Distributional Reinforcement Learning for mmWave Communications with Intelligent Reflectors on a UAV

Qianqian Zhang[1], Walid Saad[1], and Mehdi Bennis[2]

[1] Bradley Department of Electrical and Computer Engineering, Virginia Tech, VA, USA, Emails: }qqz93,walids| @vt.edu.

[2]Centre for Wireless Communications, University of Oulu, Finland, Email: mehdi.bennis@oulu.fi.

Abstract—In this paper, a novel communication framework that uses an unmanned aerial vehicle (UAV)-carried intelligent reflector (IR) is proposed to enhance multi-user downlink transmissions over millimeter wave (mmWave) frequencies. In order to maximize the downlink sum-rate, the optimal precoding matrix (at the base station) and reflection coefficient (at the IR) are jointly derived. Next, to address the uncertainty of mmWave channels and maintain line-of-sight links in a real-time manner, a distributional reinforcement learning approach, based on quantile regression optimization, is proposed to learn the propagation environment of mmWave communications, and, then, optimize the location of the UAV-IR so as to maximize the long-term downlink communication capacity. Simulation results show that the proposed learning-based deployment of the UAV-IR yields a significant advantage, compared to a non-learning UAV-IR, a static IR, and a direct transmission schemes, in terms of the average data rate and the achievable line-of-sight probability of downlink mmWave communications.

I. INTRODUCTION

Millimeter wave (mmWave) frequencies are an essential component of next-generation cellular systems, in order to meet the exponential increase of data demand and support a growing number of wireless devices [1]. Due to a large available bandwidth, mmWave frequencies have a strong potential to provide high communication rates. However, the small wavelength of mmWave spectrum yields the high susceptibility to blockage caused by common objects, such as buildings and foliage, which seriously attenuate mmWave propagation [2].

In order to bypass obstacles and prolong the communication range, signal reflectors have been considered as an energy-efficient solution for mmWave communications. Signal reflectors can establish line-of-sight (LOS) links in face of blockage, by replacing a non-line-of-sight (NLOS) mmWave channel by multiple connected LOS links. As a passive element, a signal reflector costs no energy and incurs no additional receiving noise [3]. A reflector-aided transmission enables the connected LOS links to share the same frequency band, and, thus, yields a high spectrum efficiency. Meanwhile, by aligning a large number of low-cost reflective components and jointly inducing phase shifts to incident signals, an intelligent reflector (IR) can realize beamforming with very lower energy cost for mmWave communications [4]. Moreover, to maintain LOS links in a mobile scenario, an IR can be equipped onto an unmanned aerial vehicle (UAV) platform [3], so that the location of the IR can be adjusted intelligently, based on the real-time communication environment, in order to improve the

reliability of mmWave transmissions. Compared with a UAV-aided relay station, a UAV-carried IR (UAV-IR) has a simpler antenna structure and smaller power cost. Therefore, an IR facilitate multiple-input-multiple-output transmissions for the UAV platform, which has very limited onboard energy.

The use of IRs to improve the communication performance of mmWave has been studied in [3]–[7]. In [3], we studied the problem of using a UAV-IR to maximize downlink transmissions towards a single mobile user. The authors in [4] investigated the IR-aided mmWave communications, with a deep reinforcement learning (RL) framework, to improve the coverage and increase the data rate in an indoor network. In [5] and [6], the authors proposed a hybrid precoding scheme to jointly design the transmit precoding at the base station (BS) and the reflection parameters at the IR. However, the prior works in [3]–[6] focus on the downlink transmission to a single user, and they do not consider the more challenging problem of multi-user communications. Meanwhile, the works in [7] and [8] developed a static IR-assisted mmWave communication framework for multiple downlink users, such that the weighted sum-rate is maximized. However, most of the previous works in [4]–[8] optimize the reflection transmission of the IR while assuming a fixed location. Although a static IR establishes LOS links between transmitters and receivers in the face of blockage, maintaining LOS channels is challenging, especially in a mobile communication scenario, where the movement of each user dynamically changes the channel state. In particular, given the high susceptibility of mmWave signals to human body, a single body rotation can block the LOS link and render the reflection transmission inefficient, using a static IR.

Machine learning techniques were proposed in in [4] and [9]–[11] in order to address the wireless channel dynamics and improve the performance of IR-aided communications. For instance, in [9], we studied the problem of optimizing beamforming transmissions and reconfigurable reflection of an IR to serve multiple users, using a distributional RL method, so as to maximize the downlink sum-rate. The authors in [10] proposed a deep learning approach to optimize the IR phase shift, given a limited number of active reflective elements. The work in [11] introduced an RL-based scheme to jointly optimize the precoding transmission and signal reflection. However, all the prior art in [4] and [9]–[11] does not consider a mobile IR whose location can be dynamically optimized so as to enhance the mmWave reflection performance.

The main contribution of this paper is, thus, a novel downlink framework using a UAV-IR to assist a mmWave BS for multi-user communications. First, the precoding matrix at the
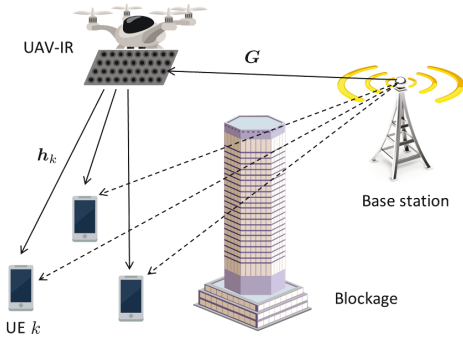
Fig. 1: A UAV-IR establishes LOS links between the BS and UEs for efficient downlink transmissions over mmWave frequencies.

BS and reflection coefficient at the IR are jointly optimized to maximize the downlink sum-rate towards multiple users. Next, to address the uncertainty of mmWave channels and maintain LOS links in a real-time manner, a distributional reinforcement learning approach [12], based on quantile regression optimization, is introduced to learn the mmWave communication environment, such that the location of the UAV-IR is optimized to maximize the communication sum-rate over a long-term horizon. Simulation results show that the proposed learning-based deployment of the UAV-IR yields a significant advantage, compared to a non-learning UAV-IR, a static IR and a direct transmission schemes, in terms of the average data rate and the achievable downlink LOS probability.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. The optimal deployment of the UAV-IR for multi-user downlink transmissions is proposed in Section III. Simulation results are shown in Section IV. Conclusions are drawn in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a wireless BS serving a set $\mathcal{L}$ of K cellular user equipment (UE) over the downlink via mmWave frequencies. The BS has M transmit antennas, and each downlink UE has a single antenna. As shown in Fig. 1, we assume that the direct transmission channel between the BS and each UE is blocked, and the received signal via the direct link is negligible. In order to bypass the obstacle and improve the received power at each UE, a UAV-IR with N reflective elements is deployed to assist downlink transmissions towards NLOS UEs. By leveraging the mobility of the UAV, the UAV-IR can potentially replace each direct NLOS link with two connected LOS links, by adjusting its position and reflecting mmWave signals from the BS towards each UE. Since an IR is a passive device, it cannot acquire the channel state information (CSI) or process received signals. To enable information exchange, an active antenna is embedded onto the UAV to receive a control signal from the BS and feedback information from served UEs.

A. Communications Capacity and Power Cost

Consider a multi-user multiple-input-single-output downlink communications, in which the BS serves downlink UEs via a common mmWave band. The BS-IR channel is denoted as $G \in \mathbb{C}^{N*M}$, and the IR-UE link for each UE k is given as

$h_k \in \mathbb{C}^{1*N}$. Let $\mathcal{O}$ be the set of N IR components. For each component $n \in \mathcal{O}$, the phase shift is denoted by $\phi_n \in [0, 2\pi)$ and the amplitude attenuation is $\beta_n \in [0, 1]$. Consequently, the IR's reflection coefficient is $\Phi = \mathrm{diag}(\beta_1 e^{j\phi_1}, \times\!\times\!\times, \beta_N e^{j\phi_N})$. In order to provide downlink communications to multiple UEs, the BS precodes the transmit signal as an $M \leq 1$ vector $\boldsymbol{a} = \sum_{k=1}^{K} \boldsymbol{w}_k s_k$, where $\boldsymbol{w}_k \in \mathbb{C}^{M*1}$ is the precoding vector, and s_k is the unit-power information symbol for UE k. The received signal at UE k will thus be: $y_k = h_k \Phi G \boldsymbol{a} + u_k$, where $u_k \leftarrow \mathcal{KO}(0, \sigma^2)$ is the receiver noise at UE k. In order to separate the mmWave propagation environment with the IR's reflection, we rewrite the received signal at UE k equivalently in the following form:

$$y_k = \boldsymbol{\theta} \boldsymbol{D}_k \boldsymbol{a} + u_k, \tag{1}$$

where $\boldsymbol{\theta} = [\beta_1 e^{j\phi_1}, \times\!\times\!\times, \beta_N e^{j\phi_N}] \in \mathbb{C}^{1*N}$ is a vector of the UAV-IR reflection coefficients and $\boldsymbol{D}_k = \mathrm{diag}(\boldsymbol{h_k})G \in \mathbb{C}^{N*M}$ is the CSI of the connected BS-IR-UE link towards UE k without any phase shift. The channel measurement approach for IR-aided cellular communications has been investigated in [9]. Here, we assume that the CSI $\boldsymbol{D}_k(\boldsymbol{x})$ of each BS-IR-UE link only depends on the location $\boldsymbol{x}$ of the UAV-IR, and the channel matrix $\boldsymbol{D}_k(\boldsymbol{x})$ is known to both BS and UAV-IR, as long as $\boldsymbol{x}$ is given. Therefore, the signal-to-interference-and-noise-ratio (SINR) of downlink communications from the BS, reflected by the IR, to each UE k is

$$\eta_k(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{x}) = \frac{\|\boldsymbol{\theta} \boldsymbol{D}_k(\boldsymbol{x})\boldsymbol{w}_k\|^2}{\sum_{i \neq k, i/\{} \|\boldsymbol{\theta} \boldsymbol{D}_k(\boldsymbol{x})\boldsymbol{w}_i\|^2 + \sigma^2}, \tag{2}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \times\!\times\!\times, \boldsymbol{w}_K] \in \mathbb{C}^{M*K}$ is a precoding matrix at the BS. Consequently, the total achievable rate that the IR-assisted communication can provide to all UEs is

$$C(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{x}) = \sum_{k=1}^{K} b \log_2(1 + \eta_k(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{x})), \tag{3}$$

where b is the downlink bandwidth.

In order to maintain LOS links with both the BS and UEs, the UAV-IR needs to frequently adjust its location, based on the real-time CSI. Thus, the power cost of a UAV-IR includes the UAV's hovering power p_h, mobility power p_m, and the adjustment power p_r for the IR's reflection coefficients. In order to facilitate beamforming transmissions at the BS and ensure a reliable reflection at the IR, we assume that downlink communication only happens when the UAV-IR has a fixed location. Let v be the speed of the UAV-IR, and $\mathbb{1}_{v=0}$ is an indicator function which equals to one when the UAV is in the hovering state with a zero speed. Therefore, the power cost of the UAV-IR for providing downlink reflection service is $p(v) = \mathbb{1}_{v=0} \times (p_h + p_r) + (1 \quad \mathbb{1}_{v=0}) \times p_m$.

B. Problem Formulation

As shown in Fig. 2, in order to provide efficient and reliable mmWave communications to downlink UEs, a dynamic deployment for the UAV-IR is considered, where the deployment process is divided into two sequential and alternating stages: communication and movement. In the communication stage,
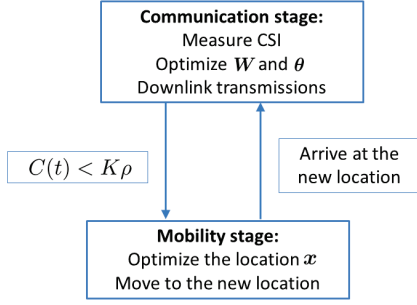
Fig. 2: The two-stage deployment process of the UAV-IR.

the UAV-IR stays at a fixed location while reflecting mmWave signals towards downlink UEs. However, once the average downlink rate is lower than a threshold ρ, blockage occurs in the downlink channels for most UEs. In this case, the current communication stage ends, and the UAV-IR moves to a new position, so as to establish LOS links for efficient downlink communications. Here, we assume that the CSI $\boldsymbol{D}_k$ is constant within each coherence time slot ΔT. Given that the duration of a communication stage depends on the real-time CSI and the length of a movement stage is determined by the movement distance and the UAV's speed, both stages can last for several coherence time slots.

Our goal is to jointly optimize the precoding matrix at the BS, the reflection coefficient of the IR, and the location of the UAV, such that, before the UAV's onboard energy is exhausted, the total achievable data transmissions that the UAV-IR provides to downlink UEs can be maximized, i.e.:

$$\max_{\|\boldsymbol{W}(t),\boldsymbol{\theta}(t),\boldsymbol{x}(t)\|\forall t} \quad \sum_{t=1}^{T} \mathbb{1}_{v(t)=0} \times C\left(\boldsymbol{W}(t),\boldsymbol{\theta}(t),\boldsymbol{x}(t)\right) \times \Delta T \tag{4a}$$

$$\text{s. t.} \quad \sum_{k=1}^{K} \left\|\boldsymbol{w}_k(t)\right\|^2 \geq P_{\max}, \tag{4b}$$

$$\|\boldsymbol{\theta}_n(t)\| \geq 1, \quad \mathcal{C}n \,\forall\, \mathcal{O}, \tag{4c}$$

$$\sum_{t=1}^{T} p(v(t)) \times \Delta T \geq E, \tag{4d}$$

where the random integer $T \,\forall\, \mathbb{N}^+$ denotes the final time slot of the UAV's service, $v(t) = \frac{\boldsymbol{x}(t-1) - \boldsymbol{x}(t)}{\Delta T}$ is the average speed of the UAV during the coherence time t, $P_{\max}$ is the maximal transmit power at the BS, and E is the initial onboard energy of the UAV. Therefore, the objective function (4a) is the summation of downlink data transmissions that the UAV-IR provides to all UEs before the end of its service, (4b) is the power limitation at the BS, (4c) is the reflection constraints at the IR, and (4d) is the energy constraint of the UAV.

The optimization problem in (4) is challenging to solve for two reasons. First, during each coherence time, the objective function (4a) is non-convex with respect to the optimal variables $\boldsymbol{W}$, $\boldsymbol{\theta}$ and $\boldsymbol{x}$. Second, the relationship between the location $\boldsymbol{x}$ of the UAV-IR with the CSI $\boldsymbol{D}_k(\boldsymbol{x})$ is not explicitly known. Meanwhile, it is impractical to apply a sweeping

search by moving the UAV-IR to all possible locations and measuring the real-time CSI $\boldsymbol{D}_k$ for each UE k. In order to address these challenges, in Section III-A, the beamforming matrix $\boldsymbol{W}$ and reflection coefficient $\boldsymbol{\theta}$ will be optimized, given a fixed location and known CSI. Next, the location optimization of the UAV-IR will be addressed using a learning-based approach to model the mmWave communication environment.

III. Optimal Deployment of UAV-IR

In this section, the precoding matrix at the BS, the reflection coefficients at the IR, and the location of the UAV-IR will be jointly optimized in a real-time manner to maximize the downlink transmission capacity. In particular, a distributional reinforcement learning (DRL) framework is proposed to model the downlink CSI during each communication stage, based on the UEs' feedback.

A. Optimal precoding and reflection coefficients

First, we focus on the communication stage, where the UAV-IR has a fixed location $\boldsymbol{x}$ and the downlink CSI $\boldsymbol{D}_k$ for each UE k is known. Then, for each coherence time, (4) is reduced to optimize the precoding matrix and the reflection coefficients, so as to maximize the downlink sum-rate, i.e.:

$$\max_{\boldsymbol{W},\boldsymbol{\theta}} \quad C\left(\boldsymbol{W},\boldsymbol{\theta}\right) \tag{5a}$$

$$\text{s. t.} \quad \sum_{k=1}^{K} \left\|\boldsymbol{w}_k\right\|^2 \geq P_{\max}, \tag{5b}$$

$$\|\boldsymbol{\theta}_n\| \geq 1, \quad \mathcal{C}n \,\forall\, \mathcal{O}. \tag{5c}$$

To solve this non-convex problem, we apply the Lagrangian dual transform method [13], introduce an auxiliary variable $\boldsymbol{\alpha} \,\forall\, \mathbb{R}^{K*1}$, and equivalently rewrite the objective function (5a) into the following form:

$$C_\alpha(\boldsymbol{W},\boldsymbol{\theta},\boldsymbol{\alpha}) = b\sum_{k=1}^{K} \left(\log_2(1+\alpha_k) - \alpha_k + \frac{(1+\alpha_k)\eta_k}{1+\eta_k} \right), \tag{6}$$

By holding $\boldsymbol{W}$ and $\boldsymbol{\theta}$ fixed and setting $\frac{\partial C_\alpha}{\partial \alpha_k} = 0$, we have the optimal value of α_k as $\alpha_k^o = \eta_k$. Then, for a fixed $\boldsymbol{\alpha}$, the optimization problem of $\boldsymbol{W}$ and $\boldsymbol{\theta}$ is reduced to

$$\max_{\boldsymbol{W},\boldsymbol{\theta}} \quad \sum_{k=1}^{K} \frac{\hat{\alpha}_k \eta_k}{1+\eta_k} \tag{7a}$$

$$\text{s. t.} \quad (5b), (5c), \tag{7b}$$

where $\hat{\alpha}_k = b(1+\alpha_k)$. Given that (7) is a multiple-ratio fractional programming problem, we can fix the value of $\boldsymbol{W}$ and $\boldsymbol{\theta}$ alternatively, and solve the optimization problem via an iterative approach, detailed as follows.

1) *Optimal precoding matrix:* For a fixed $\boldsymbol{\theta}$, the optimal precoding problem becomes

$$\max_{\boldsymbol{W}} \quad f(\boldsymbol{W}) = \sum_{k=1}^{K} \hat{\alpha}_k \frac{\|\boldsymbol{\theta}\boldsymbol{D}_k\boldsymbol{w}_k\|^2}{\sum_{i=1}^{K} \|\boldsymbol{\theta}\boldsymbol{D}_k\boldsymbol{w}_i\|^2 + \sigma^2} \tag{8a}$$

$$\text{s. t.} \quad \sum_{k=1}^{K} \left\|\boldsymbol{w}_k\right\|^2 \geq P_{\max}. \tag{8b}$$

The multiple-ratio fractional programming function in (8a) is equivalent to $f_\lambda(\boldsymbol{W}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \|\lambda_k\|^2 (\sum_{i=1}^{K} \|\boldsymbol{\theta D}_k \boldsymbol{w}_i\|^2 + \sigma^2) + \sum_{k=1}^{K} 2 \overline{\hat{\alpha}_k} \text{Re}\} \lambda_k \boldsymbol{\theta D}_k \boldsymbol{w}_k |$, where $\boldsymbol{\lambda} \forall \mathbb{R}^{K*1}$ is an auxiliary vector [13, Theorem 2]. Since f_λ is a convex function with respect to both $\boldsymbol{w}_k$ and λ_k, $\mathcal{C}k$, an iterative approach can be applied to optimize $\boldsymbol{W}$ and $\boldsymbol{\lambda}$ alternatively. First, we fix $\boldsymbol{W}$ and set $\frac{\partial f_\lambda}{\partial \lambda_k} = 0$. Then, the optimal value of λ_k is

$$\lambda_k^o = \frac{\overline{\hat{\alpha}_k} \boldsymbol{\theta D}_k \boldsymbol{w}_k}{\sum_{i=1}^{K} \|\boldsymbol{\theta D}_k \boldsymbol{w}_i\|^2 + \sigma^2}. \tag{9}$$

Then, by fixing $\boldsymbol{\lambda}$, the optimal $\boldsymbol{w}_k$ can be given by

$$\boldsymbol{w}_k^o = \sqrt{\hat{\alpha}_k} \lambda_k \boldsymbol{\theta D}_k \left(\kappa_o \boldsymbol{I}_M + \sum_{i=1}^{K} \|\lambda_i\|^2 (\boldsymbol{\theta D}_i)(\boldsymbol{\theta D}_i)^H \right)^{1}, \tag{10}$$

where $\kappa_o \sim 0$ is the minimum value such that $\sum_{k=1}^{K} \boldsymbol{w}_k^o{}^2 \geq P_{\max}$ holds. By alternating between (9) and (10), the precoding matrix will eventually converge to the unique and optimal value $\boldsymbol{W}^\in$, with a computational complexity of $\{(M^4)$.

2) Optimal reflection coefficients: Next, we fix the value of $\boldsymbol{W}$ and optimize the reflection coefficient $\boldsymbol{\theta}$ in (7), i.e.:

$$\max_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) = \sum_{k=1}^{K} \hat{\alpha}_k \frac{\|\boldsymbol{\theta D}_k \boldsymbol{w}_k\|^2}{\sum_{i=1}^{K} \|\boldsymbol{\theta D}_k \boldsymbol{w}_i\|^2 + \sigma^2} \tag{11a}$$

$$\text{s. t.} \quad \|\theta_n\| \geq 1, \quad \mathcal{C}n \forall \mathcal{O}. \tag{11b}$$

Similarly, an auxiliary vector $\boldsymbol{\delta}$ is introduced, so that (11a) equivalently becomes $f_\delta(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{k=1}^{K} 2 \overline{\hat{\alpha}_k} \text{Re}\} \delta_k \boldsymbol{\theta D}_k \boldsymbol{w}_k | \sum_{k=1}^{K} \|\delta_k\|^2 (\sum_{i=1}^{K} \|\boldsymbol{\theta D}_k \boldsymbol{w}_i\|^2 + \sigma^2)$, which is convex with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$. Therefore, the unique and optimal value of the reflection coefficients $\boldsymbol{\theta}^\in$ can be obtained via a similar alternative approach as the precoding optimization. First, the optimal δ_k for a given $\boldsymbol{\theta}$ is

$$\delta_k^o = \frac{\overline{\hat{\alpha}_k} \boldsymbol{\theta D}_k \boldsymbol{w}_k}{\sum_{i=1}^{K} \|\boldsymbol{\theta D}_k \boldsymbol{w}_i\|^2 + \sigma^2}. \tag{12}$$

Then, the optimization of the reflection coefficient for a fixed δ_k becomes

$$\max_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) = \boldsymbol{\theta U \theta}^H + 2\text{Re}\} \boldsymbol{\theta v} | \quad C \tag{13a}$$

$$\text{s. t.} \quad \|\theta_n\| \geq 1, \quad \mathcal{C}n \forall \mathcal{O}, \tag{13b}$$

where $C = \sum_{k=1}^{K} \|\delta_k\|^2 \sigma^2$, $\boldsymbol{v} = \sum_{k=1}^{K} \delta_k{}^H \boldsymbol{D}_k \boldsymbol{w}_k$, and $\boldsymbol{U} = \sum_{k=1}^{K} \|\delta_k\|^2 \sum_{i \neq k} \boldsymbol{D}_k \boldsymbol{w}_i (\boldsymbol{D}_k \boldsymbol{w}_i)^H$. Given that $\boldsymbol{U}$ is a positive-definite matrix, $f(\boldsymbol{\theta})$ is quadratic concave with respect to $\boldsymbol{\theta}$. Meanwhile, the constraint (13b) is a convex set. Thus, (13) is solvable using a Lagrange dual decomposition [8] with a computational complexity of $\{(N^6)$.

Therefore, given the CSI $\boldsymbol{D}_k$ of each BS-IR-UE link, the precoding matrix $\boldsymbol{W}^\in$ and reflection coefficient $\boldsymbol{\theta}^\in$ can be optimally and uniquely determined. Next, the optimal location of the UAV-IR will be studied to guarantee a LOS channel for each BS-IR-UE link.

B. Optimal location of the UAV-IR

During the communication stage, at the end of each coherence time t, the UAV-IR can get feedback from each UE about the downlink transmission performance. Whenever the average rate per UE is smaller than ρ, i,e, $C(t) < K\rho$, blockage occurs in the downlink channels for most UEs. In this case, the UAV-IR needs to move and rebuild LOS links. The optimization problem of the UAV-IR's location is given as

$$\max_{\} \boldsymbol{x}(t)\|_{\forall t}} \quad \sum_{t=1}^{T} \mathbb{1}_{v(t)=0} \times C(\boldsymbol{x}(t)) \times \Delta T \tag{14a}$$

$$\text{s. t.} \quad \sum_{t=1}^{T} p(v(t)) \times \Delta T \geq E. \tag{14b}$$

To determine the optimal location for each communication stage, a DRL framework is designed to learn the dynamic communication environment and model the relationship between the UAV's location and downlink CSI. In our DRL framework, the downlink CSI $\}\boldsymbol{D}_k | \mathcal{K}k$ is the communication *environment*, the UAV-IR is the *agent* that takes *action* $\Delta \boldsymbol{x}$ to change its location from $\boldsymbol{x}$ to $\boldsymbol{x} + \Delta \boldsymbol{x}$, and the communication *state* $\boldsymbol{e}$ is a vector of the received signal power at each UE, where $\boldsymbol{e} = [\|y_1\|^2, \times\times, \|y_K\|^2]$. At the end of each coherence time, the UAV-IR receives *reward* $r(\boldsymbol{x}) = \mathbb{1}_{v=0} C(\boldsymbol{x}) \Delta T$, which is the total received data in the downlink transmission. Due to the small-scale fading of mmWave channels, the downlink CSI $\boldsymbol{D}_k(\boldsymbol{x})$ may vary between different time slots, even for a fixed location $\boldsymbol{x}$ of the UAV-IR. Thus, it is more suitable to consider the reward r as a random variable with respect to $\boldsymbol{x}$, rather than a determined value. Meanwhile, $P(\boldsymbol{e}' | \boldsymbol{e}, \Delta \boldsymbol{x})$ is the state transition probability from $\boldsymbol{e}$ to $\boldsymbol{e}'$ after taking action $\Delta \boldsymbol{x}$.

To properly capture the relationship between the UAV's movement and downlink transmission performance, first, a *policy* $\pi(\Delta \boldsymbol{x} | \boldsymbol{e})$ is introduced to define the probability that the UAV-IR will move by $\Delta \boldsymbol{x}$, under a current state $\boldsymbol{e}$. Meanwhile, to quantify the potential of each action $\Delta \boldsymbol{x}$ for improving the downlink rate under a state $\boldsymbol{e}$, we define the return function of each state-action pair for any time slot t as

$$Z^\pi(\boldsymbol{e}_t, \Delta \boldsymbol{x}_t) = \sum_{i=t}^{T} \gamma^{i} {}^{t} r(\boldsymbol{e}_i, \Delta \boldsymbol{x}_i), \tag{15}$$

where $\Delta \boldsymbol{x}_i \leftarrow \pi(*|\boldsymbol{e}_i)$, $\boldsymbol{e}_{i+1} \leftarrow P(*|\boldsymbol{e}_i, \Delta \boldsymbol{x}_i)$ and $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \Delta \boldsymbol{x}_i$. Here, $\gamma \forall (0, 1)$ discounts the future rewards in the current estimation for each state-action pair. If $\gamma \in 1$, the return function Z^π will approximate (14a). Thus, the return function (15) defines a cumulative discounted reward that the UAV-IR can achieve by reflecting mmWave signals at location $\boldsymbol{x}_t + \Delta \boldsymbol{x}_t$ for the next communication stage. Meanwhile, given that r is a random variable, it is necessary to model a distribution function of (15) to identify the return value for each state-action pair. Once the return distribution is known, the optimal policy π that maximizes the expectation of the cumulative rewards can be defined by $\Delta \boldsymbol{x}^\in = \arg \max_{\Delta \boldsymbol{x}} \mathbb{E}(Z^\pi(\boldsymbol{e}_t, \Delta \boldsymbol{x}_t))$. Thus, the optimal location of the UAV-IR for the next time slot will be $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \Delta \boldsymbol{x}^\in$.

In order to model the return distribution for each state-action pair, a quantile regression (QR) method [12] is applied. A Q-quantile model Z_Q approximates the target distribution Z^π, using a discrete function with variable locations of Q supports and fixed quantile of $\frac{1}{Q}$ probabilities [9], for a fixed integer $Q \; \forall \; \mathbb{N}^+$. Mathematically, a Q-quantile model is denoted by $Z_Q(e, \Delta x) = [z_1(e, \Delta x), \times\times\times, z_Q(e, \Delta x)]$, with a cumulative probability $F_{Z_Q}(z_q) = \frac{q}{Q}$ for $q = 1, \times\times\times, Q$. The objective is to find the optimal location for each support, such that the "distance" between the target distribution Z^π and the Q-quantile model Z_Q can be minimized. However, given that the actual return distribution Z^π is not explicitly known, an empirical distribution Z will be formed, based on the UEs' feedback during each time slot, and Z is used as the target distribution to model the return approximation Z_Q. To quantify the "distance" between two distribution functions, the quantile regression loss is defined as [12],

$$\mathcal{N}_Z(Z_Q) = \sum_{q=1}^{Q} \mathbb{E}_Z \left[\|\omega_q \quad \mathbb{1}_{z<z_q}\| \times (z \quad z_q)^2 \right], \qquad (16)$$

where $\omega_q = \frac{2q \; 1}{2Q}$, $\|\omega_q \quad \mathbb{1}_{z<z_q}\|$ is the weight of regression loss penalty, and $(z \quad z_q)^2$ is the square of approximation error. Thus, the problem of the return distribution modeling becomes to minimize the quantile regression loss, i.e.,

$$\min_{z_1, \times\times\times, z_Q} \mathcal{N}_Z(Z_Q). \qquad (17)$$

Since the objective function in (17) is convex with respect to Z_Q, the minimizer $\}z_q^{\in}|_{q=1, \times\times\times, Q}$ can be found by conventional gradient-descent approaches with a computational complexity of $\{ (Q^2)$. As a result, for each state-action pair, its return distribution $Z_Q(e, \Delta x)$ can be approximated by a Q-quantile $\}z_1^{\in}(e, \Delta x), \times\times\times, z_Q^{\in}(e, \Delta x)|$ via (17).

Consequently, in the location optimization problem of the UAV-IR, after observing a communication state e_t, the UAV-IR can estimate the expected return value for each action Δx, by computing the marginal distribution of $Z_Q(e, \Delta x)$, and choose the optimal location $x_{t+1} = x_t + \Delta x^{\in}$ that maximize the summation of future downlink transmissions via

$$\Delta x^{\in} = \arg\max_{\Delta x} \mathbb{E}\left[Z_Q(e_t, \Delta x)\right] = \arg\max_{\Delta x} \frac{1}{Q}\sum_{q=1}^{Q} z_q(e_t, \Delta x). \qquad (18)$$

After the UAV-IR arrives at the new location x_{t+1} and provides downlink service to UEs, a new state e_{t+1} and reward r_{t+1} will be updated at the end of the $l+1$ time slot. Given the downlink transmission, the empirical distribution can be updated via a Q-learning approach, where $z_i(e_t, \Delta x_t) \rightarrow r_{t+1} + \gamma z_i^t(e_{t+1}, \Delta x_{t+1}), \mathcal{C} i = 1, \times\times\times, Q$. As a result, the return distribution Z_Q^t is updated by minimize the distance from the target distribution Z, based on (17). The training and update algorithm of the DRL model for the real-time optimal deployment of the UAV-IR is summarized in Algorithm 1. The convergence property of this iterative algorithm has been proved in [9, Theorem 1].

Algorithm 1 DRL-optimized deployment for the UAV-IR

Initialize the precoding matrix W_0, the reflection coefficient θ_0, the location x_0, the onboard energy E_0, and the DRL model function $Z_Q^0(e, \Delta x)$ for each state-action pair.
For $t = 1, \ggg, T$:
  A. If $v_t > 0$, the UAV-IR continues its movement;
  B .If $v_t = 0$, the UAV-IR is in the communication stage;
  Repeat:
    B1. Update the auxiliary variable $\alpha_k = \eta_k$;
    B2. Optimize W_t^* by alternating between (9) and (10);
    B3. Optimize θ_t^* , by alternating between (12) and (13);
  Until the value of C_α in (6) converges;
    B4. Receive the reward r_t and the state e_t. If $r_t \leftarrow K\rho\Delta T$, $x_{t+1} = x_t$; Otherwise, the UAV-IR moves by
      $\Delta x^* = \arg\max_{\Delta x} \frac{1}{Q}\sum_{q=1}^{Q} z_q^{t-1}(e_t, \Delta x)$.
    B5. Update the empirical distribution Z via
      $z_i(e_{t-1}, \Delta x_{t-1}) \; \forall \; r_t + \gamma z_i^{t-1}(e_t, \Delta x_t), \{i = 1, \ggg, Q$.
    B6. Update the DRL model Z_Q^t via,
      $\arg\min_{\{z_q\}_{\forall q}} \sum_{q=1}^{Q}\sum_{i=1}^{Q} \frac{2q-1}{Q} \quad \mathbb{1}_{z_i<z_q} \; \trianglelefteq (z_i \quad z_q)^2$.
  C. Update onboard energy via $E_{t+1} = E_t \quad p(v_t)\Delta T$.
Until $E_t = 0$.

IV. SIMULATION RESULTS AND ANALYSIS

For our simulations, we consider a uniform square array of antennas at both the BS and the UAV-IR with $M = 16$ and $N = 16$, and the number of downlink UEs is $K = 4$. The BS is located at $(0, 0, 20)$, each UE's location follows an i.i.d. two-dimensional Gaussian $\mathcal{O}((20, 0), 8I_2)$ with zero height, and the mobility pattern of each UE follows a Markov decision process in [3]. A building located at $(10, 0, 0)$ with a height of 18 meters permanently blocks direct BS-UE links. The path loss and mmWave channel model are based on [9]. For communication parameters, we set $f = 30$ GHz, $b = 2$ MHz, $P_{\max} = 40$ dBm, and $\Delta T = 0.1$ second. The average speed of the UAV during the mobility stage is $v = 10$ m/s, the maximal onboard energy is $E_{\max} = 20$ Wh, and the power cost are $p_m = 20$ W, $p_h = 16$ W, and $p_r = 0.16$ W. In the DRL model, we set $Q = 40$, and $\gamma = 0.9$. In order to have a finite state-action space, we discretize the communication state e to be a binary vector, where if the received power of UE k is smaller than a threshold τ, $e_k = 0$; otherwise $e_k = 1$. The discrete action space is defined as: Δx_1 is "ascend by one meter", Δx_2 is "descend by one meter", and Δx_3 to Δx_{2+K} are "move towards UE k by one meter" for each $k \; \forall \; \mathcal{L}$.

In Fig. 3a, we first show the empirical probability of having a LOS downlink channel towards each UE. To evaluate the performance of the proposed DRL approach for the UAV-IR deployment, a direct transmission, a static IR, and a UAV-IR without learning are introduced as baselines. The static IR is placed at $(20, 10, 20)$ to bypass the building and establish LOS BS-IR-UE links. However, the bodies of human users may block mmWave channels. For the non-learning scheme, the UAV-IR moves towards a UE by one meter, every time downlink blockage occurs. As shown in Fig. 3a, the static IR scheme yields a LOS probability of around 50%. Due to the mobility of the UAV, the proposed UAV-IR scheme can maintain a LOS probability of over 90%, and the non-learning
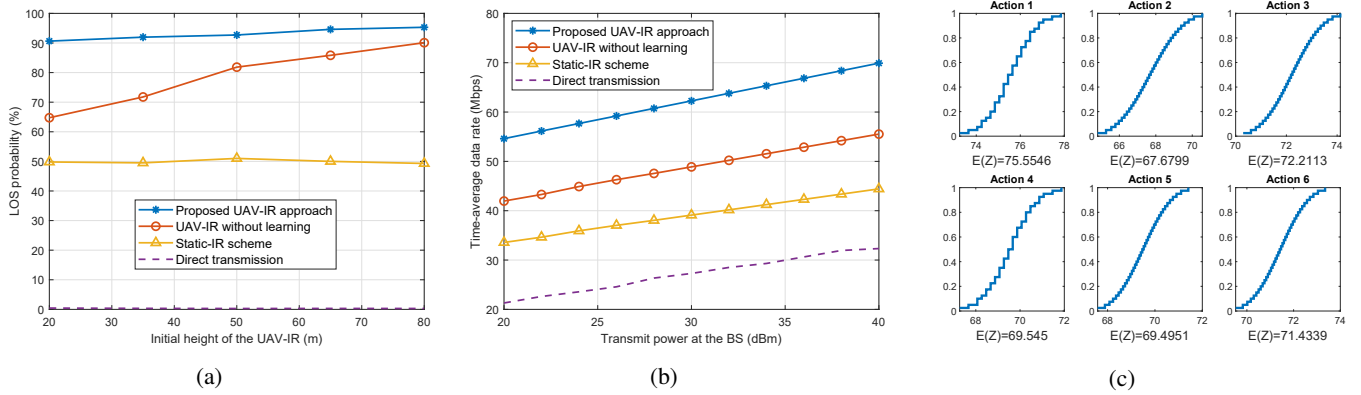
Fig. 3: (a) The achievable LOS probability of downlink mmWave channels. (b) The downlink rate increases, as the transmit power at the BS increases. (c) The return distribution of each action under the worst-case communication state $e = \mathbf{0}_K$.

UAV-IR results in a probability of over 60%. Moreover, as the altitude of the UAV-IR increases, the LOS downlink probability will naturally increase for both UAV-IR schemes. However, since the DRL-based deployment can estimate the potential of each action in improving the mmWave communication over a long term, the proposed method always yields a higher LOS probability than the non-learning scheme.

Fig. 3b shows the time-average data rate of mmWave downlink communications, as the transmit power of the BS increases. When the transmit power increases from 20 to 40 dBm, the downlink data rates of all schemes become higher. First, compared with the direct transmission scheme, the proposed UAV-IR approach yields a performance gain of over two-folds in the downlink data rate, due to a higher LOS probability. Meanwhile, the DRL-enabled deployment of the UAV-IR improves the communication performance by over 25% and 50%, compared to the non-learning UAV-IR and the static IR, respectively. For the non-learning scheme, its short-sighted strategy causes a frequent movement of the UAV-IR, thus yielding a lower downlink rate.

Fig. 3c shows how to choose the optimal action under a worst-case state $e = \mathbf{0}_K$, where the received power at each UE is lower than the threshold. In this case, Fig. 3c shows the return distribution and the expected return value of each action. Since $\Delta \boldsymbol{x}_1$ yields the highest expected reward, the optimal action under the worst-case communication state $e = \mathbf{0}_K$ is to increase the altitude of the UAV-IR by one meter.

V. CONCLUSION

In this paper, we have proposed a novel DRL-enabled approach to the deploy a UAV-IR for efficient downlink transmissions over mmWave frequencies towards multiple UEs. To maximize the downlink sum-rate, the optimal precoding matrix at the BS and reflection coefficient of the IR have been derived. In order to model the propagation environment of mmWave communications, the DRL method has been proposed to optimize the location of the UAV-IR, so as to maximize the downlink communication capacity. Simulation results show that the proposed DRL-based deployment of the UAV-IR yields a significant advantage, compared to a non-learning

UAV-IR, a static IR, and a direct transmission schemes, in terms of the average data rate and the achievable downlink LOS probability. Future research will focus on multiple UAV-IRs in outdoor communication scenarios with mobile users.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134 – 142, May/June 2020.

[2] O. Semiari, W. Saad, and M. Bennis, "Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4802–4816, May 2017.

[3] Q. Zhang, W. Saad, and M. Bennis, "Reflections in the sky: Millimeter wave communication with UAV-carried intelligent reflectors," in *Proc. of IEEE Global Communications Conference*, Hawaii, USA, Dec 2019, pp. 1–6.

[4] M. N. Soorki, W. Saad, and M. Bennis, "Ultra-reliable millimeter-wave communications using an artificial intelligence-powered reflector," in *Proc. of IEEE Global Communications Conference*, Hawaii, USA, Dec 2019, pp. 1–6.

[5] P. Wang, J. Fang, X. Yuan, Z. Chen, H. Duan, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *arXiv preprint arXiv:1908.10734*, 2019.

[6] D. Yue, H. H. Nguyen, and Y. Sun, "Analysis of intelligent reflecting surface-assisted mmWave doubly massive-MIMO communications," *arXiv preprint arXiv:2003.00282*, 2020.

[7] Y. Cao and T. Lv, "Intelligent reflecting surface aided multi-user millimeter-wave communications for coverage enhancement," *arXiv preprint arXiv:1910.02398*, 2019.

[8] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3064–3076, Feb 2020.

[9] Q. Zhang, W. Saad, and M. Bennis, "Millimeter wave communications with an intelligent reflector: Performance optimization and distributional reinforcement learning," *arXiv preprint arXiv:2002.10572*, 2020.

[10] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Deep learning for large intelligent surfaces in millimeter wave and massive MIMO systems," in *Proc. of IEEE Global Communications Conference*, Hawaii, USA, Dec 2019, pp. 1–6.

[11] C. Huang, R. Mo, C. Yuen *et al.*, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *arXiv preprint arXiv:2002.10072*, 2020.

[12] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. of Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, Feb 2018.

[13] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, Mar 2018.