This article has been accepted for publication in IEEE Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2024.3376713

Adaptive Obstacle-Aware RIS Switch for Mobile mmWave Access Networks

Bibo Zhang, Ilario Filippini, Zhuxian Lian, Yinjie Su

***Abstract*—Thanks to its capability of manipulating electromagnetic signals, the reconfigurable intelligent surface (RIS) is gaining momentum in alleviating the impact of blockages on mmWave signals, by providing redirected transmission paths. However, obstacles can also inevitably appear in the redirected paths. This can be solved by installing multiple RISs and switching among them. In this letter, for the first time, we adaptively switch among RISs for a mobile user in real time to optimize its achievable rate, without need for a priori knowledge on potential obstacles. We present an actor-critic based approach to learn unknown obstacles and variational spatial correlations originated by the user mobility, which is followed by the analysis on ergodic achievable rate. Experimental results have shown that the approach can achieve rates about 15% less than the optimum and 76% more than the state-of-the-art.**

***Index Terms*—RIS-aided mmWave access networks, obstacles, user mobility, adaptive RIS switch.**

I. Introduction

As one of the main reliefs to explosively increasing global mobile traffic, millimeter-wave (mmWave) bands have been considered for radio access networks (RANs) to enable large bandwidths and make underutilized spectrum portions available at high frequencies. However, the consequent benefits can vanish in the harsh propagation environments characterized by obstacles, where mmWave networks are typically deployed. Recently, the reconfigurable intelligent surfaces (RISs) have emerged as a tool to control the electromagnetic waves propagation. RISs are meta-surfaces containing reflecting elements, whose reactance and resistance can be adjusted, resulting in a tunable reflection / redirection of the mmWave signals [1, 2]. Therefore, they can be exploited to bypass obstacles and designed to be integrated into many applications of 6G communications[3]. With a careful design of the power pattern, even a quasi-static broad coverage can be achieved [4]. However, in practice, obstacles can inevitably appear also in the redirected path, either in the path between the base station (BS) and the RIS or in the path between the RIS and the user equipment (UE). Therefore, one potential solution is to deploy multiple RISs to provide more alternatives.

In this letter, we investigate a mmWave access network equipped with multiple RISs installed in the service area where a UE randomly moves and a few obstacles with unknown locations and sizes are randomly deployed. We consider a control framework such that the RIS that can maximize the signal-to-noise ratio (SNR) perceived by the UE is constantly activated as the UE moves in the service area. In such scenario, the SNR depends not only on the UE's position and the RIS's setups (e.g., location, orientation, hardware configuration, etc.), but also on the blockage condition of the reflected path. Theoretically, the constantly-changing user position and the lack of obstacles information require frequent and timely SNR estimation and comparison across different RISs, which is infeasible to obtain in real time. Therefore, a new appropriate RIS switch approach is required.

To tackle unknown and dynamic factors and enable real-time operations in such networks, we take advantage of the adaptability of reinforcement learning (RL), which can obtain knowledge about uncertain environments through successive interactions. Once the RL agent is trained, it can make adaptive and real-time decisions based on the current network state, without a priori information about obstacles or continuous time-costly and redundant channel estimation procedures.

In the literature, most of the RIS-related works [1, 5] focus on the scenarios with only one RIS. Some works consider multiple RISs and give the RIS selection logic [6, 7]. A general RIS selection strategy is presented in [6] via sorting SNRs, and the ergodic capacity is derived considering outdated channel information. In [7], the authors investigate optimum location-based RIS selection policies, considering product-scaling and sum-scaling path-loss models. These works set good examples of RIS selection. However, none of them focus on providing real-time switch operations that can avoid unknown potential obstacles standing in redirected propagation paths.

There are some existing works resorting to RL to address diverse technical issues in RIS-aided wireless access networks such as [8, 9]. However, as far as we know, there's no work applying RL to deal with the RIS switch problem above described. Moreover, we derive the frame-wise ergodic achievable rate for the proposed RL-based RIS selection approach. The experimental results show that our approach can provide rates approaching the optimum with the gap of only 15.35% and exceeding those of the state-of-the-art by 75.9%.

II. System Model and Problem Formulation

We consider a mmWave access network that consists of a BS, a set of RISs $\mathcal{R}$, each with E_r ($r \in \mathcal{R}$) elements, and a mobile UE. The RISs are installed at diverse locations and with different orientations. The UE's movement follows the well-known Random Waypoint model, which generates random trajectories. Several obstacles are randomly deployed

This work is partially supported by the National Natural Science Foundation of China under Grant 62001194. (*Corresponding author: Bibo Zhang*)

Bibo Zhang, Zhuxian Lian and Yinjie Su are with the Ocean College, Jiangsu University of Science and Technology, 212100 Zhenjiang, China (e-mail: bibo.zhang@just.edu.cn; zhuxianlian@just.edu.cn; yinjiesu@just.edu.cn). Ilario Filippini is with Dipartimento di Elettronica, Informazione e Biongegneria, Politecnico di Milano, 20133 Milan, Italy (e-mail: ilario.filippini@polimi.it).

This article has been accepted for publication in IEEE Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2024.3376713

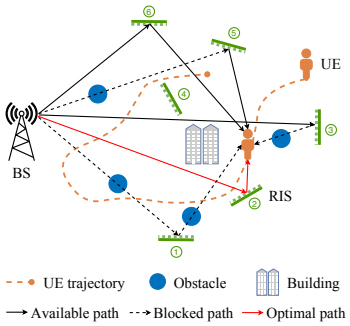


Fig. 1: A mobile mmWave network with multiple RISs.

and possibly obstruct links. An example of such scenario is depicted in Fig. 1. The time domain $\mathcal{T}$ of the system is divided into frames, each of which consists of T slots of equal length δ. In a slot t, an RIS is selected, if needed, as passive relay of the transmission from the BS to the mobile UE[1].

We consider a downlink scenario and adopt the path loss model in [10] to characterize the line-of-sight (LoS) link budget, while we model non-line-of-sight (NLoS) conditions considering 3D obstacles standing in transmission paths. In particular, the path loss experienced by the UE at slot t, if served by the r-th RIS, is denoted as $\rho_{t,r}$. 3D obstacles standing in the BS-RIS and RIS-UE paths are modeled as rectangular screens and induce a blockage attenuation according to the knife-edge diffraction model indicated in 3GPP specs [11]. The attenuation caused by an obstacle to the path between the BS and the r-th RIS is denoted as χ_r^{BR}, and that between r-th RIS and the UE at t-th time slot is denoted as $\chi_{t,r}^{\mathrm{RU}}$. Fig. 1 illustrates different blockage cases where the obstacles stand in either the BS-RIS and RIS-UE paths (RIS ①), or the RIS-UE path (RIS ③), or the BS-RIS path (RIS ⑤).

In such a scenario, we aim to identify, in each slot, the optimal RIS, thus the optimal redirected path, considering physical conditions (i.e., RIS hardware conditions, obstacles, UE mobility), to maximize the average UE SNR in a frame:

$$\max_{\mathbf{I}} \frac{1}{T} \sum_{t=1}^{T} \sum_{r=1}^{|\mathcal{R}|} \frac{P^{\mathrm{B}} \rho_{t,r} \chi_r^{\mathrm{BR}} \chi_{t,r}^{\mathrm{RU}} \mathbf{I}_{t,r}}{P^{\mathrm{N}}}, \tag{1}$$

where P^{B} and P^{N} are the transmit power of BS and the noise power. $\mathbf{I}$ is a matrix of binary variables whose element $\mathbf{I}_{t,r}$ takes value 1, if the r-th RIS is selected at slot t; 0, otherwise.

The RIS-aided mmWave access network described above is characterized by a UE moving with arbitrary directions and speeds, RISs having diverse hardware configurations, and unknown obstacles randomly deployed. Such a complex and dynamic network, full of uncertain factors, makes it tough to decide the best RIS in advance or in time, based on the limited channel information. We believe that RL can allow appropriate and adaptive RIS switch, thanks to its high efficiency in learning unknown environments with dynamic statistics.

III. Adaptive Obstacle-Aware RIS Switch

In this section, we propose an Advantage Actor Critic (A2C) [12] based RIS switch approach, which trains a policy that can be applied to adaptively switch across RISs in real time, based only on the information of the latest UE position captured[2], without any further need for channel estimation or information about potential obstacles. Then, we analyze the ergodic achievable rate when applying the trained RIS switch policy throughout the operation process.

A. RL Components

Acting as a central controller at the BS, the RL agent interacts with the physical network, in order to learn 1) the latent channel characteristics of different reflected paths as UE moves, and 2) the unknown obstacles that attenuate the signal. We consider one step in RL equivalent to one time slot of the physical frame, and design the following RL components.

State Space The state at step t is defined as the latest captured UE's 2D position (x_t, y_t): $S_t = [x_t, y_t]$.

Action Space The set of candidate actions is the set of available RISs, i.e., $A_t \in \mathcal{R}$.

Reward Function The design of the reward function is based on the two considerations below. First, as we aim to maximize the average per-frame SNR perceived by the UE, the immediate reward should be proportional to the instantaneous SNR. Second, the UE's movement can lead to considerable differences of BS-RISs-UE spatial correlations, thus SNRs, across different time slots, which can further result in an unstable and tough learning process. To mitigate such an impact, we multiply the perceived SNR by a distance-based factor. In particular, the reward is defined as:

$$R_t = \varsigma \cdot \mu_t \cdot \min_{r \in \mathcal{R}} \{ d_r^{\mathrm{BR}} \cdot d_{t,r}^{\mathrm{RU}} \}, \tag{2}$$

where μ_t is the SNR perceived by the UE at slot t, whose magnitude depends on the size of the service area, the number of RISs, the sizes of RISs, etc. This can lead to diverse reward magnitudes among different scenarios. To maintain the reward to a specific range for all scenarios, normalizing factor ς is introduced, which is adjusted empirically for different scenarios. The factor $\min_{r \in \mathcal{R}} \{ d_r^{\mathrm{BR}} \cdot d_{t,r}^{\mathrm{RU}} \}$ counterbalances the path loss difference across sparse UE locations. d_r^{BR} is the distance between the BS and r-th RIS, which can be computed in advance based on the knowledge on the BS and r-th RIS positions. $d_{t,r}^{\mathrm{RU}}$ is the approximate distance between the UE and r-th RIS at t, which can be computed beforehand as well. First, the service area is divided into a fine-grained grid. Then, the distances between cell centers and RISs are computed and recorded. Finally, $d_{t,r}^{\mathrm{RU}}$ is the distance between the r-th RIS and the cell center closest to the UE position at t.

B. Details of Learning Procedures

The A2C neural network (NN) model is composed of a fully connected (FC) preprocessing layer, actor layers and critic layers, with connection weight vectors $\boldsymbol{\zeta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. The FC layer preprocesses the input state and feeds the

[1]We here use frame-slot time scale to measure the UE throughput, however any other time scale could be applied to the proposed RIS switch approach, if needed.

[2]Considering the time cost in capturing the UE position, there could exist some error between the latest obtained UE position and the current one, which has a limited impact on the proposed approach, as demonstrated in Sec. IV.

Algorithm 1 *Learning Procedures for RIS Switch*

Parameters: T_{train}, t_{max}.

1: Initialize NN weight vector $\boldsymbol{\kappa} = [\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\zeta}]$;
2: Initialize step $t \leftarrow 1$;
3: **while** $t \leqslant T_{train}$ **do**
4: $t_{start} \leftarrow t$; Get state S_t; Reset gradient $d\boldsymbol{\kappa} \leftarrow 0$;
5: **while** $t - t_{start} < t_{max}$ and S_t is not terminal **do**
6: Decide an RIS r based on $\pi(r|S_t)$ and transmit;
7: Obtain R_{t+1} and S_{t+1};
8: $t \leftarrow t + 1$;
9: **if** S_t is not terminal **then** $G \leftarrow v(S_t)$;
10: **else** $G \leftarrow 0$;
11: $i \leftarrow t - 1$;
12: **while** $i \geqslant t_{start}$ **do**
13: $G \leftarrow R_i + \gamma G$;
14: Update $d\boldsymbol{\kappa} \leftarrow d\boldsymbol{\kappa} + \nabla_{\boldsymbol{\kappa}} L$ based on Eq. (7);
15: $i \leftarrow i - 1$;
16: Update $\boldsymbol{\kappa}$ using $d\boldsymbol{\kappa}$ based on Eq. (8);

extracted features to the parallel actor and critic layers. The actor part $\pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})$, identified with $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, is expected to approximate the policy function that picks an action based on the current state. The critic part $v_\pi(S_t;\boldsymbol{\phi},\boldsymbol{\zeta})$, parameterized with $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$, is aimed at approximating the value function that evaluates the policy's performance. Collaborative training of actor and critic parts ultimately produces the best RIS switch policy.

In particular, the critic part approximates the value function $v(S_t;\boldsymbol{\phi},\boldsymbol{\zeta})$ by minimizing the *value loss* as in (3).

$$\min_{\boldsymbol{\phi},\boldsymbol{\zeta}} L_v = a(S_t, A_t;\boldsymbol{\phi},\boldsymbol{\zeta})^2, \tag{3}$$

$$a(S_t, A_t;\boldsymbol{\phi},\boldsymbol{\zeta}) = G^k(S_t) - v(S_t;\boldsymbol{\phi},\boldsymbol{\zeta}), \tag{4}$$

$$G^k(S_t) = \sum_{i=0}^{k-1} \gamma^i R_{t+i} + \gamma^k v(S_{t+k}), \tag{5}$$

where $a(S_t, A_t;\boldsymbol{\phi},\boldsymbol{\zeta})$ computes the difference between the return $G^k(S_t)$ and the value $v(S_t;\boldsymbol{\phi},\boldsymbol{\zeta})$, thus used to constitute the value loss.

The actor part approximates the policy function by updating $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ in the direction of increasing the expected return $\mathbb{E}[G^k(S_t)]$, i.e.,$\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}}\mathbb{E}[G^k(S_t)]$, whose unbiased estimate is $\nabla_{\boldsymbol{\theta},\boldsymbol{\zeta}} \log \pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta}) a(S_t, A_t)$. In addition, to prevent a premature convergence to sub-optimal policies, the policy entropy $H(\pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})) = -\sum_{A_t} \pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta}) \log \pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})$ is included in the *policy loss* minimization:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\zeta}} L_p = &- \log \pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta}) a(S_t, A_t) \\ &- \eta H(\pi(A_t|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})),\end{aligned} \tag{6}$$

where η controls the importance of entropy regularization.

We concatenate $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ into $\boldsymbol{\kappa}$, i.e., $\boldsymbol{\kappa} = [\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\zeta}]$, and iteratively update $\boldsymbol{\kappa}$ through Eq. (8) to minimize the total loss L in Eq. (7), which is the sum of policy loss and value loss:

$$\begin{aligned}\min_{\boldsymbol{\kappa}_t} L = &- \log \pi(A_t|S_t;\boldsymbol{\kappa}_t) a(S_t, A_t) - \eta H(\pi(A_t|S_t;\boldsymbol{\kappa}_t)) \\ &+ \beta a(S_t, A_t;\boldsymbol{\kappa}_t)^2,\end{aligned} \tag{7}$$

$$\boldsymbol{\kappa}_{t+1} = \boldsymbol{\kappa}_t + o_t \nabla_{\boldsymbol{\kappa}_t} L. \tag{8}$$

The learning procedures are summarized in Algorithm 1. The training phase spans T_{train} steps, and the NN model is updated every t_{max} steps. First of all, the connection weight vector $\boldsymbol{\kappa}$ is initialized (Line 1). Before each model update, the pointer to the beginning of the training sequence t_{start} is updated and the gradient $d\boldsymbol{\kappa}$ is set to 0 (Line 4). The data used in iterative updates of gradients and NN weights is collected through the loop in Lines 5-8. At each step, an RIS r is adopted as the passive relay, according to the current policy $\pi(r|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})$ (Line 6). The reward and the next state are set respectively based on Eq. (2) and the latest UE's location obtained (Line 7). Once data has been collected, return $G^k(S_t)$ is computed in Line 13, which is initialized in Lines 9-10. The gradients are computed based on Eq. (7) (Lines 12-15). Finally, $\boldsymbol{\kappa}$ is updated using Eq. (8) (Line 16). The above operations are performed iteratively till the end of the training phase.

C. Algorithm Convergence and Complexity

1) Convergence: We let $\boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\zeta}]$ and express the value function using linear function approximation $\hat{v}(S_t;\boldsymbol{\varphi}) = \nu(S_t)^\intercal \boldsymbol{\varphi}$, whose resulting error is ϵ. Based on [13], we have

$$\frac{1}{(1+t-\tau_t)} \sum_{k=\tau_t}^{t} \mathbb{E}\|\boldsymbol{\varphi}_k - \boldsymbol{\varphi}_k^\star\|^2 = \mathcal{O}(\frac{1}{t^{1-\sigma}}) + \mathcal{O}(\frac{\log t}{t^\sigma}), \tag{9}$$

where τ_t is the mixing time of an ergodic Markov chain. $\sigma \in (0, 1)$ is a constant. $\boldsymbol{\phi}_k$ and $\boldsymbol{\phi}_k^\star$ are, respectively, the parameter and unknown parameter of value function at iteration k. $\mathcal{O}(\cdot)$ is the operation of the order of magnitude.

Let $\boldsymbol{\vartheta} = [\boldsymbol{\theta}, \boldsymbol{\zeta}]$ and $\nabla J(\boldsymbol{\vartheta}) = \nabla_{\boldsymbol{\vartheta}} \log \pi(A_t|S_t;\boldsymbol{\vartheta}) a(S_t, A_t)$, then based on [13], we have

$$\min_{0 \le k \le t} \mathbb{E}\|\nabla J(\boldsymbol{\vartheta}_k)\|^2 = \mathcal{O}(\epsilon) + \mathcal{O}(\frac{1}{t^{1-\sigma}}) + \mathcal{O}(\frac{\log^2 t}{t^\sigma}). \tag{10}$$

Based on the convergence of both critic and actor parts shown in Eqs. (9) and (10), the proposed approach can converge.

2) Complexity: Let U denote the total number of units in the NN model. The computational complexity of model updates $f(U)$ and inferences $g(U)$ depends on forward propagation and backpropagation operations on units and connection weights [14]. As Algorithm 1 consists of parts of data collection (Lines 5-8) and model updates (Lines 9-16), the computational complexity of the algorithm is $\mathcal{O}(T_{train} \cdot (f(U) + g(U)))$.

D. Ergodic Achievable Rate

The ergodic achievable data rate C, in the operation process of length $|\mathcal{T}|$, depends on the RL agent's trained policy π^*:

$$C = \mathbb{E}_{\pi^*}\left[\frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} \log_2(1 + \frac{P^{\mathrm{B}} \cdot \sigma_{\pi^*}(t)}{P^{\mathrm{N}}})\right], \tag{11}$$

where $\mathbb{E}[\cdot]$ is the statistical expectation operation. $\sigma_{\pi^*}(t)$ is the total path attenuation at slot t, which includes the path loss and obstacle attenuation, and is determined by the policy π^*. Based on the Jensen's inequality, i.e., $\mathbb{E}[\log_2(\cdot)] \le \log_2(\mathbb{E}[\cdot])$, the upper bound of the ergodic achievable rate can be obtained:

$$C \le \frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} \log_2(1 + \frac{P^{\mathrm{B}} \cdot \mathbb{E}_{\pi^*}[\sigma_{\pi^*}(t)]}{P^{\mathrm{N}}}). \tag{12}$$

This article has been accepted for publication in IEEE Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2024.3376713

TABLE I: Achievable rates for different UE speeds (bits/s/Hz).

		Optimal		Proposed	
		2-20m/s	20-60m/s	2-20m/s	20-60m/s
100×100	15 RISs	8.146	8.132	6.896	6.881
	30 RISs	9.372	9.367	7.352	7.348
200×200	15 RISs	12.133	12.119	9.962	9.948
	30 RISs	13.364	13.359	10.918	10.912

We abbreviate $\pi^*(r|S_t;\boldsymbol{\theta},\boldsymbol{\zeta})$ to $\pi^*_{t,r}$ and let $\boldsymbol{\pi}^*_t = [\pi^*_{t,r}]_{r\in\mathcal{R}}$ be the decision probability vector at slot t. In addition, suppose that the r-th RIS is sampled from $\boldsymbol{\pi}^*_t$, then the total path attenuation $\sigma_{\pi^*}(t)$ can be specified as $\sigma_{t,r}$. Further, denoting $\boldsymbol{\sigma}_t = [\sigma_{t,r}]_{r\in\mathcal{R}}$, we can obtain $\mathbb{E}_{\pi^*}[\sigma_{\pi^*}(t)] = \boldsymbol{\pi}^*_t\boldsymbol{\sigma}^\intercal_t$. After we submit it to Eq. (12), the upper bound becomes:

$$C \leq \frac{1}{|\mathcal{T}|}\sum_{t=1}^{|\mathcal{T}|}\log_2(1+\frac{P^{\mathrm{B}}\cdot\boldsymbol{\pi}^*_t\boldsymbol{\sigma}^\intercal_t}{P^{\mathrm{N}}}). \tag{13}$$

As the policy generated by the proposed A2C-based approach follows the categorical distribution, which is a special case of the general RL-based approach, rather than stopping at the upper bound, we can further derive the exact ergodic achievable rate. Let $\omega_{t,r} = \log_2(1+\frac{P^{\mathrm{B}}\sigma_{t,r}}{P^{\mathrm{N}}})$ be the achievable rate if r-th RIS is employed at slot t, then the ergodic rate through the whole operation process can be computed as:

$$C = \frac{1}{|\mathcal{T}|}\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left[(\sum_{t\in\mathcal{T}}\omega_{t,\boldsymbol{\xi}(t)})\cdot(\prod_{t\in\mathcal{T}}\pi^*_{t,\boldsymbol{\xi}(t)})\right], \tag{14}$$

where $\boldsymbol{\xi}(t)\sim\pi^*(S_t)$ denotes the adopted RIS at slot t. $\boldsymbol{\xi} = [\boldsymbol{\xi}(t)]_{t\in\mathcal{T}}$ is an action sequence over time domain $\mathcal{T}$, which is composed of a combination of the RISs sampled at different slots. $\boldsymbol{\Xi}$ is the whole set of possible action sequences. By introducing the rate matrix $\boldsymbol{\Omega} = [\omega_{t,r}]_{t\in\mathcal{T},r\in\mathcal{R}}$ and the policy matrix $\boldsymbol{\Pi}^* = [\pi^*_{t,r}]_{t\in\mathcal{T},r\in\mathcal{R}}$, the ergodic rate can be derived:

$$C = \frac{1}{|\mathcal{T}|}tr(\boldsymbol{\Pi}^*\boldsymbol{\Omega}^\intercal), \tag{15}$$

where $tr(\cdot)$ is the trace of a matrix.

IV. Numerical Results

In this section, we first compare the achievable rates of the proposed and baseline approaches, considering different numbers and sizes of RISs. Then we experiment the proposed approach with different blockage cases, UE speeds and UE position estimation errors. The values in the figures and tables are averages over 10 randomly generated network instances.

A. Network Scenario Settings

In the experiments, we consider a 100 m × 100 m service area where 1 BS is located at the left-side midpoint. We respectively consider 15 and 30 RISs deployed with random locations and orientations facing the BS. The RISs can contain either 100×100 or 200×200 elements whose length and width are both equal to half of the wave length and reflection coefficient amplitude is 1. And we consider 10 and 20 randomly placed small obstacles with (radius, height) of $(2.5, 6)$ m or large ones with $(7.5, 12)$ m. The UE moves according to the Random Waypoint model with speeds uniformly sampled from $[2, 20]$ m/s or $[20, 60]$ m/s, moving intervals in $[2, 6]$ s, and pause intervals in $[0, 1]$ s. The heights of the BS, RISs and UE are 10 m, 2 m and 1.7 m, respectively. The network operates at the frequency of 26 GHz. The BS is equipped with 1 antenna panel with transmission power of 34 dBm. The noise power at the UE follows $-174 + 10\log_{10}(\mathrm{BW}) + \mathrm{NF}$, where the bandwidth BW is 200 MHz and the noise figure NF is 9 dB.

B. RL Model Settings

The critic and actor networks are both composed of 5 FC layers of 32 hidden units, whose output layers use Softmax and linear functions, respectively. We train an RL model based on the experience data of 100000 episodes, each with 100 steps. It takes approximately 2.3 hours on our Intel(R) Core(TM) i5-12500 @3.00GHz and 24.0GB RAM machine. The discount factor γ is set to 0.99. The coefficients η and β in Eq. (7) are set to 0.01 and 0.5. The RMSProp optimizer is used to minimize the total loss, with a learning rate of 0.0007.

C. Performance Analysis

We first compare the proposed adaptive A2C-based RIS switch approach (referred to as *Proposed*) against the following three baselines, on the scenarios with 15 and 30 RISs that are equipped with 100×100 and 200×200 elements:

- **Optimal Scheme** (referred to as *Optimal*): provides the achievable data rate under the ideal condition where the best RIS is adopted in each slot, assuming that the BS stands in the god's view and is aware of the obstacles.
- **Ergodic Achievable Rate** (referred to as *Ergodic*): is the ergodic rate that can be achieved by the proposed A2C-based approach, derived in Sec. III-D.
- **Location-Based Scheme [7]** (referred to as *MinProdDis*): uses the RIS with the minimum product of the BS-RIS and RIS-UE distances. We adapt it to a real-time version, using the same distances recorded in Sec. III-A.

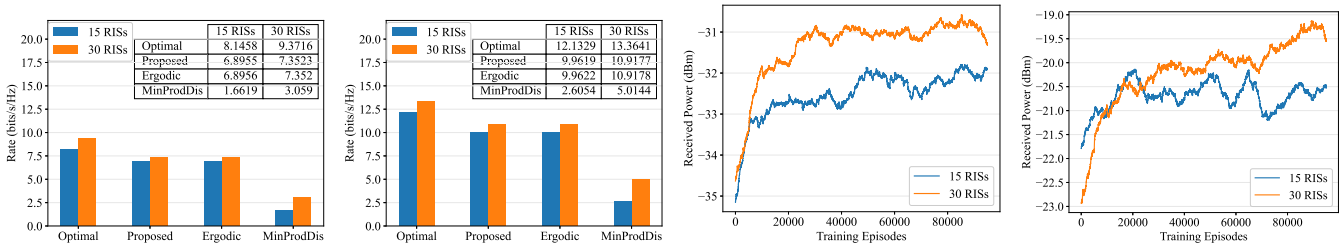
In Figs. 2a and 2b, the tiny difference between the rates achievable with *Proposed* and *Ergodic* confirms the effectiveness of our approach, and that the number of samples in the testing phase is sufficient so that numerical results adhere to the theoretical analysis. And *Proposed* provides rates very close to those of *Optimal* and outperforms *MinProdDis*, regardless of the number of RISs and the number of elements. These two observations show that *Proposed* can learn spatial information about network entities and unknown obstacles.

Moreover, if maintaining the same number of elements, installing more RISs can help to increase UE rates. This is reasonable because an increased number of RISs provide a larger action space for the RL agent and a denser relay system for the UE. Finally, equipping the same number of RISs with more elements can considerably increase the UE rates.

The training curves of A2C models applied in Figs. 2a and 2b are shown in Figs. 2c and 2d, w.r.t. the UE's received power. As can be seen, despite the number of RIS elements, the models trained with 30 RISs can gain greater improvement than those with 15 RISs, because a increased number of RISs lead to a larger action space thus more possibilities.

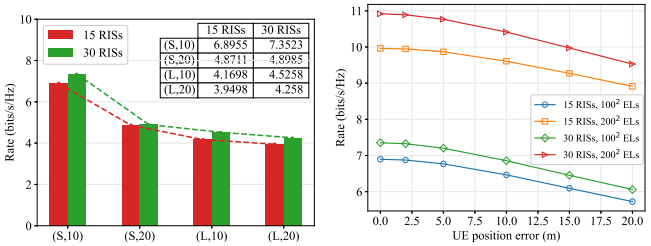
Subsequently, we examine the performance of the *Proposed* on different numbers (10 and 20) of obstacles with different

This article has been accepted for publication in IEEE Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2024.3376713



(a) 100×100 elements (b) 200×200 elements (c) 100×100 elements (d) 200×200 elements

Fig. 2: Comparison on the achievable rates and the training curves of the proposed approach.



(a) Different blockage cases. (b) Different UE position errors.

Fig. 3: The impact of different blockage cases and UE position errors on the performance of the proposed approach.

sizes (small (S) and large (L)). Fig. 3a shows the achievable rates in scenarios where 15 and 30 RISs, both equipped with 100×100 elements, are deployed. We can see that even severe blockages have very limited impact ($\leq$ 3 bits/s/Hz rate decrease) on the performance of the *Proposed*. When the number or size of obstacles is small, increasing the other factor considerably cuts down the rates, whereas when either factor is large, the impact of increasing the other is quite limited.

Then, we test the *Optimal* and *Proposed* on different UE speed ranges ($[2, 20]$ m/s and $[20, 60]$ m/s) in scenarios with 10 small obstacles. The achievable UE rates are listed in Table I. As expected, both approaches see slight rate decreases as UE speeds are increased. Despite the number of RIS elements, both approaches exhibit about 0.015 and 0.005 bits/s/Hz rate decrease for 15 and 30 RISs, respectively. 30 RISs enable less rate decrease, because more RISs can serve more UE positions.

Finally, we test the sensitivity of the *Proposed* to the UE position estimation errors (from 0 m to 20 m), in the scenarios with 10 small obstacles. As can be seen from Fig. 3b, despite the RIS settings, as the UE position error increases, the UE rates decrease at a close rate. The rate decrease is not evident when the UE position error is less than 5 m. The rate decreases corresponding to 10 m and 20m are 0.5 and 1 bits/z/Hz. This demonstrates that the *Proposed* is not sensitive to the UE position error, therefore, can be used to achieve satisfactory rates even with outdated UE position information due to time costs of positioning and RIS configuration procedures.

V. Conclusion

In this letter, we have investigated a real-time RIS switch approach for mobile mmWave network scenarios where multiple RISs are installed and the associated redirected propagation paths could be blocked by obstacles. We have resorted to the actor-critic technique to learn uncertain obstacles and constantly changing spatial conditions caused by the mobile user. Finally, we have derived the ergodic rate achievable with the proposed approach. The experimental results carried out in different scenarios have shown that the proposed RIS switch approach can effectively learn unknown environmental features and approximate the optimal strategy. It can even approach the ideal optimum with only around 15% gap and outperform the state-of-the-art by 76%.

References

[1] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (ris) assisted wireless coverage extension: Ris orientation and location optimization," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 269–273, 2020.

[2] Z. Lian *et al.*, "A novel geometry-based 3-d wideband channel model and capacity analysis for irs-assisted uav communication systems," *IEEE Trans. Wireless Commun.*, 2023.

[3] J. Xu *et al.*, "Reconfiguring wireless environments via intelligent surfaces for 6g: Reflection, modulation, and security," *Science China Information Sciences*, vol. 66, 2023.

[4] M. He, J. Xu, W. Xu, H. Shen, N. Wang, and C. Zhao, "Ris-assisted quasi-static broad coverage for wideband mmwave massive mimo systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2551–2565, 2023.

[5] K. Feng, Q. Wang, X. Li, and C. K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for miso communication systems," *IEEE Wireless Commun. Lett.*, no. 99, 2020.

[6] N. Mensi and D. B. Rawat, "Reconfigurable intelligent surface selection for wireless vehicular communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1743–1747, 2022.

[7] Y. Fang, S. Atapattu, H. Inaltekin, and J. Evans, "Optimum reconfigurable intelligent surface selection for wireless networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6241–6258, 2022.

[8] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, 2020.

[9] Q. Zhang, W. Saad, and M. Bennis, "Millimeter wave communications with an intelligent reflector: Performance optimization and distributional reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1836–1850, 2021.

[10] W. Tang *et al.*, "Path loss modeling and measurements for reconfigurable intelligent surfaces in the millimeter-wave frequency band," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6259–6276, 2022.

[11] 3GPP, *Study on channel model for frequencies from 0.5 to 100 GHz, TR 38.901*.

[12] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016, pp. 1928–1937.

[13] Y. F. Wu, W. ZHANG, P. Xu, and Q. Gu, "A finite-time analysis of two time-scale actor-critic methods," in *NIPS*, 2020, pp. 17 617–17 628.

[14] R. Livni, S. Shalev-Shwartz, and O. Shamir, "On the computational efficiency of training neural networks," in *NIPS*, 2014, pp. 855–863.