The inner maximization problem can be expressed as

$$\begin{aligned}
&\max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left(-\mathcal{T}^{\tilde{\pi}} v(s) - \lambda(s) D_{\mathrm{KL}}\left(\tilde{\pi}(\cdot|s)\left\|\pi(\cdot|s)\right.\right)\right) \\
&= \max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \lambda(s) \left(-\frac{1}{\lambda(s)} \mathcal{T}^{\tilde{\pi}} v(s) - D_{\mathrm{KL}}\left(\tilde{\pi}(\cdot|s)\left\|\pi(\cdot|s)\right.\right)\right) \\
&= \max_{\tilde{\pi} \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \lambda(s) \left(\left\langle -\frac{Q_v(s, \cdot)}{\lambda(s)}, \tilde{\pi}(\cdot|s)\right\rangle - D_{\mathrm{KL}}\left(\tilde{\pi}(\cdot|s)\left\|\pi(\cdot|s)\right.\right)\right) \\
&= \lambda(s)\Omega^*\left(-\frac{Q_v(s, \cdot)}{\lambda(s)}\right),
\end{aligned} \tag{33}$$

where $\Omega^*\left(-\frac{Q_v(s,\cdot)}{\lambda(s)}\right)$ is the Fenchel duality [42] of $\Omega(\tilde{\pi}(\cdot|s)) = D_{\mathrm{KL}}\left[\tilde{\pi}(\cdot|s)\|\pi(\cdot|s)\right]$, which is expressed as

$$\Omega^*\left(-\frac{Q_v(s, \cdot)}{\lambda(s)}\right) = \log \mathbb{E}_{a \in \tilde{\pi}} \exp\left(-\frac{Q_v(s, a)}{\lambda(s)}\right), \tag{34}$$

The solution of the problem in (33) is given by

$$\pi_{\tilde{\mathbf{e}}}^* \propto \exp\left(-\frac{Q_v\left(s, a\right)}{\lambda^*(s)}\right) \pi(a|s). \tag{35}$$

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

$$\lambda^*(s) = \underset{\lambda(s)>0}{\arg\min} \left(\lambda(s)\Omega^*\left(-\frac{Q_v(s, \cdot)}{\lambda(s)}\right) + \lambda(s)\tilde{\mathbf{e}}(s)\right), \tag{36}$$

which is a typical convex optimization problem. The solution can be obtained by standard convex program solvers, such as CVX [46].

Lemma 1. *The construction of the policy error $\tilde{\mathbf{e}}(s)$ is in the form of $\tilde{\mathbf{e}}(s) = Cn(s)^{-\eta}$ with the constants $C > 0$ and $\eta > 0$, while $n(s)$ indicates how many times the state was visited. This construction implies that the estimation error should decrease with the amount of experience collected.*

As discussed in Section IV.A, the policy improvement strategy for SAC following the per-state entropy bonus is given by

$$\pi(a|s) \propto \exp\left(\frac{1}{\alpha} Q_v(s, a)\right), \tag{37}$$

where α is the entropy temperature. Substituting (37) into (35), the adversarial policy in SAC is obtained as

$$\pi_{\tilde{\mathbf{e}}}^* \propto \exp\left[\left(\frac{1}{\alpha} - \frac{1}{\lambda^*(s)}\right) Q_v(s, a)\right]. \tag{38}$$

Based on the adversarial Bellman operator $\mathcal{T}^{\pi_{\tilde{\mathbf{e}}}^*} v(s)$, the term $\hat{Q}\left(S_l, A_l\right)$ in the loss function (46) of the critic network should be rewritten as

$$\begin{aligned}
&\hat{Q}(S_l, A_l) = \mathcal{R}_{S_l}^{A_l} + \\
&\gamma \mathbb{E}_{A_{l+1} \sim \pi_{\tilde{\mathbf{e}}}^*}\left[Q_{\hat{\omega}}(S_{l+1}, A_{l+1}) - \alpha \log\left(\pi(A_{l+1}|S_{l+1})\right)\right].
\end{aligned} \tag{39}$$

The details of the DRSAC algorithm developed are summarized in **Algorithm 2**. The main difference between DRSAC and SAC lies in the policy evaluation phase. Instead of updating the critic network towards the true action-value function following the current policy π, the adversarial policy $\pi_{\tilde{\mathbf{e}}}^*$ is adopted for action-value estimation (lines 11 - 14) for providing a lower-bound performance guarantee.

Remark 4. *We design the policy error $\tilde{\mathbf{e}}(s)$ in the way that it gets smaller with accumulated experience. This is reflected in the construction of $\tilde{\mathbf{e}}(s)$ in the form $\tilde{\mathbf{e}}(s) = Cn(s)^{-\eta}$ with $C > 0$ and $\eta > 0$. More intuitively, we can say that the radius of the uncertainty set shrinks as the learning process proceeds. Consequently, when $m \rightarrow +\infty$, the adversarial policy $\pi_{\tilde{\mathbf{e}}}^*$ converges to the policy π. Although the algorithm performs conservatively in a short-term, it acts optimistically in a long*

$$\begin{aligned}\mathcal{T}^{\pi^*_{\tilde{\mathbf{e}}}} v(s) &= \max_{\lambda(s)>0} \min_{\tilde{\pi} \in \Delta^{\mathcal{S}}_{\mathcal{A}}} \Big(\mathcal{T}^{\tilde{\pi}} v(s) + \lambda(s) D_{\text{KL}} \left(\tilde{\pi}(\cdot|s) \, \| \pi(\cdot|s) \right) - \lambda(s) \tilde{\mathbf{e}}(s) \Big) \\ &= \min_{\lambda(s)>0} \max_{\tilde{\pi} \in \Delta^{\mathcal{S}}_{\mathcal{A}}} \Big(-\mathcal{T}^{\tilde{\pi}} v(s) - \lambda(s) D_{\text{KL}} \left(\tilde{\pi}(\cdot|s) \, \| \pi(\cdot|s) \right) + \lambda(s) \tilde{\mathbf{e}}(s) \Big)\end{aligned} \tag{32}$$

Algorithm 2 Sum-Rate Maximization Algorithm Based on DRSAC

```
Initialize environment;
Initialize critic network, actor network, replay buffer $\mathcal{D} = \emptyset$;
Set $\mathcal{H}_{\min}$, $C$, $\eta$, $n(s) = 0, \forall s \in \mathcal{S}$;
for each iteration do
    for each environment step do
        Execute action based on the current policy;
        Store transition $(S_l, A_l, \mathcal{R}_{S_l}^{A_l}, S_{l+1})$ in $\mathcal{D}$;
    end for
    for each gradient step do
        Sample a random minibatch of transitions $(S_l, A_l, \mathcal{R}_{S_l}^{A_l}, S_{l+1})$ from $\mathcal{D}$;
        $\tilde{\mathbf{e}}(s) \leftarrow Cn(s)^{-\eta}$;
        Solve convex optimization problem (36) to obtain $\lambda^*(s)$;
        Obtain $\pi^*_{\tilde{\mathbf{e}}}(a|s)$ via (38);
        Obtain $\hat{Q}(S_l, A_l)$ via (39);
        Update actor network, critic network, and $\alpha$ as in Algorithm 1;
    end for
end for
return optimal policy $\pi^*$.
```

run, hence the convergence and optimality of DRSAC can then be guaranteed, similarly to that of SAC.

V. Numerical Results

In this section, we evaluate the performance of our proposed algorithm for STAR-RIS assisted UAV communications in terms of the sum-rate. We show how the system performance is influenced by the number of UAV antennas and the number of STAR-RIS elements. For comparison, the following benchmark algorithms are used:

- *SAC*: We utilize Algorithm 1 to solve the sum-rate maximization problem, which serves as a benchmark to show that the DRSAC algorithm achieves robust performance in the face of estimation errors caused by limited sampling of the environment.
- *Deep Q-network (DQN)*: We adopt the conventional DQN algorithm using the reward defined in (10) without considering the entropy. The initial exploration probability, the minimum exploration probability, and the learning rate are set to 0.9, 0.05, and 0.00001, respectively.
- *Reflecting/Transmitting-only RIS case*: This case serves as a benchmark for demonstrating the merits of the STAR-RIS, where a reflecting-only RIS and a transmitting-only RIS are deployed adjacent to each other, each of which consists of $N/2$ elements. The resultant optimization problem is solved by Algorithm 2 while setting the reflection amplitude coefficients to 1 for the first $N/2$ elements, and 0 for the latter $N/2$ elements.
- *STAR-RIS with equal energy splitting*: In this case, the reflection and transmission amplitude coefficients of each STAR-RIS sub-surface are both set to $\sqrt{0.5}$. The resultant optimization problem is solved by Algorithm 2.
- *Random phase shift*: Algorithm 2 is performed for the joint optimization of the UAV's trajectory, the active beamforming at the UAV, and the transmission/reflection amplitude coefficients of the STAR-RIS sub-surfaces, with the transmission/reflection phase shifts of the STAR-RIS sub-surfaces generated randomly.
- *UAV communications w/o STAR-RIS*: In this case, the STAR-RIS is not deployed, which means that the communication links between the UAV and GUs only include the direct links. The resultant joint UAV's trajectory and beamforming design problem is solved by Algorithm 2.

A. Simulation Setup

In the simulation, the narrow-band quasi-static fading channels spanning from the UAV to STAR-RIS and from the STAR-RIS to GUs are modeled as Rician fading links as follows:

$$\mathbf{H}_{U,R}[l] = \sqrt{\frac{\beta_0}{(d_{U,R}[l])^{\alpha_1}}} \left(\sqrt{\frac{\kappa_{U,R}}{1+\kappa_{U,R}}} \mathbf{H}_{U,R}^{\text{LoS}}[l] + \sqrt{\frac{1}{1+\kappa_{U,R}}} \mathbf{H}_{U,R}^{\text{NLoS}}[l] \right), \tag{40a}$$

$$\mathbf{h}_{R,j}^{H}[l] = \sqrt{\frac{\beta_0}{(d_{R,j}[l])^{\alpha_1}}} \left(\sqrt{\frac{\kappa_{R,j}}{1+\kappa_{R,j}}} \mathbf{h}_{R,j}^{\text{LoS}}[l] + \sqrt{\frac{1}{1+\kappa_{R,j}}} \mathbf{h}_{R,j}^{\text{NLoS}}[l] \right), \tag{40b}$$

where β_0 is the path loss at the reference distance of $d_0 = 1$ m, α_1 is the corresponding path loss exponent, $d_{U,R}$ and $d_{R,j}$ denote the distance from the UAV to STAR-RIS and that from the STAR-RIS to GU j, respectively, $\kappa_{U,R}$ and $\kappa_{R,j}$ represent the Rician factor, $\mathbf{H}_{U,R}^{\text{LoS}}$ and $\mathbf{h}_{R,j}^{\text{LoS}}$ denote the LoS components, and $\mathbf{H}_{U,R}^{\text{NLoS}}$ and $\mathbf{h}_{R,j}^{\text{NLoS}}$ denote the non line-of-sight (NLoS) components modeled by Rayleigh fading.

The channel spanning from the UAV to GU j is modeled as the Rayleigh fading link of $\mathbf{h}_{U,j}^{H}[l] = \sqrt{\beta_0 \left(d_{U,j}[l]\right)^{-\alpha_2}} \hat{\mathbf{h}}_{U,j}[l]$, where $d_{U,j}$ denotes the distance between the UAV and GU j, α_2 is the path loss exponent of the Rayleigh fading channel, and $\hat{\mathbf{h}}_{U,j}$ denotes the small-scale fading component, where the elements are independently drawn from the circularly symmetric complex Gaussian (CSCG) distribution with unit variance.

In our simulation, we set the initial and final locations of the UAV to $(-30, -15, 50)$ and $(30, -15, 50)$ meters, respectively. The STAR-RIS is located at $(0, 0, 20)$ meters. We consider 2 GUs in the network, each having a random location generated at one side of the STAR-RIS with the distance of 10 m. The number of quantization bits for discrete phase shift is set to be 1 . The UAV's maneuver direction is discretized into 4 actions, i.e. left, right, forward, and backward movement. The reflection amplitude coefficient is discretized into 2 levels, i.e. $\sqrt{0.3}$ and $\sqrt{0.7}$. We consider an obstacle appearing in

TABLE I: System Parameters

α_1	Path loss parameter for LoS transmissions	2.2	d_{min}	Minimum separation distance	18 m
α_2	Path loss exponent for NLoS transmissions	4	λ_Q, λ_π, λ_α	Gradient descent step length	0.00001
κ	Rician factor	10 dB	C	Constant	1
β_0	Path loss at 1 m	-20 dB	η	Constant	0.5
σ_b^2	Noise power	-90 dBm		Number of training steps	1000000
δ_t	Duration of each time slot	1 s		Replay memory size	20000
$\overline{N}$	Number of elements in each sub-surface	10		Mini-batch size	128
P_t^{max}	UAV's maximum transmit power	30 dBm		Optimizer	Adam
T^{max}	Maximum flight duration	8.5 s		Activation function	ReLU

the airspace with a random location. Other specific system parameters and the DRL's hyperparameters are summarized in Table I, unless otherwise specified.

B. Performance of the proposed algorithm

In Fig. 2, we compare the performance of DRSAC against SAC and DQN. The curves are obtained for five runs of each algorithm with different random seeds. Fig. 2 shows the system sum-rate during training, where the solid curves denote the averaged values and the shaded regions represent the lower and upper bounds over the five trials. It can be observed that DRSAC and SAC outperform DQN in terms of learning efficiency, which can be explained by the sampling-efficiency boosting scheme by considering the entropy term in the SAC framework. Moreover, the sum-rate lower-bound achieved by DRSAC is shown to be higher than that of SAC, which indicates that DRSAC improves the worst-case performance. This is consistent with our proof in Section IV showing that DRSAC is robust to estimation errors.

Fig. 3 demonstrates the performance of the proposed DRSAC algorithm under the training and implementations stages. Particularly, Fig. 3(a) portrays the probability of the UAV reaching its destination safely with DRSAC during the training process under different seeds. Observe that the success probability emerges from 0, which means that the UAV cannot avoid any obstacle or find the right path to reach its destination at the beginning of the learning process. However, upon increasing the number of training episodes, the DRSAC algorithm gathers experience and optimizes the policy, thereby improving the success probability. After about 400 evaluation steps, the success probability approaches 100%. Observe from Fig. 3(a) that the proposed DRSAC algorithm learns effectively and converges under different seeds, which illustrates the robustness of the DRSAC algorithm to the choice of its hyperparameters.

Fig. 3(b) depicts the UAV's trajectory obtained by DRSAC under the implementation stage. It is observed that the UAV exploits its mobility to adaptively adjust its trajectory to move closer to the STAR-RIS for stronger communication links at the beginning. Meanwhile, the safety distance from the UAV to obstacles is maintained autonomously throughout onboard sensing. To meet the maximum allowed flight time, the UAV also seeks to fly towards the final location, while keeping its distance to the STAR-RIS as low as possible. It is observed that the UAV can move flexibily at both sides ($x > 0$ and $x < 0$ regions) of the STAR-RIS thanks to its full-space coverage. Observe from Fig. 3(b) that, after offline training, the DRSAC algorithm enables the UAV to make smart decisions at the online implementation stage.

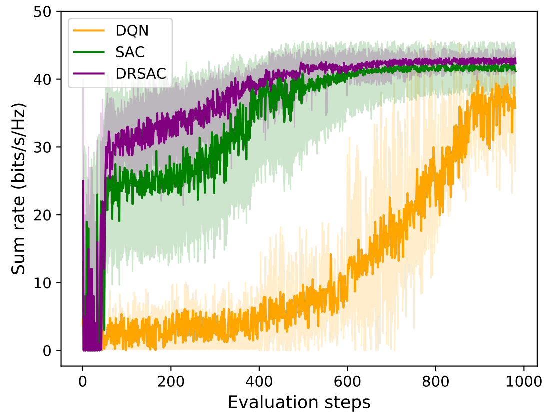
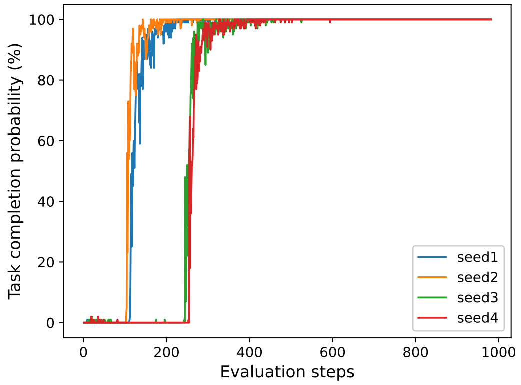


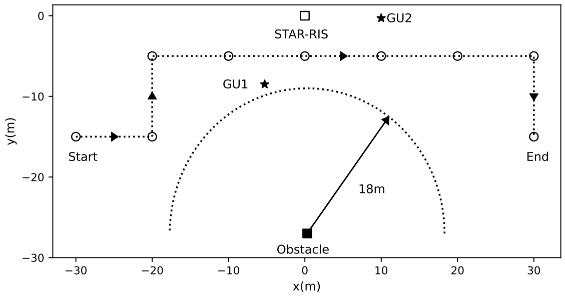
Fig. 2: Performance comparison among DQN, SAC and DRSAC.

C. Comparison with conventional reflecting/transmitting-only RIS

In Fig. 4, we investigate the performance gain of STAR-RIS in UAV communications. We set $N = 40$ and obtain the sum-rate with different numbers of UAV antennas. As seen from Fig. 4, the sum-rate increases with K, which is expected due to the enhanced beamforming gain at the UAV side for more antennas. Furthermore, one can also observe from Fig. 4 that the STAR-RIS outperforms the conventional reflecting/transmitting-only RIS by a large margin. This is because compared to the conventional reflecting/transmitting-only RIS, the STAR-RIS benefits from added flexibility by co-designing the phase shift and amplitude coefficients of both the transmission and reflection. Besides, the proposed STAR-RIS scheme also achieves a valuable performance improvement over the case, where all elements adopt the equal-energy splitting ratio. This is because the latter term is a special case of the proposed STAR-RIS scheme, which fails to offer optimized performance guarantee. However, we also add that the equal-energy splitting is easier to implement in practice [23].



(a) Task completion probability of DRSAC under thetraining stage.



(b) The UAV's trajectory obtained by DRSAC under the implementation stage. The discretized UAV's locations sampled every 1 s are marked with '○'.

Fig. 3: Performance of DRSAC under the training and implementation stages.

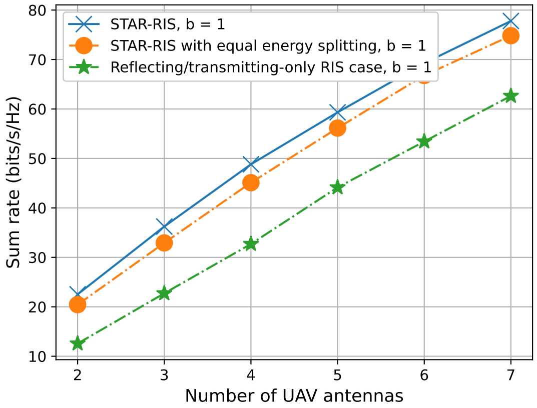
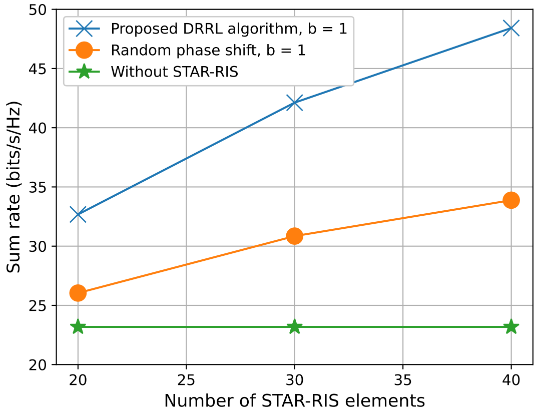


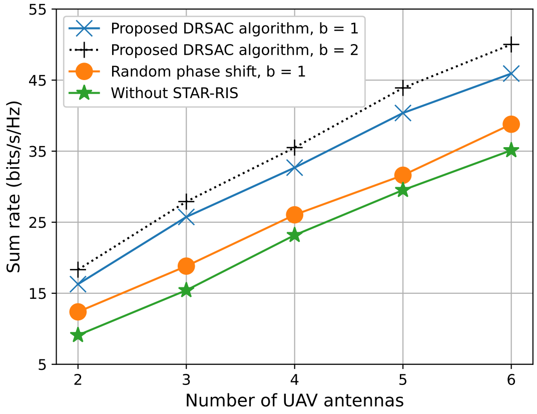
Fig. 4: Sum-rate versus K with $N = 40$.

D. Illustration on the impact of STAR-RIS phase shift configuration

In Fig. 5, we study the impact of STAR-RIS phase shift configurations. In Fig. 5(a), we set $K = 4$ and observe that the sum-rate increases with the number of STAR-RIS elements. This is because the larger number of STAR-RIS elements leads to higher beamformer design flexibility, thereby enhancing the desired signals and mitigating the multi-user interference more efficiently. It can also be observed that the proposed DRSAC algorithm achieves much higher sum-rate than both the random phase shift design and the case where no STAR-RIS is deployed, which shows the effectiveness of our proposed algorithm for STAR-RIS phase shift configuration. Moreover, the 1-bit phase shift is also shown to achieve significantly higher sum-rate than the no STAR-RIS case. This demonstrates the effectivenss of the STAR-RIS for signal enhancement even with low-cost phase shifters.



(a) Sum-rate versus N with $K = 4$.



(b) Sum-rate versus K with $N = 20$.

Fig. 5: Illustration on the impact of STAR-RIS phase shift configuration.

Fig. 5(b) depicts the sum-rate versus the number of UAV antennas under $N = 20$. Similarly, we can observe that the proposed DRSAC algorithm achieves higher sum-rate than the two benchmarks. We also compare the system sum-rate of 1-bit and 2-bit phase quantization. It is observed that the sum rate of using 2-bit phase shifters is higher than that of the 1-bit case. This is expected, since due to having less quantization levels, aligning the multi-path signals becomes less accurate, thus resulting in performance loss.

VI. Conclusions

STAR-RIS assisted UAV communications have been proposed. In contrast to conventional reflecting-only RISs, STAR-RISs facilitate simultaneous transmission and reflection of the incident signals. The achieved full-space coverage is particularly suitable for high-mobility UAV communications in the face of random obstacles constituted by the urban paraphernalia. Considering the tight coupling among the UAV's trajectory, active beamforming at the UAV, and passive transmission/reflection beamforming at the STAR-RIS, a sum-rate maximization problem has been formulated, subject to the constraints on the UAV's flight safety, the maximum flight duration, as well as the GUs' minimum data rate requirements. To solve the resultant online decision making problem in the face of uncertainties caused by the limited sampling of the environment, we have proposed a novel DRRL algorithm to satisfy a certain worst-case performance. Numerical results have been provided for demonstrating the proposed algorithm compared to traditional DRL schemes in terms of its learning efficiency and robustness. The obtained results have showed that the STAR-RIS achieved significant sum-rate gain over its conventional reflecting/transmitting-only counterparts. Moreover, it has been revealed that the UAV's trajectory become more flexible thanks to the full-space coverage of the STAR-RIS. For future work, the deployment of the STAR-RIS should be discussed in more detail, as the location and orientation of the STAR-RIS may affect the quality of the transmission/reflection links. Moreover, considering the interference caused by the UAV to ground cellular networks, the employment of the STAR-RIS for simultaneous signal enhancement between the UAV and its serving GUs as well as the interference mitigation between the UAV and nearby ground cells is expected to be an interesting research topic.

Appendix A
Proof of Proposition 1

For ZF, the transmit precoding (TPC) matrix should satisfy the following constraints:

$$\left|\mathbf{v}_j^{\kappa}\mathbf{g}_j\right| = \sqrt{p_j}, \tag{41}$$

$$\left|\mathbf{V}_{j'}^{\kappa}\mathbf{g}_j\right| = 0, \forall j' \neq j, \tag{42}$$

where p_j is the power received by GU j. According to the results given in [37], the ZF beamformer is given by

$$\mathbf{G} = \mathbf{V}^H\left(\mathbf{V}\mathbf{V}^H\right)^{-1}\mathbf{P}^{\frac{1}{2}} = \tilde{\mathbf{V}}\mathbf{P}^{\frac{1}{2}}, \tag{43}$$

Based on the constraints (41) and (42), the optimization problem (7) can be rewritten as

$$\max_{\{p_j\}} \sum_{j=1}^{J} \log_2\left(1+\frac{p_j}{\sigma^2}\right), \tag{44a}$$

$$s.t. \quad \log_2\left(1+\frac{p_j}{\sigma^2}\right) \geq R_j^{\min}, \forall j, \tag{44b}$$

$$\operatorname{Tr}\left(\mathbf{P}^{\frac{1}{2}}\tilde{\mathbf{V}}^H\tilde{\mathbf{V}}\mathbf{P}^{\frac{1}{2}}\right) \leq P_t^{\max}. \tag{44c}$$

By applying the classic water-filling algorithm, the closed-form optimal solution of problem (44) can be obtained as

$$p_j = \frac{1}{\nu_j}\max\left\{\frac{1}{\mu} - \nu_j\sigma^2, \nu_j p_j^{\min}\right\}, \tag{45}$$

where ν_j is the j-th diagonal element of $\tilde{\mathbf{V}}^H\tilde{\mathbf{V}}$, μ is a normalization factor which is selected for ensuring that $\sum\limits_{1\leq j\leq J}\max\{\frac{1}{\mu}-\nu_j\sigma^2, \nu_j p_j^{\min}\} = P_t^{\max}$, and $p_j^{\min} = \sigma^2(2^{R_j^{\min}} - 1)$ is the minimum received power constraint of GU j.

Appendix B
Proof of Proposition 2

Since the objective of the Q-network is to approximate the true state values, the loss function should be designed for guiding the output towards (19). Therefore, the loss function is given by the soft Bellman residual as follows [39]:

$$\mathcal{L}_Q\left(\omega\right) = \mathbb{E}_{(S_m,A_m)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\omega(S_m,A_m) - \hat{Q}(S_m,A_m)\right)^2\right], \tag{46}$$

where $\hat{Q}(S_m,A_m)$ is given by (47). In contrast to the SAC framework of [39], which is aimed for continuous action settings, the action space in this work is discrete. Therefore, (47) becomes tractable by employing discrete action probabilities for the expectation calculation. Specifically, (47) can be rewritten as (48).

Appendix C
Proof of Proposition 3

Based on the policy improvement principle given in (22), the loss function of the policy network is defined as (49), where D_{KL} represents the Kullback-Leibler (KL) divergence used for quantifying the similarity of a pair of distributions. The loss function (49) is defined for guiding the network output to be updated towards the improved policy (22). Since $X_{\pi_\vartheta}\left(S_m\right)$ depends only on the state, the loss function can be reduced to (multiplied by α)

$$\begin{aligned}&\mathcal{L}_\pi\left(\vartheta\right) = \\ &\mathbb{E}_{S_m\in\mathcal{D}}\mathbb{E}_{A_m\sim\pi_\vartheta}\left(\alpha\log\left[\pi_\vartheta(A_m|S_m)\right] - Q_\omega\left(S_m,A_m\right)\right).\end{aligned} \tag{50}$$

For discrete action spaces, the expectation over the actions in (50) can be calculated based on the specific action probabilities. Therefore, (50) can be rewritten as (51).

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] Z. Xiao, L. Zhu, Y. Liu, P. Yi, R. Zhang, X.-G. Xia, and R. Schober, "A survey on millimeter-wave beamforming enabled UAV communications and networking," *IEEE Commun. Surveys Tuts., Early Access*, 2021.

[3] Z. Han, A. L. Swindlehurst, and K. 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.

[4] 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, Sept. 2020.

[5] 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. 1, pp. 1–15, Apr. 2019.

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

[7] S. Li, B. Duo, X. Yuan, Y. 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, May 2020.

[8] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

$$\begin{aligned}\hat{Q}(S_m, A_m) &= \mathcal{R}_{S_m}^{A_m} + \gamma V_{\hat{\omega}}\left(S_{m+1}\right)\\ &= \mathcal{R}_{S_m}^{A_m} + \gamma \mathbb{E}_{A_{m+1}\sim\pi}\left[Q_{\hat{\omega}}\left(S_{m+1}, A_{m+1}\right) - \alpha \log\left(\pi\left(A_{m+1}|S_{m+1}\right)\right))\right]\end{aligned} \tag{47}$$

$$\hat{Q}(S_m, A_m) = \mathcal{R}_{S_m}^{A_m} + \gamma \sum_{A_{m+1}\in\mathcal{A}} \pi\left(A_{m+1}|S_{m+1}\right)\Big[Q_{\hat{\omega}}\left(S_{m+1}, A_{m+1}\right) - \alpha \log\left(\pi\left(A_{m+1}|S_{m+1}\right)\right))\Big] \tag{48}$$

$$\begin{aligned}\mathcal{L}_{\pi}\left(\vartheta\right) &= \mathbb{E}_{S_m\in\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi_{\vartheta}\left(\cdot|S_m\right)\middle\|\frac{\exp\left(\frac{1}{\alpha}Q_{\omega}\left(S_m,\cdot\right)\right)}{\alpha X_{\pi_{\vartheta}}\left(S_m\right)}\right)\right]\\ &= \mathbb{E}_{S_m\in\mathcal{D}}\left[\int_{A_m} \pi_{\vartheta}\left(A_m|S_m\right) \times \left(\log\left(\pi_{\vartheta}(A_m|S_m)\right) - \frac{1}{\alpha}Q_{\omega}\left(S_m,\cdot\right) + \log X_{\pi_{\vartheta}}(S_m)\right)\right]\\ &= \mathbb{E}_{S_m\in\mathcal{D}}\mathbb{E}_{A_m\sim\pi_{\vartheta}}\left(\log\left(\pi_{\vartheta}(A_m|S_m)\right) - \frac{1}{\alpha}Q_{\omega}\left(S_m, A_m\right) + \log X_{\pi_{\vartheta}}(S_m)\right)\end{aligned} \tag{49}$$

$$\mathcal{L}_{\pi}\left(\vartheta\right) = \mathbb{E}_{S_m\in\mathcal{D}} \sum_{A_m\in\mathcal{A}} \pi_{\vartheta}(A_m|S_m)\Big(\alpha \log\left[\pi_{\vartheta}(A_m|S_m)\right] - Q_{\omega}\left(S_m, A_m\right)\Big) \tag{51}$$

[9] 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.

[10] 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.

[11] D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Multiuser MISO UAV communications in uncertain environments with no-fly zones: Robust trajectory and resource allocation design," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3153–3172, May 2020.

[12] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[13] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[14] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

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

[16] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[17] 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.

[18] 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, Jun. 2021.

[19] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. D. Renzo, "On the performance of RIS-assisted dual-hop UAV communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 385–10 390, Jun. 2020.

[20] 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.

[21] J. Xu, Y. Liu, X. Mu, and O. A. Dobre., "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 3134–3138, May 2021.

[22] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, and H. V. Poor, "STAR: Simultaneous transmission and reflection for 360 coverage by intelligent surfaces," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 102–109, Dec. 2021.

[23] H. Zhang and et al., "Intelligent omni-surfaces for full-dimensional wireless communications: Principles, technology, and implementation," *IEEE Commun. Mag.*, vol. 60, no. 2, pp. 39–45, Feb. 2022.

[24] NTT DOCOMO, "Docomo conducts worlds first successful trial of transparent dynamic metasurface," *[Online]. Available:www.nttdocomo.co.jp/english/info/mediacenter/pr/2020/011700.html.*

[25] S. Zhang and et al., "Intelligent omni-surface: Ubiquitous wireless transmission by reflective-transmissive metasurfaces," *IEEE Trans. Wireless Commun., Early Access*, 2021.

[26] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, Early Access, doi: 10.1109/TWC.2021.3118225.

[27] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[28] S. Eom, H. Lee, J. Park, and I. Lee, "UAV-aided wireless communication designs with propulsion energy limitations," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 651–662, Jan. 2020.

[29] 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.

[30] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2020.

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

[32] Y. Liu, X. Mu, R. Schober, and H. V. Poor, "Simultaneously transmitting and reflecting (STAR)-RISs: A coupled phase-shift model," in *Proc. of the IEEE Int. Conf. on Commun. (ICC)*, 2022, to be published.

[33] Z. 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.

[34] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 4659–4663.

[35] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski

et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[36] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[37] C. Peel, B. Hochwald, and A. Swindlehurst, “A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part I: channel inversion and regularization,” *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Feb. 2005.

[38] Y. Duan, X. Chen, X. Houthooft, R. Schulman, and P. Abbbeel, “Benchmarking deep reinforcement learning for continuous control,” in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, USA, Jul. 2016.

[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 *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

[40] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, “Soft actor-critic algorithms and applications,” *arXiv preprint arXiv:1812.05905*, 2018.

[41] Y. S. Nasir and D. Guo, “Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

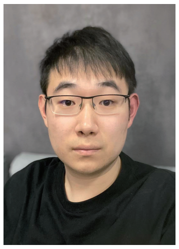
[44] E. Delage and Y. Ye, “Distributionally robust optimization under moment uncertainty with application to data-driven problems,” *Operations research*, vol. 58, no. 3, pp. 595–612, Jan. 2010.

[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.

[46] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” *[Online] Available: http://cvxr.com/cvx*, 2014.

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 in Amazon, London, U.K. Currently, she is an associate professor with the Research Institute for Frontier Science, Beihang University, Beijing, China. Her current research interests include non-orthogonal multiple access, reconfigurable intelligent surfaces, aeronautical broadband communications, and machine learning.

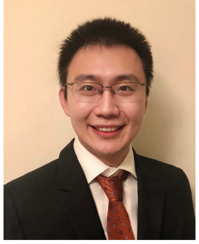
Yanbo Zhu (Member, IEEE) received the B.S. and Ph.D. degrees from the Beihang University, in 1995 and 2009, respectively. He is currently the vice president of the Aviation Data Communication Corporation, China. He is also a part-time professor with the School of Electronic and Information Engineering,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



Xidong Mu (Graduate Student Member, IEEE) received the B.S. degree in information engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree. From May 2021, he is a visiting student with the the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K.

His research interests include non-orthogonal multiple access, IRSs/RISs aided communications, UAV communications, and optimization theory. He received the Exemplary Reviewer Certificate of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2020.

Kaiquan Cai (Member, IEEE) received the B.S. and Ph.D. degrees from the 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.



Yuanwei Liu (Senior Member, IEEE, http://www.eecs.qmul.ac.uk/ yuanwei) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications in 2011 and 2014, respectively, and the PhD degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, Kings College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Senior Lecturer (Associate Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since Aug. 2021, where he was a Lecturer (Assistant Professor) from 2017 to 2021. His research interests include non-orthogonal multiple access, 5G/6G networks, RIS, integrated sensing and communications, and machine learning.

Yuanwei Liu is a Web of Science Highly Cited Researcher 2021. He is currently a Senior Editor of IEEE COMMUNICATIONS LETTERS, an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He serves as the leading Guest Editor for IEEE JSAC special issue on Next Generation Multiple Access, a Guest Editor for IEEE JSTSP special issue on Signal Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks. He received IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020. He received the 2020 IEEE Signal Processing and Computing for Communications (SPCC) Technical Early Achievement Award, IEEE Communication Theory Technical Committee (CTTC) 2021 Early Achievement Award. He received IEEE ComSoc Young Professional Outstanding Nominee Award in 2021. He has served as the Publicity Co-Chair for VTC 2019-Fall. He is the leading contributor for “Best Readings for Non-Orthogonal Multiple Access (NOMA)” and the primary contributor for “Best Readings for Reconfigurable Intelligent Surfaces (RIS)”. He serves as the chair of Special Interest Group (SIG) in SPCC Technical Committee on the topic of signal processing Techniques for next generation multiple access (NGMA), the vice-chair of SIG Wireless Communications Technical Committee (WTC) on the topic of Reconfigurable Intelligent Surfaces for Smart Radio Environments (RISE), and the Tutorials and Invited Presentations Officer for Reconfigurable Intelligent Surfaces Emerging Technology Initiative.

Lajos Hanzo (Life Fellow, IEEE, http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo) received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published 2000+ contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 PhD students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is also a Fellow of the Royal Academy of Engineering (FREng), of the IET and of EURASIP. He is the recipient of the 2022 Eric Sumner Field Award.