# Ultra-Reliable Millimeter-Wave Communications using an Artificial Intelligence-Powered Reflector

Mehdi Naderi Soorki[1,2], Walid Saad[1], and Mehdi Bennis[2]

[1]Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA.
[2]Centre for Wireless Communications, University of Oulu, Finland.
Emails: {mehdin,walids}@vt.edu, mehdi.bennis@oulu.fi

***Abstract*—In this paper, a novel framework for guaranteeing ultra-reliable millimeter-wave (mmW) communications using a smart, artificial intelligence (AI)-powered mmW reflector is proposed. The use of an AI-powered reflector allows changing the propagation direction of mmW signals and, thus, improving coverage particularly for non-line-of-sight (LoS) areas. However, due to the possibility of stochastic blockage over mmW links, designing an intelligent phase shift-control policy for the mmW reflector to guarantee ultra-reliable mmW communications becomes very challenging. In this regard, first, based on the framework of risk-sensitive reinforcement learning, a parametric risk-sensitive episodic return is proposed to maximize the expected bit rate while mitigating the risk of non-LoS mmW link in the presence of future stochastic blockage over the mmW links. Then, a closed-form approximation for the gradient of the risk-sensitive episodic return is analytically derived. To *directly* find the optimal policy for the proposed phase-shift controller, a parametric functional-form policy is implemented using a deep recurrent neural network (RNN). Then, based on the derived closed-form gradient of risk-sensitive episodic return, the deep RNN-based parametric functional-form policy is trained. The efficiency of the proposed AI-powered reflector is evaluated in an office environment. Simulation results show that the root-mean-square errors between the optimal and approximate phase shift-control policies of the proposed deep RNN is $1.35\%$ in the worst case. Moreover, on average, the mean value and variance of the achievable rates resulting from the deep RNN-based policy are only $1\%$ and $2\%$ less than the optimal solution for different unknown mobile users' trajectories, respectively.**

*Index Terms*— Millimeter wave networks; reflectors; deep neural networks; risk-sensitive reinforcement learning; 5G and beyond.

## I. Introduction

Millimeter wave (mmW) communications from 30 to 300 gigahertz (GHz) band is a promising candidate solution to enable high-speed wireless access in next-generation wireless networks [1]–[3]. Nevertheless, the high attenuation and scattering of mmW propagation, even by small objects, renders the design of mmW wireless networks very challenging [2]. To overcome this challenge, integrating massive antennas for highly directional beamforming at both mmW access point (APs) and user equipments (UEs) has been proposed [1], [2]. However, applying beamforming techniques will render the use of directional mmW links very sensitive to random blockage caused by people and objects in a dense environment. This, in turn, gives rise unstable LoS mmW links and unreliable communications [2], [3]. To provide robust LoS mmW links, one proposed solution is to deploy ultra-dense mmW APs and active relay nodes to improve link quality using multi-connectivity for a given UE [3], [4]. However, the deployment of multiple mmW APs and active relay nodes is not economically feasible and can also increase the control signalling overhead. To decrease signalling overhead and alleviate economic costs while also establishing reliable mmW communications, a mmW reflector can be used between the mmW AP and UE [1], [5], [6]. The use of a mmW reflector allows changing the propagation direction of mmW signals thus improving coverage particularly for non-LoS areas.

Several recent works such as in [4]–[8] have proposed the deployment of mmW reflectors to establish reliable mmW links. In [5], the authors present efficient designs for both transmit power allocation and coefficient control of the surface reflecting elements. Their goal is to optimize spectrum or energy efficiency subject to individual link budget guarantees for the mobile users. However, the work in [5] does not consider stochastic blockage and, thus, the results cannot be generalized to a real-world mmW system. In [4], the authors implement a smart mmWave reflector to provide high data rates between a virtual reality headset and game consoles. To handle beam alignment and tracking between the mmWave reflector and the headset, their proposed method must try every combination of mirror transmit beam angle and headset receive beam angle and, thus, it incurs significant overhead. In [8], the authors designed a smart reflector consisting of 224 reconfigurable patches. Then, they proposed a two-stage phase shift-control algorithm for the smart reflector-assisted 802.11ad networks. Their proposed phase shift-control algorithm uses exhaustive search to find the optimal beam angle of the AP and phase shift of the reflector. The work in [6] used software-controlled metasurfaces as smart reflectors for indoor scenarios. In the model of [6], a central server receives incoming reflector reports, calculates the optimal configuration per reflector to increase the received power of the target user, and sends the corresponding commands. The existing works in [4]–[8] assume static reflectors and do not provide any efficient solution to intelligently control the configuration of smart reflectors in an adaptive manner. Moreover, the objective of [4]–[8] is to increase the coverage probability or signal-to-noise ratio without mitigating the risk of non-LoS mmW link. In practice, an intelligent solution is required to capture unknown future blockages, adaptively control the configuration of smart mmW reflectors, and guarantee ultra-reliable mmW communication.

The main contribution of this paper is a novel framework

This research was supported by the U.S. National Science Foundation under Grants CNS-1836802 and CNS-1814477.

for guaranteeing ultra-reliable mobile mmW communications using a smart mmW reflector called artificial intelligence (AI)-powered reflector. The proposed AI-based approach can autonomously reconfigure the patches of the smart reflector in presence of stochastic blockage over the mmW links. To solve the phase shift-control problem in a reflector-assisted mmW network and guarantee ultra-reliable mmW communications, the proposed AI-powered reflector uses a deep recurrent neural network (RNN) which learns an adaptive phase-shift control policy. First, we formulate the problem as a stochastic optimization problem whose goal is to not only maximize the expected bit rate but to also mitigate the risk of non-LoS mmW link over time. Then, we use deep and risk-sensitive reinforcement learning (RL) to solve the problem in an adaptive manner. The parametric functional-form policy is implemented using a deep RNN which *directly* searches the optimal policy of the phase-shift controller. In this regard, a closed-form approximation for the gradient of risk-sensitive episodic return is analytically derived, and the RNN-based policy is trained using this derived closed-form gradient. Simulation results show that the root-mean-square errors (RMSEs) between optimal and RNN-based parametric functional-form policy is $1.35\%$ in the worst case. Moreover, on average, the mean value and the variance of the achievable rates from the RNN-based policy are only $1\%$ and $2\%$ less than the optimal solution, respectively.

The rest of the paper is organized as follows. Section II presents the system model and the stochastic and risk-sensitive optimization problem in the smart reflector-assisted mmW networks. In Section III, based on the framework of deep and risk-sensitive RL, we propose a deep RNN to intelligently solve the stochastic and risk-sensitive optimization problem of reflector configuration. Then, in Section IV, we numerically evaluate the the proposed policy-gradient approach which intelligently controls the reflector configuration. Finally, conclusions are drawn in Section V.

## II. System Model and Problem Formulation

### A. System model

Consider the downlink of an indoor reflector-assisted mmW network, composed of one mmW access point (AP). In this network, due to the blockage of mmW signals, there exist areas in which it is not possible to establish LoS mmW links between the mmW AP and UE, particularly when users are mobile. We call these areas as *dark areas*. Each mmW AP and UE have $N_a$ and $N_u$ antennas to form their beams, respectively. In our model, there is a AI-powered smart mmW reflector that can intelligently reflect the mmW signals from the mmW AP to the target UE in the dark area. Each mmW reflector has $L$ patches of small patch reflectors indexed by $l$. Without loss of generality, we focus on one mmW AP and a user device connected to it. We consider discrete time slots indexed by $t$. The angle-of-departure (AoD) from the mmW AP to the patch $l$ of the reflector at time slot $t$ is represented by $\theta_{l,t}$. Then, the mmW reflector establishes a LoS mmW link using a controlled reflected path to bypass an obstacle. Let $\psi_{l,t} \in [-\pi, \pi]$ be the amount of shift that the patch $l$ of the mmW reflector changes in the phase of received signal at time slot $t$. We also define $B$ as the total number of possible discrete values for phase shifting per patch. Thus, the angle-of-arrival (AoA) $\phi_{l,t}$ from a UE to the patch $l$ of the mmW reflector at time slot $t$ is given by $\psi_{l,t} + \theta_{l,t}$. The vector $\Psi^{(i)} \in \mathbb{R}^L$ for $i = 1, ..., B$ includes the $i$-th possible value for all controlled phase shifts of the $L$ patches at the mmW reflector. We define $p_t^{(i)} = \Pr(\Psi_t = \Psi^{(i)})$ as the phase shift control policy of the mmW reflector which is defined as the probability that the mmW reflector selects the $i$-th phase shift vector to reflect the received signal from the mmW AP at time slot $t$. For a given vector $\Psi^{(i)}$ at time slot $t$, an $N_a \times N_u$ channel matrix between mmW AP and user is defined as follows [9]:

$$\boldsymbol{H}_t(\Psi^{(i)}) = [\boldsymbol{a}_{\text{Tx,a}}(\theta_{1,t}), ..., \boldsymbol{a}_{\text{Tx,a}}(\theta_{L,t})] \times \text{diag}(\boldsymbol{\alpha}) \times [\boldsymbol{a}_{\text{Rx,u}}(\phi_{1,t}), ..., \boldsymbol{a}_{\text{Rx,u}}(\phi_{L,t})]^H, \quad (1)$$

where $\boldsymbol{a}_{\text{Tx,a}}(\theta_{l,t}) = \frac{[e^{-j\frac{N_a-1}{2}\frac{2\pi}{\lambda}d\sin(\theta_{l,t})}, ..., e^{j\frac{N_a-1}{2}\frac{2\pi}{\lambda}d\sin(\theta_{l,t})}]}{\sqrt{N_a}}$ and $\boldsymbol{a}_{\text{Rx,r}}(\phi_{l,t}) = \frac{[e^{-j\frac{N_u-1}{2}\frac{2\pi}{\lambda}d\sin(\phi_{l,t})}, ..., e^{j\frac{N_u-1}{2}\frac{2\pi}{\lambda}d\sin(\phi_{l,t})}]}{\sqrt{N_u}}$ respectively denote the response vectors for the AoD to the $l$-th patch and the AoA from patch $l$ [9]. Here, $\phi_{l,t} = \psi_{l,t}^{(i)} + \theta_{l,t}, \forall l = 1, ..., L$, $\lambda$ is the mmW signal wavelength, and $d = \frac{\lambda}{2}$. $\boldsymbol{\alpha} = \sqrt{N_a N_u}[\frac{\alpha_1}{\sqrt{\rho_1}}, ..., \frac{\alpha_L}{\sqrt{\rho_L}}]$ where $\rho_l$ is the path loss and $\alpha_l$ is the complex channel gain of path $l$ from the mmW AP through patch $l$ to the UE [9]. Consequently, when a controller at the mmW reflector selects the phase shift vector $\Psi^{(i)}$ at time slot $t$, the bit rate over the mmW link between the mmW AP and a user through mmW reflector is given by [9]:

$$r_t(\Psi^{(i)}) = w \log_2 \det \left[ \boldsymbol{I}_{N_a} + \frac{q}{N_a w \sigma^2} \boldsymbol{H}(\Psi^{(i)}) \boldsymbol{H}(\Psi^{(i)})^* \right], \quad (2)$$

where $q$ is the transmission power, $w$ is mmW bandwidth, and $\sigma^2$ is the density of noise.

Fig. 1 is an illustrative example that shows how one mmW AP and one mmW reflector are used to bypass the blockage during two consecutive time slots $t$ and $t+1$. As we can see in Fig. 1, since the user is in the kitchen during time slots $t$ and $t+1$, that area will be a dark area for the mmW AP, the mmW reflector is therefore used to cover the user. In this example, the mmW reflector uses 3 patches of reflectors. Thus, the AoA of mmW AP signals directed to the patches are $\theta_{1,t}, \theta_{2,t}, \theta_{3,t}$ at time slot $t$ and $\theta_{1,t+1}, \theta_{2,t+1}, \theta_{3,t+1}$ at time slot $t+1$. Then, the mmW reflector shifts the phase of the received signal following $\Psi_t$ and $\Psi_{t+1}$ during time slot $t$ and $t+1$, respectively. As we can see in Fig. 1, the user receives data at time slot $t$ over the reflected mmW link because the mmW reflector selects the phase shift values $\Psi_t$ such that the AoDs of the mmW signals $\phi_{1,t}, \phi_{2,t}$, and $\phi_{3,t}$ are matched to the user's location. Meanwhile, at time slot $t+1$, the AoDs of mmW signals $\phi_{1,t+1}, \phi_{2,t+1}$, and $\phi_{3,t+1}$ are not matched to the user's location, due to user mobility causing the reflector to select the wrong values for phase shifts, $\Psi_{t+1}$, to reflect the received signals at time slot $t+1$. In this case, the phase-shift control policy needs to not only consider

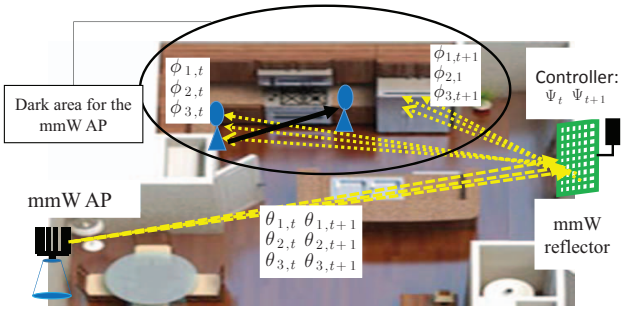


Fig. 1: An illustrative example of the system model.

the unknown future trajectory of mobile users but also to adapt itself to the possible stochastic blockages in future time slots. Thus, an intelligent phase shift-control policy, which can predict unknown stochastic blockages, is required to guarantee ultra-reliable mmW communication particularly for indoor scenarios having a high numbers of dark areas.

### *B. Phase shift control for reflector-assisted mmW networks*

Due to the stochastic changes of the mmW blockage between mmW AP or UE and mmW reflector, the phase shift-control policy at a given slot $t$ depends on unknown future changes in the LoS mmW links. To guarantee ultra-reliable mmW links, instead of maximizing the expected bit rate, we will seek to mitigate the risk which takes into account the variance and skewness of LoS mmW links over time. To this end, by using Taylor series, we define an exponential utility function of the user's bit rate before taking the expectation, yielding higher order moments [10]. Consequently, we formulate the phase shift-control problem for a reflector-assisted mmW network as follows:

$$\max_{\{[p_{t'}^{(i)}]_{B\times T}\}} \frac{1}{\mu}\log\left(\mathbb{E}_{r_{t'}}\{e^{(\mu\sum_{t'=t}^{t+T-1} r_{t'})}\}\right), \tag{3}$$

s.t.

$$0 \leq p_{t'}^{(i)} \leq 1, \forall i \in \{1,...,B\}, \forall t' \in \{t,...,t+T-1\}, \tag{4}$$

$$\sum_{i=1}^{B} p_{t'}^{(i)} = 1, \forall t' \in \{t,...,t+T-1\}, \tag{5}$$

where the parameter $\mu \leq 0$ denotes the risk sensitivity parameter and the operator $\mathbb{E}$ is the expectation operation [10]. In (3), the objective is to maximize the average of episodic sum of future bit rate, $\sum_{t'=t}^{t+T-1} r_{t'}$, while also minimizing the variance of this sum so as to capture the tail of rate distribution, using phase shift-control policies $p_{t'}^{(i)}, \forall t' \in \{t,...,t+T-1\}$. The risk sensitivity parameter penalizes the variance and skewness of the episodic sum of future bit rate. In (3), $\{r_{t'}|t'=t,...,t+T-1\}$ depends on the phase shift-control policies and the unknown AoA from user's location during $T$-consecutive future time slots.

The phase shift-control problem in (3) is a combinatorial and stochastic optimization problem whose objective is to assign the direction of reflections to each arriving mmW signals. This problem does not admit a closed-form solution and has an exponential complexity [11]. The complexity of the stochastic optimization problem in (3) becomes more significant due to the unknown probabilities for possible random network changes such as the mmW link blockage events and the user's locations [11] as well as the large size of the state-action space for decision variables. However, we are interested in developing a low-complexity policy to solve (3) that can intelligently adapt to intensive dynamics of mmW links over future time slots. Next, we propose a framework based on principles of deep RL to solve the optimization problem in (3) with low complexity and in an adaptive manner.

## III. Intelligent Phase Shift-Control Policy

In this section, we present the proposed adaptive policy based on a new deep and risk-sensitive RL framework to solve the phase shift-control problem in (3). We model the problem in (3) as a partially observable Markov decision process (POMDP) represented by the tuple $\{\mathcal{S},\mathcal{A},\mathcal{O},P,R,o_0\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $P$ is the stochastic state transition function, $P(s',s,a) = \Pr(s_{t+1}=s'|s_t=s,a_t=a)$, $R(a_t,s_t)$ is the immediate reward function, and $o_0$ is the initial observation for the controller of the mmW reflector [12]. In our problem, the action space is the set of possible shift phases $\mathcal{A} = \{\Psi^{(i)} \in \mathbb{R}^L | i = 1,...,B\}$ for the reflector patches, the observation is the bit rate over mmW link $r_t \in \mathcal{O}$, and the immediate reward is current bit rate $R(a_t,s_t) = r_t$. The state is the AoA from the patches to the user $\phi_l$ for $l = 1,...,L$ which is not observable. We represent the policy of our POMDP in a parametric functional-form $\pi_{\boldsymbol{\theta}}(a_t|r_t) = \Pr\{a=a_t|o=r_t,\boldsymbol{\theta}\}$ where $\boldsymbol{\theta}$ is a parameter vector. If $\Lambda_T = \{(a_{t'},r_{t'})|t' = t,...,t+T-1\}$ is a trajectory of the POMDP during $T$-consecutive time slots, then the stochastic episodic reward function during future $T$-consecutive time slots is defined as $R_{T,t} = \sum_{t'=t}^{t+T-1} r_t$. The unknown probability of trajectory $\Lambda_T$ is equal to $\Pi_{\boldsymbol{\theta}}(T) = \prod_{t'=t}^{t+T-1} \pi_{\boldsymbol{\theta}}(a_t|r_t)\Pr\{r_{(t+1)}|a_t,r_t\}$. We define the risk-sensitive episodic return for parameter vector $\boldsymbol{\theta}$ at time slot $t$ as $J(\boldsymbol{\theta},t) = \frac{1}{\mu}\log\left(\mathbb{E}_{\pi_{\boldsymbol{\theta}}}\{e^{(\mu R_{T,t})}\}\right)$ [10]. Given the parametric functional-form policy $\pi_{\boldsymbol{\theta}}$, the goal of the phase shift controller is to solve the following optimization problem:

$$\max_{\{\boldsymbol{\theta}\in\mathbb{R}^N\}} J(\boldsymbol{\theta},t), \tag{6}$$

s.t.

$$0 \leq \pi_{\boldsymbol{\theta}}(a_{t'}|r_{t'}) \leq 1, \forall a_{t'} \in \mathcal{A}, \forall t' \in \{t,...,t+T-1\}, \tag{7}$$

$$\sum_{\forall a_{t'}\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a_{t'}|r_{t'}) = 1, \forall t' \in \{t,...,t+T-1\}, \tag{8}$$

where $T << N$. We will define the parameter vector $\boldsymbol{\theta}$ and the value of $N$ in Subsection III-A. To solve the optimization problem in (8), the phase shift controller needs to have full knowledge about the transition probability $\Pi_{\boldsymbol{\theta}}(T)$, and all possible values of $R_{T,t}$ for all of the trajectories during $t' = t,...,t+T-1$ from the POMDP under policy $\pi_{\boldsymbol{\theta}}$. Since the explicit characterization of the transition probability and values of the episodic reward for all of the trajectories is not feasible in highly dynamic mmW neworks, we use an RL framework to

solve (8). More specifically, we use policy-search approaches to find the optimal phase shift-control policy to solve problem in (8) for the following reasons. First, value-based approaches such as $Q$-learning are oriented toward finding deterministic policies. However, the optimal policy is often stochastic and policy-search approaches can select different phase shifts with specific probabilities by adaptively tuning the parameters in $\boldsymbol{\theta}$ [11]. Second, any small change in the estimated value of an action can cause it to be (or not) selected in the value-based approaches and these small changes are highly probable in highly dynamic mmW networks. In this regard, the most popular policy-search method is the policy-gradient method. In the policy-gradient method, the gradient objective function is calculated and used in gradient-ascend algorithm. The gradient $\nabla J(\boldsymbol{\theta}, t)$ of the risk-sensitive objective function is approximated as follow.

**Theorem 1.** The gradient of the objective function, $J(\boldsymbol{\theta}, t)$, in (8) is approximated by:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, t) \approx \mathbb{E}_{\Lambda_T}\{\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T) \times \big((1-\mu \mathbb{E}_{\Lambda_T}\{R_{T,t}\})R_{T,t} + \frac{\mu}{2} R_{T,t}^2\big)\}, \tag{9}$$

where $\nabla_{\boldsymbol{\theta}} \Pi_{\boldsymbol{\theta}}(T) = \sum_{t'=t}^{t+T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|r_t)$ and $\mathbb{E}_{\Lambda_T}\{R_{T,t}\} = \sum_{\Lambda_T} \Pi_{\boldsymbol{\theta}}(T) R_{T,t}$

*Proof.* Let $\Lambda_T = \{(a_{t'}, r_{t'}) | t = t, ..., t+T-1\}$ be a trajectory during $T$-consecutive time slots which leads to the episodic reward $R_{T,t} = \sum_{t'=t}^{t+T-1} r_{t'}$. The Taylor expansion of the utility function for small values of $\mu$ yields:

$$J(\boldsymbol{\theta}, t) \simeq \mathbb{E}_{\Lambda_T}\{R_{T,t}\} + \frac{\mu}{2} \text{Var}_{\Lambda_T}\{R_{T,t}\}. \tag{10}$$

Since $\text{Var}_{\Lambda_T}\{R_{T,t}\} = \mathbb{E}_{\Lambda_T}\{R_{T,t}^2\} - \big(\mathbb{E}_{\Lambda_T}\{R_{T,t}\}\big)^2$, we can rewrite:

$$J(\boldsymbol{\theta}, t) \simeq \mathbb{E}_{\Lambda_T}\{R_{T,t} + \frac{\mu}{2} R_{T,t}^2\} - \frac{\mu}{2}\big(\mathbb{E}_{\Lambda_T}\{R_{T,t}\}\big)^2. \tag{11}$$

The probability of the trajectory $\Lambda_T$ is $\Pi_{\boldsymbol{\theta}}(T)$. Thus, we can write $J(\boldsymbol{\theta}, t) \simeq \sum_{\Lambda_T}\{\Pi_{\boldsymbol{\theta}}(T)(R_{T,t} + \frac{\mu}{2} R_{T,t}^2)\} - \frac{\mu}{2}\big(\sum_{\Lambda_T} \Pi_{\boldsymbol{\theta}}(T)\{R_{T,t}\}\big)^2$. Hence:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, t) \simeq \sum_{\Lambda_T}\{\nabla_{\boldsymbol{\theta}} \Pi_{\boldsymbol{\theta}}(T)(R_{T,t} + \frac{\mu}{2} R_{T,t}^2)\} - \mu\big(\sum_{\Lambda_T} \nabla_{\boldsymbol{\theta}} \Pi_{\boldsymbol{\theta}}(T)\{R_{T,t}\}\big)\big(\sum_{\Lambda_T} \Pi_{\boldsymbol{\theta}}(T)\{R_{T,t}\}\big). \tag{12}$$

Since $\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T) = \frac{\nabla_{\boldsymbol{\theta}} \Pi_{\boldsymbol{\theta}}(T)}{\Pi_{\boldsymbol{\theta}}(T)}$, we can write $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, t) \approx \mathbb{E}_{\Lambda_T}\{\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T)\big(R_{T,t} + \frac{\mu}{2} R_{T,t}^2\big)\} - \mu \mathbb{E}_{\Lambda_T}\{\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T) R_{T,t}\}\mathbb{E}_{\Lambda_T}\{R_{T,t}\}$. By performing additional simplifications, we have:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, t) \approx \mathbb{E}_{\Lambda_T}\{\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T) \times \big((1-\mu \mathbb{E}_{\Lambda_T}\{R_{T,t}\})R_{T,t} + \frac{\mu}{2} R_{T,t}^2\big)\}, \tag{13}$$

where $\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(T) = \sum_{t'=t}^{t+T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|r_t)$ because the probability of trajectory $\Lambda_T$ occuring is equal to $\Pi_{\boldsymbol{\theta}}(T) = \prod_{t'=t}^{t+T-1} \pi_{\boldsymbol{\theta}}(a_t|r_t) \Pr\{r_{t+1}|a_t, r_t\}$ and $\log(xy) = \log(x) + \log(y)$. ■

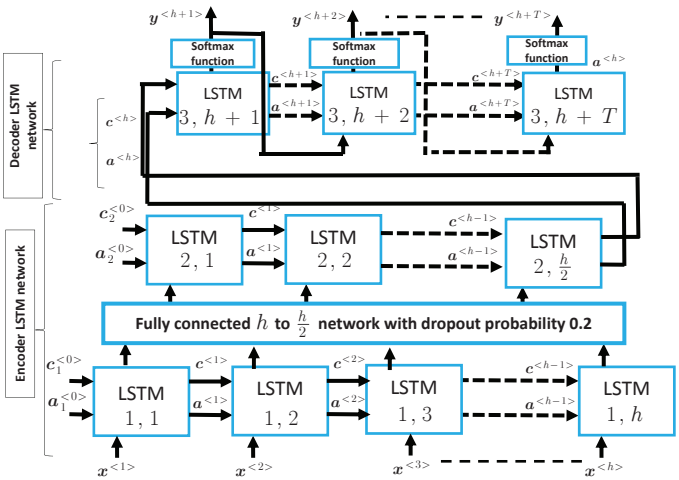


Fig. 2: The many-to-many deep RNN for implementing the mmW reflector phase-shift control policy $\pi_{\boldsymbol{\theta}}$.

Following Theorem 1, we can use (7) to solve the optimization problem in (8) using a gradient ascent algorithm and, then, find the optimal phase shift-control policy. To calculate (7), we need a lookup table of all trajectories of risk-sensitive values and policies over time. However, this lookup table is not practically available for a highly dynamic indoor mmW networks. To overcome this challenge, we propose to combine deep neural network (DNN) with the RL policy-search method. Such a combination was shown in [12], where a DNN learns a mapping from the partially observed state to an action without requiring any lookup table of all trajectories of risk-sensitive values and policies over time. Consequently, next, we propose an RL algorithm that uses a deep RNN based on policy gradient for solving (8).

### A. Architecture of the proposed deep RNN for solving (8)

We use a DNN to approximate the policy $\pi_{\boldsymbol{\theta}}(a|s)$ for solving (8). Our proposed DNN for risk-sensitive and deep RL method is presented in Fig. 2. Here, the parameters $\boldsymbol{\theta} \in \mathbb{R}^N$ includes the weights over all connections of the proposed DNN where $N$ is equal to the number of connections [12].

1) *Input Layer:* since the Markov decision process is partially observable, we consider the history of the POMDP during $h$-consecutive previous time slots as the input layer of DNN [11], [12]. Thus, the input $x_t$ is equal to $\{x^{<k>} | k = 1, ..., h\}$ in which $x^{<k>} = \{p^{(i)}_{t-(h-k+1)} | i = 1, ..., B\} \cup \{r_{t-(h-k+1)}\}$. The first and second inputs of entry in $x^{<k>}$ indicate the phase shift-control policies and received bit rate at time slot $t-(h-k+1)$.

2) *Output layer:* the output $y_{t'}$ is $T$ vectors of size $B$ for future time slots $t' = t, ..., t+T-1$, where element $i$ in the vector captures the probability $p^{(i)}_{t'}$ for selecting phase shifts vector $\Psi^{(i)}$ at time slot $t'$.

3) *Long short-term memory (LSTM) layer*: the dynamics of LoS mmW links depend on the unknown blockage due to the mobility of the target user over time. On the other hand, at a given time slot, the user's trajectory over future time slots

depends on the locations of the mobile user during previous time slots, especially for an indoor scenario. Thus, we use an RNN to deploy a deep RL that aggregates the observations of mmW blockages during previous time slots and makes a more precise prediction of the next state of the POMDP compared to traditional DNNs [13]. In particular, we use a many-to-many RNN in which the encoder network has two LSTM layers with $h$ and $\frac{h}{2}$ units, and the decoder network has one LSTM layer with $T$ units, as shown in Fig. 2. There is a fully connected $h$-to-$\frac{h}{2}$ network between two layers of the encoder network. We use a drop out probability 0.2 to prevent over-fitting of our proposed deep RNN [13]. The input of LSTM unit $k$ at layer 1 is $\boldsymbol{x}^{<k>}$ where $k = 1, ..., h$. After using a Softmax function for each LSTM unit at the encoder network, the output of each unit $h+t$ at layer 3, $\boldsymbol{y}^{<h+t>} = \pi_{\boldsymbol{\theta}}$, is a vector of size $B$ that includes the phase shift-control policy at each time slot $t$, $y_i^{<h+t>} = p_t^{(i)}$. This architecture was chosen because the encoder LSTM network maps the history of the POMDP to the *internal state* shown by $\boldsymbol{c}^{<h>}$ and $\boldsymbol{a}^{<h>}$ in Fig. 2, more precisely compared to other DNN architectures, and then the decoder LSTM network can use this internal state to predict the phase shift-control policy [13], [14].

Consider a training set $\mathcal{M}$ of $M$ samples that is available to train the RNN network. Each training sample $m$ includes a sample of policies and bit rates during $h$-consecutive time slots before time slot $t_m$, $\{\pi_{\boldsymbol{\theta}}^{(m)}(a_{t'}|r_{t'}), r_{t'}^{(m)}|t' = t_m - h + 1, ..., t_m\}$, and policies and bit rates during future $T$-consecutive time slots after time slot $t_m$, $\{\pi_{\boldsymbol{\theta}}^{(m)}(a_{t'}|r_{t'}), a_{t'}^{(m)}, r_{t'}^{(m)}|t' = t_m + 1, ..., t_m + T\}$. Consequently, based on Theorem 1 and by replacing the expectation with sample-based estimator for $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, we use the gradient-ascend algorithm to train the RNN as follows:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \frac{1}{M} \sum_{m=1}^{M} \Big( \nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}^{(m)}(T) \times$$

$$\big((1 - \mu R_M) R_{T,t_m} + \frac{\mu}{2} R_{T,t_m}^2\big)\Big), \tag{14}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \tag{15}$$

where $\nabla_{\boldsymbol{\theta}} \Pi_{\boldsymbol{\theta}}^{(m)}(T) = \sum_{t'=t_m+1}^{t_m+T} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}^{(m)}(a_{t'}^{(m)}|r_{t'}^{(m)})$, $R_{T,t_m} = \sum_{t'=t_m+1}^{t_m+T} r_{t'}^{(m)}$, and $R_M = \frac{1}{M} \sum_{m=1}^{M} R_{T,t_m}$. Here, $\alpha$ is the learning rate. In summary, to solve the optimization problem in (3), we use deep and risk-sensitive RL and solve (8) using the gradient ascent algorithm. Hence, we implement the parametric functional-form policy $\pi_{\boldsymbol{\theta}}$ with the proposed deep RNN in Fig. 2. Then, we use (14) and (15) to iteratively train the proposed RNN and optimize $\boldsymbol{\theta}$.

## IV. Simulation Results and Analysis

For our simulations, the carrier frequency is set to 73 GHz and the mmW bandwidth is 1 GHz. The number of transmit antennas at the mmW AP and receive antennas at the UE are set to 64 and 4, respectively. The duration of each time slot is 1 millisecond. The transmission power of mmW AP is 46 dBm and the power density of noise is -88 dBm. We assume that the mmW reflector assigns a square of $3 \times 3 = 9$ patches

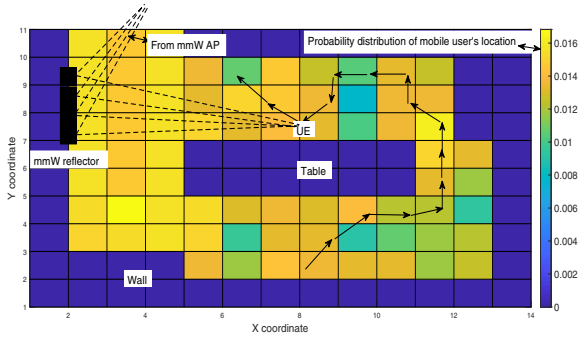


Fig. 3: The distribution probability of mobile user's location.

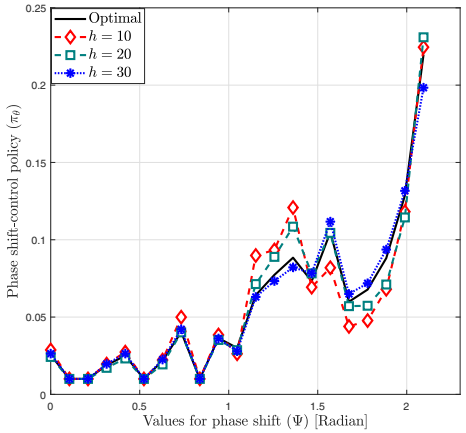


Fig. 4: Phase shift-control policy, $\pi_{\boldsymbol{\theta}}$.

to reflect the mmW signals. Each patch of the mmW reflector shifts the phase of the mmW signals with a step of $\frac{\pi}{30}$ radians from the range $[-\pi, \pi]$. For generating the data set of mobile users' trajectories to train and test our proposed RNN, we use a modified random walk model. In this case, the direction of each user's next step is chosen based on the probability of user's presence at next step location in a given indoor scenario. The probability of user's presence at each location can be obtained by indoor localization techniques [15]. We consider a 35-sq. meter office environment with a table at the center. Fig. 3 shows the probability distribution of the user's locations in the office, the location of the mmW reflector, and an illustrative example of a user trajectory. For comparison purposes, we consider the optimal solution, as a benchmark in which the exact user's locations and optimal strategies for the reflector during the next future $T$-time slots are known. Fig. 4 shows the approximated phase shift-control policy, $\pi_{\boldsymbol{\theta}}$, for different history length, $h$, when the size of the training set is 90% of the data set, with $\mu = -0.3$, and $T = 2$. From Fig. 4, we can observe that the RMSEs between the approximate and optimal policies are 0.0135, 0.0081, and 0.0057, respectively, for $h = 10$, $h = 20$, and $h = 30$. This is due to the fact that when the history length increases, the deep RNN can capture the previous dynamics over mmW links better and predict the future mobile user's trajectory more precisely. Thus, the approximated phase shift-control policy based on RNN is near the optimal solution.

In Fig. 5, we show the RMSE of the approximated phase

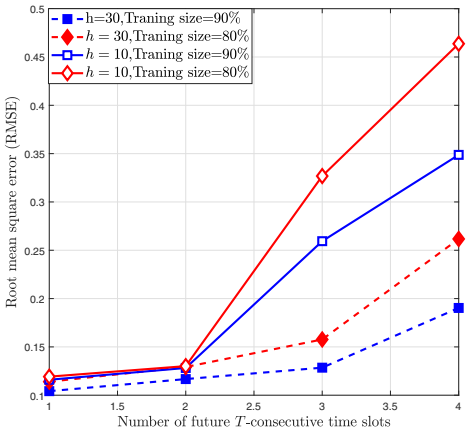


Fig. 5: RMSE for the parametric functional-form policy.

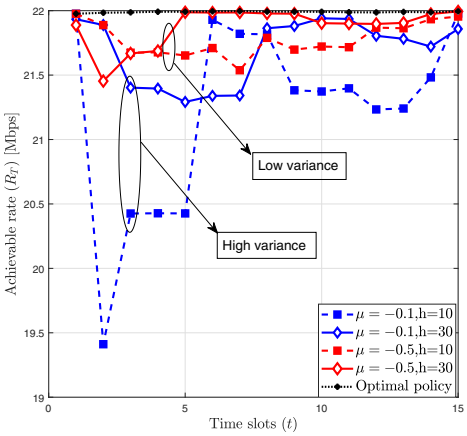


Fig. 6: Achievable rate, $R_T$.

shift-control policy with respect to the number of future consecutive time slots, $T$, for different history lengths $h$ and training sizes. Following Fig. 5, the RMSE increases as the length of future consecutive time slots used to predict the policies increases. In addition, the RMSE decreases for high value of history length. On average, the RMSE for history length 30 is 26% less than a history length of 10. Moreover, as the training size increases from 80% to 90% of the dataset, the RMSE decreases. This is because a training deep RNN with 90% compared to 80% of the dataset leads to better prediction of the phase shift-control policy.

In Fig. 6, we show the achievable rate, $R_T$, following the approximate phase shift-control policy over time for different history lengths and risk sensitivity parameters. As we can see from Fig. 6, a lower value for the risk sensitivity parameter leads to a higher reliability over the highly dynamic mmW link. On average, the variance of the achievable rate is 30% and 2% more than the variance of the optimal solution for $\mu = -0.1$ and $\mu = -0.5$, respectively. Moreover, when the history length increases, the achievable rate from our proposed RNN-based approach becomes closer to the optimal solution. On average, the mean values of the achievable rates are 3% and 1% less than the optimal solution for $h = 10$ and $h = 30$, respectively. This is due to the fact that more history length leads to lower RMSE (See Fig. 5).

## V. Conclusion

In this paper, we have proposed a novel framework for guaranteeing ultra-reliable mobile mmW communications using an AI-powered reflector that provides LoS mmW coverage by reflecting mmW signals toward mobile users. First, based on risk-sensitive RL, we have proposed a parametric risk-sensitive episodic return to maximize the expected bit rate and mitigate the risk of non-LoS mmW link in the presence of unknown future stochastic blockage over the mmW links. Then, we have analytically derived a closed-form approximation for the gradient of the risk-sensitive episodic return. We have modeled the parametric functional-form policy with a deep RNN, which is trained based on the derived closed-form gradient of the risk-sensitive episodic return. Simulation results have shown the effectiveness of the proposed approach and its ability to achieve a near-optimal solution.

## References


[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: applications, trends, technologies, and open research problems," *IEEE Network*, 2019.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: it will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] O. Semiari, W. Saad, M. Bennis, and M. Debbah, "Integrated millimeter wave and sub-6 ghz wireless networks: a roadmap for joint mobile broadband and ultra-reliable low-latency," *IEEE Wireless Communications Magazine*, to apear, 2019.

[4] O. Abari, D. Bharadia, A. Duffield, and D. Katabi, "Enabling high-quality untethered virtual reality," *in Proc. of the USENIX symposium on networked systems design and implementation*, pp. 1–5, Boston, MA, USA, March 2017.

[5] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Large intelligent surfaces for energy efficiency in wireless communication," *CoRR*, vol. abs/1810.06934, 2018. [Online]. Available: http://arxiv.org/abs/1810.06934

[6] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Realizing wireless communication through software-defined hypersurface environments," *arXiv:1805.06677*, May 2018.

[7] H. Kamoda, T. Iwasaki, J. Tsumochi, T. Kuki, and O. Hashimoto, "60-ghz electronically reconfigurable large reflectarray using single-bit phase shifters," *IEEE transactions on antennas and propagation*, vol. 59, no. 7, pp. 2524–2531, July 2011.

[8] X. Tan, Z. Sun, D. Koutsonikolas, and J. M. Jornet, "Enabling indoor mobile millimeter-wave networks based on smart reflect-arrays," *in Proc. of IEEE Conference on Computer Communications*, pp. 270–278, Honolulu, Hi, USA, April 2018.

[9] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter-wave MIMO in 5G systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1822–1835, March 2018.

[10] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, pp. 267–290, Nov. 2002.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*. second edition, MIT press, Cambridge, MA, 2018.

[12] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," *in Proc. of Association for the Advancement of Artificial Antelligence Symposium*, Arlington, Virginia, USA, July 2015.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[14] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: a tutorial," *IEEE Communications Surveys and Tutorials*, 2019.

[15] Z. Zhong, Z. Tang, X. Li, T. Yuan, Y. Yang, M. Wei, Y. Zhang, R. Sheng, N. Grant, C. Ling, K. S. K. X. Huan, and S. Lee, "XJTLUIndoorLoc: a new fingerprinting database for indoor localization and trajectory stimation based on Wi-Fi RSS and geomagnetic field," *in Proc. of International Symposium on Computing and Networking Workshops (CANDARW)*, pp. 228–234, Hida Takayama, Japan, Nov. 2018.