Online Reinforcement Learning for Beam Tracking and Rate Adaptation in Millimeter-wave Systems

Marwan Krunz, *Fellow, IEEE*, Irmak Aykin, *Member, IEEE*, Sopan Sarkar, *Student Member, IEEE*, and Berk Akgun, *Member, IEEE*

***Abstract*—In this paper, we propose MAMBA, a restless multi-armed bandit framework for beam tracking in directional millimeter-wave (mmW) cellular systems. Instead of relying on explicit control messages, MAMBA utilizes the ACK/NACK packets transmitted by user equipments (UEs) to the base station (BS) as a part of the hybrid automatic repeat request (HARQ) procedure. These packets are used to measure the quality of the currently operating downlink beam, and select a new downlink beam along with an appropriate modulation and coding scheme (MCS) for future transmissions. At its core, MAMBA implements an online reinforcement learning technique called adaptive Thompson sampling (ATS), which determines a good beam and associated MCS to be used for the upcoming transmissions. To evaluate MAMBA's performance, we conduct extensive simulations and over-the-air (OTA) experiments over the 28 GHz band using phased-array antennas. We study fixed-as well as adaptive-rate variants of MAMBA, and contrast it with four other beam tracking strategies: a beam selection scheme similar to the one used in 5G NR (called 'static oracle'), a theoretically optimal but practically infeasible beam tracking scheme (called 'dynamic oracle'), an ϵ-greedy algorithm [1], and the Unimodal Beam Alignment (UBA) algorithm [2]. Our results show that MAMBA achieves 182% throughput gain over the 'static oracle' and is reasonably close to the throughput of the 'dynamic oracle'. Compared to UBA, MAMBA achieves 25-35% gain in throughput, depending on UE mobility. Finally, when operated at a fixed MCS, MAMBA/ATS achieves 21% gain over the ϵ-greedy algorithm at the lowest applied MCS index, and 255% gain at the highest MCS index.**

***Index Terms*—Millimeter-wave, directional communications, beam tracking, reinforcement learning, multi-armed bandit.**

I. Introduction

Millimeter-wave (mmW) communications are a key aspect of next-generation wireless systems, including 5G [3] and WiGig [4]. The abundant mmW spectrum enables many users to be served by a base station (BS), with significantly higher data rates than what is possible at sub-6 GHz bands [5]. Traditionally, the mmW spectrum has not been utilized for terrestrial communications due to the harsh channel behavior and the immaturity of wireless technologies that could operate at such high frequencies. This limitation is, however, partially compensated for by the small wavelengths of mmW transmissions, which allow large antenna arrays to be implemented in small form-factor radios. By using high-dimensional phased-array antennas, transmissions/receptions can be narrowly beamed along desired directions. The resulting beamforming gain makes it possible to achieve high data rates, despite the unfavorable characteristics of the channel [6].

While beamforming allows for high gains, establishing and maintaining a directional link add new challenges [7]–[10]. Due to the limited scattering at mmW frequencies, the channel between the BS and the user equipment (UE) is typically sparse [11], [12], so the transmitted signal reaches the receiver along a few angular clusters. Identifying the directions of these clusters takes a considerable amount of time, prolonging the initial access (IA) process (e.g., [2], [13]–[17]) that takes place prior to establishing a BS-UE link. Once a link has been established, beam misalignment may occur frequently due to mobility, environmental changes, or even wind [18]. Such misalignment results in reduced data rate and link outage. Consequently, tracking UEs and maintaining the quality of their directional links are critical to ensuring seamless mmW communications [19], [20]. Frequent channel measurements used for this purpose add significant control overhead, lowering spectral efficiency and increasing latency.

In this paper, we propose a restless multi-armed bandit (MAB) framework called MAMBA for beam tracking and rate adaptation in mmW systems. Although, in general, MAMBA is applicable to any directional mmW network (with appropriate modifications), we study it in the context of 5G New Radio (NR), as specified by 3GPP. Note that beam management in 5G NR involves coarse beam selection (a.k.a. Phase 1) and beam refinement (a.k.a. Phase 3). Phase 3 uses a similar procedure to Phase 1 when selecting the best beam, and is repeated whenever an outage occurs without resorting back to Phase 1. Because MAMBA is meant to provide a generic solution, i.e., selecting a beam from a set of beams (be it coarse or narrow), we apply it to one phase of 5G NR beam management.

According to MAMBA, each beam is modeled as an arm in a MAB problem. The BS acts as the agent that interacts with these arms to learn the underlying system dynamics, i.e., changes in beam qualities over time. To quantify beam quality, we rely on the best possible modulation and coding scheme (MCS) that this beam can support in terms of achievable rate. We integrate a reinforcement learning (RL) algorithm, called adaptive Thompson sampling (ATS), into MAMBA and use it to select a good beam/MCS pair for the next downlink transmission(s). ATS aims at maximizing the link throughput while taking into account the estimated reward distributions associated with each beam. Due to the time-varying nature of the environment, keeping track of these

An abridged version of this paper appeared in the Proceedings of the INFOCOM'20 Conference, 27-30 April, 2020, Beijing, China.

M. Krunz and S. Sarkar is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, 85721, USA I. Aykin and B. Akgun are currently with Qualcomm Inc. This work was conducted as part of their Ph.D. work at the University of Arizona. (email: {krunz, sopansarkar, aykin, berkakgun}@email.arizona.edu)

reward distributions is nontrivial. To address this issue, ATS uses a priori information of beam quality, collected through IA, and updates this information at each iteration based on the feedback obtained from the UE. The beam and MCS to be used during the next downlink transmission are then selected based on the updated posterior distributions of the rewards, i.e., achievable rates of various beams. ATS can accurately estimate the best beam/MCS pair without making modeling assumptions about the channel and/or mobility pattern.

It is worth mentioning that other RL approaches besides ATS may also be considered for the design of MAMBA, including the Soft Actor-Critic (SAC) algorithm [21], Proximal Policy Optimization (PPO) [22], Twin delayed DDPG (TD3) [23], Asynchronous Advantage Actor Critic (A3C) [24], and the Asynchronous Advantage Actor Critic (A3C) algorithm [24]. SAC adds an entropy maximization term to the original RL objective. Balancing the maximization of the reward and the entropy encourages the resulting policy to converge to an optimal solution, while acting as randomly as possible. In MAMBA, we do not want the algorithm to act as randomly as possible, since the prior distributions learned during IA could still be valuable. Therefore, adding an entropy term to the objective function is not necessarily beneficial in our case. In PPO, an estimator of the policy gradient is computed and plugged into a stochastic gradient-ascent algorithm. Specifically, in each iteration, N parallel actors collect T time steps worth of data. Then, the PPO algorithm constructs a loss function on these NT time steps worth of data and optimizes this loss function using mini-batch stochastic gradient descent. Although PPO is more efficient than the TRPO algorithm proposed by the same authors [25], it is still computationally complex and its performance relies heavily on hyper-parameter optimization. TD3 is a Q-learning based deep RL algorithm that aims at learning two Q-functions instead of one. It uses the smaller of the two Q-values to form the targets in the Bellman error loss functions. In the underlying beam tracking problem, selecting a new action (beam) does not change the state space of the system. That is, selecting a beam at time t does not affect the set of beams that can be selected at time $t+1$, so the problem can be adequately modeled as a single-state Markov decision process (MDP) with one state. A Q-learning solution adds complexity without bringing much value. As for A3C, this algorithm uses critics to learn the value function. Multiple actors are trained in parallel and are periodically synchronized. For stability, the gradients are accumulated as part of training similar to parallelized stochastic gradient descent. As in TD3, this Q-learning method is designed to solve MDPs with more than one state, and hence is inherently more complex than the ATS algorithm.

The main contributions of this paper are as follows:

- We introduce MAMBA, a MAB framework for beam tracking and adaptive rate selection in 5G mmW systems. MAMBA does not incur extra messaging overhead. It utilizes the ACK/NACK feedback obtained from the UE to select the best beam/MCS pair.
- We develop an RL algorithm called ATS to be used in MAMBA. ATS selects the optimal beam/MCS pair so as to maximize the data rate of the underlying transmission. To address the nonstationarity in the environment, we introduce a *forget* factor that discounts the information obtained in the past and a *boost* factor that increases the impact of the recent observations on beam selection.
- We derive an upper bound on the Bayesian regret of the ATS algorithm. To account for the time-varying rewards, we utilize a discrete-time random walk process in our analysis.
- Through hardware experiments and software simulations at 28 GHz frequency using a 4×8 phased-array antenna, we verify the efficiency of ATS in terms of total delivered traffic, average data rate, instantaneous data rate, and outage duration in both indoor and outdoor scenarios. We also validate its high performance by comparing it with four other beam tracking algorithms.

II. Related Work

Efficient and reliable beam tracking in mmW systems is still an open research topic. Numerous techniques were proposed in the literature to address the issue (see [26] for a recent survey). In [27], the authors used extended Kalman filters (EKF) for angle-of-arrival (AoA) and angle-of-departure (AoD) tracking. Their method tracks the currently utilized channel cluster, i.e., only one AoA/AoD pair is tracked at a time. Similarly, the authors in [28] used Kalman filters to track the AoA and AoDs at the receiver (Rx) and transmitter (Tx), respectively. Both [27] and [28] assumed that the angles are randomly perturbed according to a zero-mean Gaussian distribution, which may not hold in reality.

In [29], the authors proposed *BeamSpy*, a scheme for predicting the quality of alternative beams by inspecting the channel response of the current beam. This is done by constructing a path skeleton that exploits channel sparsity. The model parameters were extracted from a one-time measurement and are invariant under blockage. This means that if the channel is highly dynamic, e.g., due to mobility, BeamSpy needs to update the path skeleton quite frequently, hence incurring high overhead. An extension of BeamSpy, called *Beam-forecast*, was proposed in [30]. Beam-forecast reconstructs the spatial channel profile from a few beam measurements and virtually tries each candidate beam to predict its quality without actual probing. The algorithm was implemented and tested on a custom 60 GHz platform, and analysis was carried out in an indoor environment. The results indicate that the throughput drops by more than 50% when mobility increases from 1 Km/h to 5 Km/h. Similar to BeamSpy, the effectiveness of Beam-forecast depends heavily on channel sparsity and blockage-invariant spatial correlation.

A generic mmW beam steering algorithm was proposed in [31], which utilizes the previous valid link information to initiate a search for a feasible BS/UE beam pair. The algorithm adaptively increases the sector search space around the BS to re-establish a link. It is a reactive algorithm, executed only after the link between the BS and the UE breaks down. The feasibility of the algorithm was analysed in an indoor environment with limited mobility. Even though the search space is limited to nearby beams, the algorithm still needs to perform an iterative search over this space.

The authors in [14], [32] presented beam tracking techniques that do not require dedicated control resources but instead utilize multiple RF chains to simultaneously collect channel information from several directions. Specifically, in [14], a scheme called *Agile-Link* was proposed, which tries to identify the best beam direction using a logarithmic number of measurements. Agile-Link manipulates the phase shifters in the antenna array to generate random multi-armed beams (hashes the beam space into bins) and samples multiple beam directions simultaneously. It then uses a voting mechanism to recover the directions over which signals from the UE are detected. Palacios et. al. proposed a pseudo-exhaustive beam training (PE-Training) and probabilistic beam tracking (P-Track) schemes [32]. Their approach leverages the ability of hybrid analog-digital transceivers to simultaneously collect CSI from multiple beam directions. It requires the UE to correctly detect the preamble and save its samples to obtain the complex power at the RF combiner. Both [32] and [14] use multiple simultaneous beam directions, relying on hybrid beamforming to observe the mmW channel. Creating multiple beams increases the effective beam width, resulting in overlapping beams and a lower beamforming gain. This eventually leads to less accurate estimates of the underlying channel compared to a single-beam approach. In contrast, MAMBA relies only on a single directional beam and readily available receiver signal strength (RSS) values, obtained from ACK/NACK messages.

In [33] the authors proposed a neighbour discovery (ND) technique called *FastND*, which accelerates the ND process by gathering channel information along different beam directions and using a Compressive Sensing based Beam Prediction (CSBP) and Maximum Distance based Beam Prediction (MDBP) modules. FastND was evaluated in an indoor environment with limited mobility. It has not been tested in outdoor scenarios at high mobility.

ML techniques have also been used to address the beam tracking problem. For example, the authors in [34] proposed a Long short-term memory (LSTM) based approach for tracking the AoA. They used an omnidirectional antenna and a simplistic mobility model, which can create a bias in AoA estimation. The authors in [35] also used LSTM but their technique is not intended for a single BS system. The RL-based beam tracking methods in [36], [37] utilize location information, which may not always be available. In [38], Koda et. al. studied the usefulness of past node position and velocity information for beam tracking in an unstable surrounding (under complex wire dynamics). The proposed system model consists of a fixed on-building node and a dynamic on-wire node, where beam misalignment occurs due to changes in the orientation of the on-wire node. The problem of tracking highly mobile nodes was not addressed. The authors in [39] tried to minimize the packet delivery latency by determining the optimal beam using deep deterministic policy gradient (DDPG) based RL approach. However, a simple mobility model was considered, where the UE moves along a straight line. The time it takes for the DDPG algorithm to converge is quite high.

MAB models have been extensively applied in the literature to address different online optimization problems [40]. The goal of a MAB is to capture the exploration versus exploitation tradeoff and to minimize the cumulative regret of deviating from the optimal strategy. In [1], the author studied three MAB algorithms, ϵ-greedy, upper confidence bound (UCB), and Thompson sampling (TS) [41], aiming to find the best BS-UE beam pair that maximizes the long-term average throughput. With probability $1-\epsilon$, the ϵ-greedy algorithm selects the action that has the highest empirical mean, or it selects a random action with probability ϵ. UCB maintains a confidence interval for each arm, along with the empirical means. In each round, the algorithm greedily picks the action that has the highest upper confidence bound. In TS, the rates of exploration and exploitation are dynamically updated with respect to the posterior distribution of each beam. Beams with higher estimated rewards are exploited more frequently. The author in [1] demonstrated the performance gain of his algorithms over location-based and channel estimation-based beam tracking. In [2], the authors relied on UCB to develop the Unimodal Beam Alignment (UBA) scheme for beam tracking. UBA uses the correlation between beam misalignment and RSS as contextual information to reduce the beam search space. It assigns each arm/beam a KL-UCB index. At any given time, the algorithm selects the arm that has the maximal index within the neighborhood of the arm that has the highest empirical reward. Results indicate that UBA improves the delay overhead over the exhaustive search method. Both [1] and [2] were evaluated under limited mobility and did not perform joint beam tracking and rate adaptation. In [42], the authors proposed a variation of TS for optimal rate selection over time-varying wireless channels with unknown channel statistics without considering directional communications.

In this paper, we adopt a TS approach but adapts it to non-stationary scenarios. This adaptation is necessary to account for UE mobility and/or environmental changes. Our proposed ATS algorithm is model-free and does not make assumptions regarding the underlying channel or user mobility. In each round, the online decision-making process involves solving a system of linear equations, which is easy to parallelize. Due to the ease of implementation, our algorithm does not require complex hardware, e.g., graphical processing units (GPUs). We evaluate the performance of MAMBA against the UCB and ϵ-greedy algorithms.

III. System Model

Without loss of generality, we consider tracking a single UE. Extending the treatment to multiple UEs is straightforward. We first briefly describe how beamforming is typically applied over a mmW link. We then present the MAMBA framework and formulate the reward-maximization problem.

A. Codebook-based Beamforming

Consider a directional link between a BS and a UE, implemented using electronically steerable uniform planar arrays (UPAs). Let the total number of antennas at the BS and the UE be A_{BS} and A_{UE}, respectively. Let $\mathbf{H}$ be the $A_{\text{UE}} \times A_{\text{BS}}$ complex channel matrix between them. To express the received signal, Tx and Rx beamforming should be applied to channel $\mathbf{H}$. In practice, the beamforming vectors are computed offline for a set of directions and stored in codebooks at the BS and the UE [6]. Denote the set of codebooks for the BS beamformer by $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_{D_{\text{BS}}}\}$ and for the UE beamformer by $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{D_{\text{UE}}}\}$, where D_{BS} and D_{UE} are the maximum number of narrow beams that can be generated at the BS and the UE, respectively. Assume that after IA, the BS and the UE agree on a directional link for which the BS uses its Tx beamforming vector $\mathbf{f}_i \in \mathbb{C}^{A_{\text{BS}} \times 1}$, and the UE uses its Rx beamforming vector $\mathbf{q}_j \in \mathbb{C}^{A_{\text{UE}} \times 1}$ (i and j are the indices of the Tx/Rx beamforming vectors in their respective codebooks). The received signal at time t, $y_{ij}(t)$, can then be written as:

$$y_{ij}(t) = \mathbf{q}_j^H \mathbf{H} \mathbf{f}_i s + \mathbf{q}_j^H \mathbf{z}(t) \tag{1}$$

where s is the transmitted signal and $\mathbf{z} \in \mathbb{C}^{A_{\text{UE}} \times 1}$ is a vector of complex circularly-symmetric white Gaussian noise. Each $(\mathbf{f}_i, \mathbf{q}_j)$ pair achieves a certain Rx power $P_{ij}(t)$ at time t, where $P_{ij}(t) = |y_{ij}|^2$. Because $\mathbf{H}$ is time-varying, the distribution of $P_{ij}(t)$ is nonstationary.

B. MAMBA Framework

A simple tracking strategy would exploit the current best beam pair, say $(\mathbf{f}_i, \mathbf{q}_j)$, for a relatively long time. In the 5G NR standard [43], if a new UE wishes to join the network, it waits for the BS to execute the IA procedure. During IA, the BS transmits synchronization signals (SS), allowing a listening UE to measure beam qualities and report them back to the BS. BS periodically reruns the IA to discover new UEs and update the best beams of already connected UEs. In between IA cycles, other periodic control messages, called channel state information-reference signals (CSI-RS), are transmitted by the BS to maintain communication. CSI-RS messages are used to obtain reference signal received power (RSRP) measurements for beam management during mobility. However, this can be quite wasteful, given that no data is transmitted/received during the IA phase (which typically lasts for 5 ms) or CSI-RS (which occupies up to 4 OFDM symbols). To support ultra-reliable low-latency communications (URLLC), the control overhead of beam tracking needs to be significantly decreased [44]. Our goal is to reduce this overhead by skipping CSI-RS transmissions and extending the period between two IA cycles, while maintaining connectivity. To do that, MAMBA exploits the ACK/NACK feedback obtained from the UE to make new beam selections (see Fig. 1). The ACK/NACK mechanism is already a part of the 5G NR hybrid automatic repeat request (HARQ) procedure [45]. We assume that the UE communicates using relatively wide beams so the tracking problem is mainly a concern at the BS side. This is a reasonable assumption, considering the smaller form-factor and fewer antenna elements in a UE device. Given our focus on the BS side only, in the subsequent sections, the subscript 'BS' will be dropped from related variables.

One approach to model the beam tracking problem is to use MDPs. At each time step, the MDP is in some state s and the agent (i.e., the decision maker) may choose any action that is available in that state. The MDP responds at the next time step by randomly transitioning to a new state and giving the decision maker a corresponding reward. In our beam tracking problem, after taking an action and observing a reward, the available set of actions does not change in the next time slot. Thus, the problem can be modeled as a single-state MDP, i.e., a MAB problem.

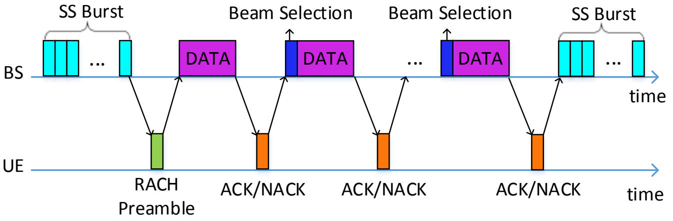


Fig. 1. Timeline of the proposed downlink communication scheme between a BS and a UE.

The MAMBA framework is specified by the tuple $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A} \triangleq \{\mathbf{f}_1, \cdots, \mathbf{f}_D\}$ is the set of actions referring to the possible BS beams at a given time and $\mathcal{R}$ is the set of rewards (i.e., achievable rates) associated with these actions. At time t, an action $a_t \in \mathcal{A}$ is taken and a reward $\mathbf{r}_t = [r_t^{(0)}, \cdots, r_t^{(M-1)}] \in \mathcal{R}$ is observed. This $\mathbf{r}_t$ is a random vector, sample drawn from the selected beam's underlying reward distribution. Let $\mathbf{\Theta}_{i,t}$ denote the reward distribution associated with beam i at time t, and let $\mathbb{E}[\mathbf{\Theta}_{i,t}] = \boldsymbol{\theta}_{i,t}$, where $\boldsymbol{\theta}_{i,t}$ is *unknown*. Note that there are D distributions in total, associated with various BS beams. With some abuse of notation we use $a_t = \mathbf{f}_i$ to mean that beamformer $\mathbf{f}_i$ is selected at time t, and hence the BS receives a reward $\mathbf{r}_t \sim \mathbf{\Theta}_{i,t}$. In MAMBA, the BS obtains the reward by measuring the RSS of ACK/NACK packets transmitted back by the UE, and determining the optimal MCS index that can be supported based on the measured RSS. Assuming channel reciprocity, the BS then uses this information to perform beam/MCS selection for the subsequent downlink data transmission.

After IA is completed, the BS designs a beam tracking policy to be used until the next IA period. A policy is defined as a T-element vector that specifies the actions to be taken at subsequent times $t = 1, \cdots, T$. The most common metric to measure the performance of a given policy is the cumulative regret, defined as the lost reward as a result of deviating from the optimal strategy. The goal of MAMBA is to find a policy that maximizes the cumulative reward, which is equivalent to minimizing the cumulative regret up to time T. We will analyze the regret performance of our policy in Section V.

Uplink Variation: In the uplink scenario, BS performs beam selection based on the RSS information gathered from the data packets transmitted by the UE. To apply MAMBA to uplink communications, the BS needs to know whether the UE is scheduled to transmit or not. This way, the BS can differentiate between two events: the UE is not transmitting or the transmitted packet is not being received due to improper beam selection. Fortunately, this information is already available at the BS thanks to the UE scheduling request (SR). SR is a control message used by a UE to ask the network for an uplink grant so that the UE can transmit data on the physical uplink shared channel (PUSCH). The SR message is transmitted over the physical uplink control channel (PUCCH) using a simple

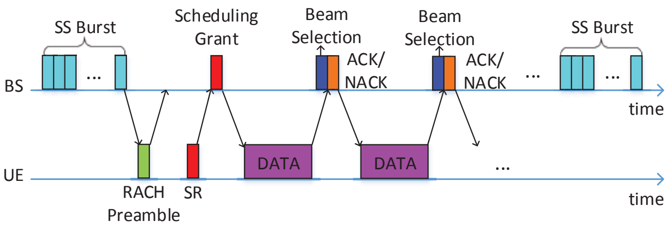


Fig. 2. Timeline of the proposed uplink communication scheme between a BS and a UE.

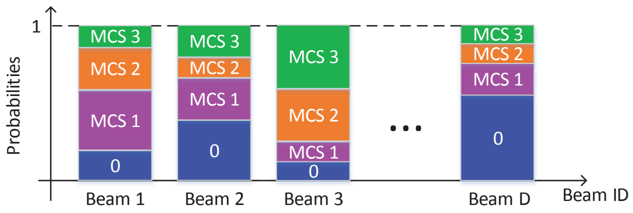


Fig. 3. Illustration of qualities of different BS beams in terms of achievable MCS indices ($M = 4$). Probabilities refer to $\theta_{i,t}^{(m)}$, $\forall i \in \mathcal{A}$ and $\forall m \in \{1, \cdots, M-1\}$ for a given t.

on-off keying, with the UE transmitting a symbol 1 to request a PUSCH resource, and transmitting nothing when it does not request to be scheduled [46]. Fig. 2 shows the proposed uplink communication timeline.

C. Problem Formulation

In MAMBA, the BS has some prior "belief" about the reward distribution of each beam, obtained through the IA. An effective method to update these beliefs during data transmission is Bayesian inference. Using Bayesian inference, the posterior distribution $\Pr(\boldsymbol{\theta}|\mathbf{x})$, i.e., the distribution of $\boldsymbol{\theta}$ given the observation $\mathbf{x}$, can be computed as:

$$\Pr(\boldsymbol{\theta}|\mathbf{x}) = \Pr(\mathbf{x}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) / \Pr(\mathbf{x}) \qquad (2)$$

where $\Pr(\mathbf{x}|\boldsymbol{\theta})$ is the *likelihood*, i.e., the distribution of the observed data, $\Pr(\boldsymbol{\theta})$ is the prior distribution, i.e., the distribution of $\boldsymbol{\theta}$ before any data is observed, and $\Pr(\mathbf{x})$ is the marginal distribution of the evidence, which normalizes the posterior distribution. Using (2), the BS continuously updates its belief of each arm's mean rewards, i.e., $\boldsymbol{\theta}_{i,t}$, $\forall i \in \mathcal{A}$ and $\forall t \in \{1, \cdots, T\}$, while transmitting/receiving data.

In our setup, the rewards are modeled as M-dimensional variables. For each transmission, the BS chooses a beam as well as a transmission rate for that beam from the set $\{v_0, v_1, \cdots, v_{M-1}\}$. Specifically, for a given beam, the BS can establish communication with the UE using one of the $M-1$ available MCS indices, each of which has an associated rate v_m, $m \in \{1, \cdots, M-1\}$, or it cannot establish any communication, i.e., $v_0 = 0$. Based on the feedback received from the UE (i.e., ACK, NACK, or no reply), the BS decides whether the selected rate is attainable on the selected beam or not. If an ACK or NACK is received, the BS measures the RSS of the received packet and determines the MCS index that can be supported over that beam. If neither an ACK nor a NACK is received, the reward is set to 0.

The expected reward for each beam is drawn from a likelihood distribution associated with that beam. A suitable reward distribution to be used here is the categorical distribution, a.k.a., generalized Bernoulli distribution. This discrete distribution describes the possible results of a random variable that can take one of M possible categories, with the probability of each category separately specified. Note that such a distribution is quite general and can fit any underlying empirical distribution by properly setting the mean values for each associated category/event. Given that the reward for each beam should follow a discrete event (as the reward represents the throughput associated with one of several MCS indices), the generalized Bernoulli distribution can be applied, irrespective of the underlying channel model.

The pmf of the categorical random variable $x \sim \text{Cat}(\boldsymbol{\theta}_{i,t})$ with M categories can be written as $\Pr(x = m|\boldsymbol{\theta}_{i,t}) = \theta_{i,t}^{(m)}$, where $\boldsymbol{\theta}_{i,t} \triangleq [\theta_{i,t}^{(0)}, \cdots, \theta_{i,t}^{(M-1)}]$. Here, $\theta_{i,t}^{(m)}$ refers to the mth element of vector $\boldsymbol{\theta}_{i,t}$, such that $\theta_{i,t}^{(m)} \geq 0$ $\forall m$ and $\sum_{m=0}^{M-1} \theta_{i,t}^{(m)} = 1$. An illustrative example is shown in Fig. 3 using $M = 4$. Each beam $i \in \mathcal{A}$ has an associated mean reward vector $\boldsymbol{\theta}_{i,t}$.

At any time t, the observed reward vector $\hat{\mathbf{r}_t} = [\hat{r_t}^{(0)}, \cdots, \hat{r_t}^{(M-1)}]$ contains a single 1 at the highest attainable MCS index (based on the RSS of ACK/NACK packets) and 0's elsewhere. For convenience, we assign $\hat{r_t}^{(0)} = 1$ for an unsuccessful communication and $\hat{r_t}^{(m)} = 1$ for a communication whose highest attainable MCS index is m, $\forall m \in \{1, \cdots, M-1\}$. Therefore, the observed data rate at time t can be written as $\hat{\mathbf{r}_t}\mathbf{v}^T$, where $\mathbf{v} \triangleq [v_0, v_1, \cdots v_{M-1}]$ is the value vector whose entries correspond to the rates associated with different MCS indices ($v_0 \triangleq 0$).

Given the above, the goal of the BS is to select a policy $\boldsymbol{\xi} = [a_1, \cdots, a_T]$, i.e., sequence of Tx beams at times $t = 1, \cdots, T$, that maximizes the expected throughput. If the expected reward vectors $\boldsymbol{\theta}_{i,t} = [\theta_{i,t}^{(0)}, \cdots, \theta_{i,t}^{(M-1)}]$ of each beam i at each time t are known, this translates into solving the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\xi}}{\text{maximize}} \quad & \sum_{t=1}^{T} \boldsymbol{\theta}_{i,t}\mathbf{v}^T \\ \text{s.t.} \quad & \sum_{m=0}^{M-1} \theta_{i,t}^{(m)} = 1, \quad \theta_{i,t}^{(m)} \geq 0, \quad \forall i, t, m. \end{aligned} \qquad (3)$$

The challenge here is that the expected reward vectors are unknown and nonstationary. As a result, we cannot solve (3) directly. Our goal is to design an RL algorithm that learns the expected rewards of different beams and outputs a policy that converges to the optimal one.

IV. Proposed Beam Tracking and MCS Selection Algorithm

In this section, we explain our TS-based MAMBA algorithm used by a BS. The process of adapting the BS beam should be seamless from the UE's perspective, i.e., the UE should not be required to know about BS beam switching and should not expect control packets regarding that. The flowchart of

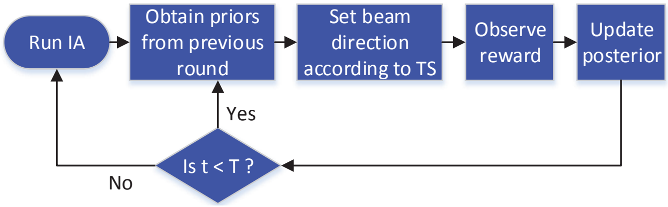
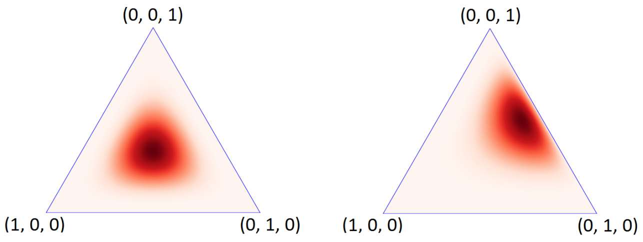

Run IA
Obtain priors from previous round
Set beam direction according to TS
Observe reward
Update posterior
Is t < T ?
Yes
No

Fig. 4. Flowchart of the proposed beam tracking and MCS selection at the BS.



(a) (b)

Fig. 5. Visualization of 3D Dirichlet distributions as a heatmap, where darker areas denote higher probabilities and lighter areas denote lower probabilities. (a) $\boldsymbol{\alpha}_{i,t} = [5, 5, 5]$, (b) $\boldsymbol{\alpha}_{i,t} = [2, 5, 6]$.

the proposed method is shown in Fig. 4. We first consider a stationary system where the expected rewards of various arms do not change in time. We then extend our treatment to time-varying systems.

TS is a posterior sampling technique. Therefore, before taking an observation, we need a suitable prior that represents our belief on an arm's reward. Because the reward distribution of arm i is modeled as a categorical distribution and the Dirichlet distribution is the conjugate prior of this distribution, we model the prior of the expected rewards as a Dirichlet distribution with parameter $\boldsymbol{\alpha}_{i,t}$, $\text{Dir}(\boldsymbol{\alpha}_{i,t})$. As a result, the posterior obtained at each round is also a Dirichlet distribution, following (2).

The Dirichlet distribution is a multivariate generalization of the beta distribution. The set of points in the support of an M-dimensional Dirichlet distribution is the standard $(M-1)$-simplex. For $M = 3$, the support is an equilateral triangle with vertices at $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. The pdfs of two example 3D Dirichlet distributions are shown in Fig. 5.

At each round, after an action is taken, a reward is observed and the posterior distribution is updated according to (2). When the prior is the conjugate distribution of the likelihood, the update rule is much simpler. Specifically, for the case with $\text{Cat}(\boldsymbol{\theta}_{i,t})$ rewards and $\text{Dir}(\boldsymbol{\alpha}_{i,t})$ priors $\forall i \in \mathcal{A}$, the update rule for the posterior is as follows:

$$\boldsymbol{\alpha}_{i,t+1} = \begin{cases} \boldsymbol{\alpha}_{i,t} + \hat{\mathbf{r}}_t, & \text{if } a_t = i \\ \boldsymbol{\alpha}_{i,t}, & \text{if } a_t \neq i. \end{cases}$$

The first case ($a_t = i$) is when beam i is selected for transmission at time t and a reward $\hat{\mathbf{r}}_t$ is observed. The posterior distribution of beam i is then updated accordingly. The second case is when beam i is not selected for transmission at time t, and thus, its posterior is not changed.

Algorithm 1 Thompson Sampling

```
for $t = 1, 2, \cdots, T$ do
    *Take Samples*:
        for $i \in \mathcal{A}$ do
            Sample $\mathbf{s}_{i,t} \sim \text{Dir}(\boldsymbol{\alpha}_{i,t})$
    *Choose and Apply Action*:
        $a_t = \text{argmax}_{i \in \mathcal{A}} \mathbf{s}_{i,t}\mathbf{v}^T$
        Select $a_t$ and observe $\hat{\mathbf{r}}_t$
    *Update Distributions*:
        for $i \in \mathcal{A}$ do
            if $a_t = i$ then
                $\boldsymbol{\alpha}_{i,t+1} \leftarrow \boldsymbol{\alpha}_{i,t} + \hat{\mathbf{r}}_t$
            else
                $\boldsymbol{\alpha}_{i,t+1} \leftarrow \boldsymbol{\alpha}_{i,t}$
```

After the distributions are updated, the arm to be selected for the next round is determined based on random samples taken from the current posterior distributions of the arms. Specifically, at each time t, the BS samples from each arm's updated distribution to obtain $\mathbf{s}_{i,t} \sim \text{Dir}(\boldsymbol{\alpha}_{i,t})$, $\forall i \in \mathcal{A}$, and selects the action as:

$$a_t = \underset{i \in \mathcal{A}}{\text{argmax}}\, \mathbf{s}_{i,t}\mathbf{v}^T. \tag{4}$$

Therefore, even though the arms with currently high estimated means are more likely to be selected, other arms also get a chance to be picked and updated, i.e., exploration versus exploitation. This is called the Thompson sampling and its pseudocode is provided in Algorithm 1. Note that $|\mathbf{s}_{i,t}| = 1$, $\forall i \in \mathcal{A}$, $\forall t \in \{1, \cdots, T\}$.

Algorithm 1 works well in stationary scenarios, where beam qualities do not change over time. However, for nonstationary scenarios, we need to adapt this algorithm.

A. Adaptive Thompson Sampling (ATS) Algorithm

In nonstationary scenarios, the algorithm should never stop exploring, since it needs to keep track of changes. With some modification, TS remains an effective approach, as long as the channel characteristics change relatively slowly.

To address nonstationary scenarios, we model the evolution of the belief distributions in a way that *discounts* the relevance of past observations and increases the impact of recent observations. This can be done by implementing a “forgetting” factor γ_1 that slowly alters the posterior distributions and a “boost” factor γ_2. For $i \in \mathcal{A}$, the update rule is now written as:

$$\boldsymbol{\alpha}_{i,t+1} = \begin{cases} \gamma_1\boldsymbol{\alpha}_{i,t} + \gamma_2\hat{\mathbf{r}}_t, & \text{if } a_t = i \\ \gamma_1\boldsymbol{\alpha}_{i,t}, & \text{if } a_t \neq i \text{ and } \max\{\gamma_1\boldsymbol{\alpha}_{i,t}\} > 1 \\ \mathbf{1}, & \text{otherwise.} \end{cases}$$

Here, the operation $\max\{\gamma_1\boldsymbol{\alpha}_{i,t}\}$ returns the largest element of the vector $\gamma_1\boldsymbol{\alpha}_{i,t}$. Note that multiplying $\boldsymbol{\alpha}_{i,t}$ by a constant γ_1 effectively increases the variance (given that $0 < \gamma_1 < 1$),

Algorithm 2 Adaptive Thompson Sampling

```
for $t = 1, 2, \cdots, T$ do
  *Take Samples*:
    for $i \in \mathcal{A}$ do
      Sample $\mathbf{s}_{i,t} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{i,t})$
  *Choose and Apply Action*:
    $a_t = \mathrm{argmax}_{i \in \mathcal{A}}\, \mathbf{s}_{i,t}\mathbf{v}^T$
    Select $a_t$ and observe $\hat{\mathbf{r}_t}$
  *Update Distributions*:
    for $i \in \mathcal{A}$ do
      if $a_t = i$ then
        $\boldsymbol{\alpha}_{i,t+1} \leftarrow \gamma_1 \boldsymbol{\alpha}_{i,t} + \gamma_2 \hat{\mathbf{r}_t}$
      else if $a_t \neq i$ and $\max\{\gamma_1 \boldsymbol{\alpha}_{i,t}\} > 1$ then
        $\boldsymbol{\alpha}_{i,t+1} \leftarrow \gamma_1 \boldsymbol{\alpha}_{i,t}$
      else
        $\boldsymbol{\alpha}_{i,t+1} \leftarrow \mathbf{1}$
```

but does not alter the mean of the Dirichlet distribution. To show that, we first calculate $\boldsymbol{\mu}_{i,t+1} \triangleq \mathbb{E}[\mathrm{Dir}(\boldsymbol{\alpha}_{i,t+1})]$:

$$\begin{aligned}\boldsymbol{\mu}_{i,t+1} &= \left[\frac{\gamma_1 \alpha_{i,t}^{(0)}}{\sum_{j=0}^{M-1} \gamma_1 \alpha_{i,t}^{(j)}}, \cdots, \frac{\gamma_1 \alpha_{i,t}^{(M-1)}}{\sum_{j=0}^{M-1} \gamma_1 \alpha_{i,t}^{(j)}}\right] \\ &= \left[\frac{\alpha_{i,t}^{(0)}}{\sum_{j=0}^{M-1} \alpha_{i,t}^{(j)}}, \cdots, \frac{\alpha_{i,t}^{(M-1)}}{\sum_{j=0}^{M-1} \alpha_{i,t}^{(j)}}\right] = \boldsymbol{\mu}_{i,t}.\end{aligned}$$

Next, we calculate $\boldsymbol{\sigma}_{i,t+1}^2 \triangleq \mathrm{Var}[\mathrm{Dir}(\boldsymbol{\alpha}_{i,t+1})]$:

$$\begin{aligned}\boldsymbol{\sigma}_{i,t+1}^2 &= \left[\frac{\mu_{i,t+1}^{(0)}(1 - \mu_{i,t+1}^{(0)})}{1 + \sum_{j=0}^{M-1} \gamma_1 \alpha_{i,t}^{(j)}}, \cdots, \frac{\mu_{i,t+1}^{(M-1)}(1 - \mu_{i,t+1}^{(M-1)})}{1 + \sum_{j=0}^{M-1} \gamma_1 \alpha_{i,t}^{(j)}}\right] \\ &= \left[\frac{\mu_{i,t}^{(0)}(1 - \mu_{i,t}^{(0)})}{1 + \gamma_1 \sum_{j=0}^{M-1} \alpha_{i,t}^{(j)}}, \cdots, \frac{\mu_{i,t}^{(M-1)}(1 - \mu_{i,t}^{(M-1)})}{1 + \gamma_1 \sum_{j=0}^{M-1} \alpha_{i,t}^{(j)}}\right] > \boldsymbol{\sigma}_{i,t}^2\end{aligned}$$

for $0 < \gamma_1 < 1$. Thus, the variances of the unexplored arms increase at each iteration. Note that the effects of γ_1 and γ_2 are different. Specifically, γ_1 determines the rate at which the prior information is forgotten, whereas γ_2 determines how much the new information is valued. Finally, the last condition ensures that if arm i has not been selected for a long time, $\boldsymbol{\alpha}_{i,t+1}$ is updated in a way that our belief on arm i's distribution converges to a multi-dimensional uniform distribution, i.e., $\mathrm{Dir}(\mathbf{1})$. We incorporate this new update rule into an algorithm called ATS (see Algorithm 2).

Prior Selection: An important design issue is the initialization of prior distributions right after IA. In general, a uniform prior works well with most TS algorithms. For our problem formulation, this would correspond to $\mathrm{Dir}(\mathbf{1})$. However, this prior ignores any useful knowledge obtained through IA. Taking past knowledge into account and choosing an informative prior reduce what must be newly learned. Specifically, if the best MCS index that beam i can satisfy during IA is m, we assign $\alpha_{i,0}^{(m)} = P$ and $\alpha_{i,0}^{(j)} = 1, \forall j \in \{0, \cdots, M-1\}, j \neq m$. Here, $P \geq 1$ is an adjustable design parameter called the *prior strength*. By selecting informative prior parameters $\boldsymbol{\alpha}_{i,0}$ according to IA, the convergence time can be reduced and the average data rate can be significantly improved, as we show in Section VI.

Selection of γ_1 and γ_2: The BS needs to select appropriate values for γ_1 and γ_2 before running Algorithm 2. For this purpose, we let the BS estimate the distances of the UEs through the RSS of the received packets, and calculate the optimum γ_1 and γ_2 offline for each UE. RSS-based distance estimation has been widely applied in sub-6 GHz wireless systems (e.g., [47]–[49]). For distance estimation at mmW spectrum, we use a commonly employed path-loss model:

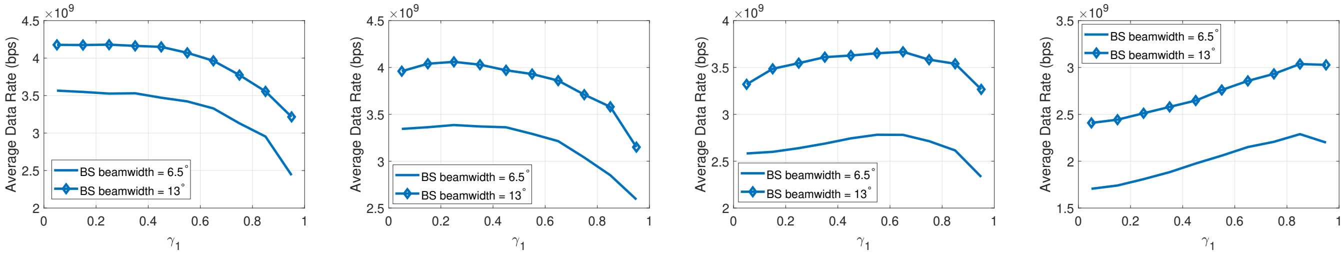
$$PL(d)[\mathrm{dB}] = c_1 + c_2 10 \log_{10}(d) + \psi, \;\; \psi \sim \mathcal{N}(0, \sigma_n^2) \quad (5)$$

where d is the BS-UE distance in meters, c_1 and c_2 are the floating intercept and slope of the model (obtained via regression of measured data), respectively, and σ_n^2 is the log-normal shadowing variance. This model has been adopted by many researchers (see [11], [32], for example). It does not consider the (directional) beamforming pattern at the Tx/Rx or the small-scale (multi-path) effects. Yet, as shown later, it can still provide good estimates of γ_1 and γ_2 that result in improved throughput. Note that for the simulation-based evaluation of MAMBA, we rely on a much detailed channel model that accounts for antenna directionality, clustering, and multi-path effects, as discussed in Section VI-B.

Knowing the UE transmit power and the RSS it observes, the BS can easily compute the path loss and solve (5) for d. Our intuition here is that UEs that are closer to the BS are more likely to switch between beams, as even small displacements can result in large angular changes. Conversely, when the UE is further away from the BS, it is more likely to be served by the same beam for a long period of time. With this intuition, we conduct simulations to study the effect of distance on the optimal γ_1 and γ_2, for two different beamwidths (the values of other simulation parameters are provided in Section VI-B). In Fig. 6, we can observe that as the distance increases, the optimum γ_1 also increases, since the angular mismatch between the UE and the BS beam does not change quickly. Note that the average data rate decreases with d, due to lower received power at the UE. In addition, the trend is the same for different beamwidths, which makes us conclude that the optimum γ_1 does not depend on the beamwidth. On the other hand, ATS is not as sensitive to changes in γ_2 as it is to changes in γ_1, once it exceeds a certain value. Specifically, as seen in Fig. 7(a), for $\gamma_2 > 30$, the average data rate remains approximately the same.

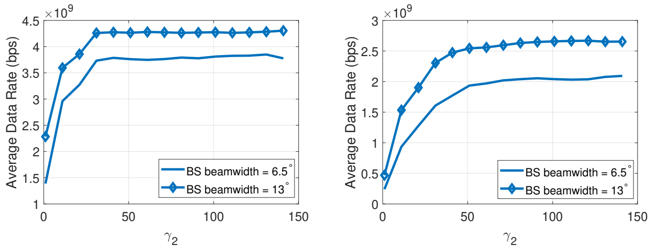
B. Rate Selection

After a beam has been selected via ATS, the BS needs to determine an appropriate MCS to be used during data transmission. MCS selection is particularly important, as the effective data rate of a given transmission would be 0 if the MCS that the BS selects cannot be supported at the UE. Conversely, if the BS selects a lower MCS than the maximum one that the UE can support, the link would be underutilized. Taking this tradeoff into account, we propose two techniques for MCS selection: greedy and conservative.



(a) (b) (c) (d)

Fig. 6. Effect of γ_1 on ATS performance for different distances and beamwidths. (a) $d = 50$ m, (b) $d = 100$ m, (c) $d = 150$ m, (d) $d = 200$ m.



(a) (b)

Fig. 7. Effect of γ_2 on ATS performance for different distances and beamwidths. (a) $d = 50$ m, (b) $d = 200$ m.

Greedy MCS Selection: Here, the MCS index that attains the maximum *expected* rate over the selected beam is used. Specifically, after sampling $\mathbf{s}_{i,t}, \forall i \in \mathcal{A}$, and taking action a_t, the MCS index m^* is selected as:

$$m^* = \underset{m\in\{0,\cdots,M-1\}}{\operatorname{argmax}} \quad s_{a_t,t}^{(m)} v^{(m)}. \tag{6}$$

Therefore, even when an MCS index is less likely to be attained than others, depending on $\mathbf{v}$, the BS may decide to choose it due to its higher associated rate.

Conservative MCS Selection: In this scheme, the MCS index that is most likely to be attained and that can achieve a non-zero rate over the selected beam is used for transmission. In other words, after the BS collects $\mathbf{s}_{i,t}, \forall i \in \mathcal{A}$, and selects the beam, it will select a transmission rate based on the probabilities of attaining different MCS indices on the selected beam. Specifically, given action a_t, the selected MCS index is given by:

$$m^* = \underset{m\in\{0,\cdots,M-1\}}{\operatorname{argmax}} \quad s_{a_t,t}^{(m)}. \tag{7}$$

V. Regret Analysis

In this section, we compute an upper bound on the Bayesian regret of ATS. Let $\mathcal{I}$ denote an instance of the MAB problem drawn initially from some known distribution $\mathbb{P}$ (a.k.a. the prior) over a set of possible problem instances. A problem instance is specified by $\boldsymbol{\theta}_{i,t}$ $\forall i \in \mathcal{A}$ and $\forall t \in \{1, 2, \cdots\}$. (For a stationary bandit problem, in which $\boldsymbol{\theta}_{i,1} = \boldsymbol{\theta}_{i,2} = \cdots = \boldsymbol{\theta}_{i,t}$ $\forall i \in \mathcal{A}$ and $\forall t \in \{1, 2, \cdots\}$, a problem instance is specified only by $\boldsymbol{\theta}_i$, $\forall i \in \mathcal{A}$.) Then, Bayesian regret within a time horizon of T is defined as:

$$\mathrm{BR}(T) = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{I}\sim\mathbb{P}} \left[\mathbb{E}\left[\boldsymbol{\theta}_{a_t^*,t}\mathbf{v}^T - \boldsymbol{\theta}_{a_t,t}\mathbf{v}^T \mid \mathcal{I} \right] \right] \tag{8}$$

where $\boldsymbol{\theta}_{a_t,t}$ denotes the expected reward vector of the action selected by our algorithm at time t and $a_t^* = \operatorname{argmax}_{i\in\mathcal{A}} \boldsymbol{\theta}_{i,t}\mathbf{v}^T$. The inner expectation in (8) is the expected regret for a given problem instance $\mathcal{I}$, and the outer expectation is over the set of all problem instances. Let BR_t denote the instantaneous regret at time t, i.e., $\mathrm{BR}_t = \mathbb{E}[\boldsymbol{\theta}_{a_t^*,t}\mathbf{v}^T - \boldsymbol{\theta}_{a_t,t}\mathbf{v}^T]$, where inner and outer expectations in (8) are merged into a single expectation. Then, BR_t can be also written as:

$$\mathrm{BR}_t = \sum_{m=0}^{M-1} \mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)} \right] v_m. \tag{9}$$

Now, we focus on providing a theoretical bound on $\mathbb{E}[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}]$, $\forall m \in \mathcal{M}$. We use a random walk process to model the nonstationarity of the rewards obtained from various beams [50] (note that the MAMBA scheme itself does not rely on this model). Specifically, the expected reward vector of each beam follows a discrete-time random walk in an $(M-1)$-dimensional space with reflecting boundaries. We assume that the step sizes $\epsilon_{i,t}$ of this walk at each time interval t are uniformly distributed: $\epsilon_{i,t} \sim \mathcal{U}[0, \sigma]$ $\forall i \in \mathcal{A}$ and $\forall t \geq 0$. Here, σ denotes the maximum step size, which is also called the *volatility* of an arm in MAB context [50]. The direction of the walk is also determined by a uniform distribution within all the possible directions in $(M-1)$ dimensions. See Fig. 5 for a visualization of this model when $M = 3$. Let $\boldsymbol{\omega}_{i,t} \in \mathbb{R}^{1\times 3}$ denote the unit vector towards the selected step direction. Given the triangle in Fig. 5, whose corners are located on the x-, y- and z-axes, if $\boldsymbol{\theta}_{i,t} + \epsilon_{i,t}\boldsymbol{\omega}_{i,t}$ does not hit the edge, then $\boldsymbol{\theta}_{i,t+1} - \boldsymbol{\theta}_{i,t} = \epsilon_{i,t}\boldsymbol{\omega}_{i,t}$. Otherwise, $|\boldsymbol{\theta}_{i,t+1} - \boldsymbol{\theta}_{i,t}| \leq \epsilon_{i,t}$ due to the reflecting boundaries (where $|.|$ denotes the length of a vector).

Let $S_{i,t}^{(m)}$ denote the empirical summation of the rewards observed when using beam i and MCS index m from time 0 up to time t. Also, let $n_{i,t}$ denote the number of times beam i is selected up to time t, based on our ATS algorithm. Note that when beam i is selected at time t, we observe a reward vector $\mathbf{r}_t$, which includes rewards of all MCS indices belonging to that beam (1 or 0). That is, for each MCS index, there are

$n_{i,t}$ observations. Accordingly, $S_{i,t}^{(m)} = \gamma_2 \sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)}$ where $\tau_{i,j}$ denotes the time of the jth selection of beam i. Then, the expected value of $S_{i,t}^{(m)}$ is given by $\mathbb{E}[S_{i,t}^{(m)}] = \gamma_2 \sum_{k=1}^{n_{i,t}} \mathbb{E}[\gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)}] = \gamma_2 \sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} \theta_{i,\tau_{i,k}}$.

Lemma 1: For a given beam i and $t \leq T$,

$$\Pr\left(\left|\theta_{i,t}^{(m)} - \gamma_1^t \theta_{i,0}^{(m)}\right| \geq \min\{1,\sigma\}\sqrt{8T\log T}\right) = \mathcal{O}(T^{-4}).$$

Proof: To simplify the proof, we drop from the notation the MCS index m. Let $X_n = \gamma_1^{T-n}\theta_{i,n}$, $n = 0, 1, \cdots, T$, denote a sequence of random variables. This sequence is a supermartingale, as $\mathbb{E}[X_{n+1}|X_0, X_1, \cdots, X_n] \leq X_n$, $n = 0, 1, \cdots, T-1$ (recall that $\gamma_1 < 1$). Therefore, we can apply Azuma-Hoeffding inequality as in Claim 3.6 of [50]. First, it is clear that $|X_{n+1} - X_n| < \min\{1,\sigma\}$ almost surely. Following Azuma-Hoeffding inequality, $\Pr(|\theta_{i,t} - \gamma_1^t\theta_{i,0}| \geq \min\{1,\sigma\}\sqrt{8T\log T}) \leq \Pr(|\theta_{i,T} - \gamma_1^T\theta_{i,0}| \geq \min\{1,\sigma\}\sqrt{8T\log T}) \leq 2T^{-4} = \mathcal{O}(T^{-4})$. ■

Lemma 2: Let $\hat{\theta}_{i,t}^{(m)} \triangleq \sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)}/n_{i,t}$ denote our empirical estimate of $\theta_{i,t}^{(m)}$. Then,

$$\Pr\left(\left|\hat{\theta}_{i,t}^{(m)} - \theta_{i,t}^{(m)}\right| \geq \delta_{i,t}\right) = \mathcal{O}(T^{-4}) \tag{10}$$

where $\delta_{i,t} = \sqrt{2\log T/n_{i,t}} + \min\{1,\sigma\}\sqrt{8T\log T}$ and $t \leq T$.

Proof: We utilize Azuma-Hoeffding inequality to prove this lemma. Let $Y_k = \gamma_2\gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)}$, $k = 1, \cdots, n_{i,t}$, denote each term in $S_{i,t}$, which consists of independent random variables that are strictly bounded by the interval $[0, \gamma_2]$. Following Azuma-Hoeffding inequality, we obtain (11). In the next step, each term of the inequality inside the probability expression is divided by γ_2. (13) follows from Lemma 1. Finally, by dividing the terms of the inequality inside the probability expression by $n_{i,t}$, we obtain (10). ■

In the rest of the analysis, we exploit similar techniques to bound the Bayesian regret as in [51]. Given a problem instance $\mathcal{I}$, let a history H_t denote all selected beams of our algorithm and the corresponding observed rewards up to time t, i.e., a particular run of the algorithm. Given this history, let $U_t^{(m)}(i)$ and $L_t^{(m)}(i)$ denote the upper and lower confidence bounds on action i's expected reward at time t for MCS index m, respectively, such that:

$$U_t^{(m)}(i) = \hat{\theta}_{i,t}^{(m)} + \delta_{i,t} \quad \text{and} \quad L_t^{(m)}(i) = \hat{\theta}_{i,t}^{(m)} - \delta_{i,t}. \tag{14}$$

Lemma 3: For any $t \leq T$,

$$\mathrm{BR}_t \leq 2Mv_{M-1}\mathbb{E}\left[\delta_{a_t,t}\right] + \mathcal{O}(T^{-4}). \tag{15}$$

Proof: Conditioned on a certain history H_t, the optimal action a_t^* and the action a_t (selected by ATS) are identically distributed, and $U_t^{(m)}(a_t^*) = U_t^{(m)}(a_t)$ (please refer to Proposition 1 in [51] for further details). Hence,

$$\begin{aligned}
&\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}\right] = \mathbb{E}_{H_t}\left[\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}|H_t\right]\right] \\
&= \mathbb{E}_{H_t}\left[\mathbb{E}\left[U_t^{(m)}(a_t) - U_t^{(m)}(a_t^*) + \theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}|H_t\right]\right] \\
&= \mathbb{E}_{H_t}\left[\mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}|H_t\right] + \mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)|H_t\right]\right] \\
&= \mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] + \mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right]. \qquad (16)
\end{aligned}$$

We separately investigate the two terms in (16). Let $(a)^+ \triangleq \max\{0, a\}$ for any real number a. First, consider the second term in (16):

$$\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right] \leq \mathbb{E}\left[\left(\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right)^+\right] \tag{17}$$

$$\leq \Pr\left(\theta_{a_t^*,t}^{(m)} \geq U_t^{(m)}(a_t^*)\right) = \mathcal{O}(T^{-4}). \tag{18}$$

The first inequality in (18) follows from the fact that the largest possible value for $(\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*))^+$ is 1. The second inequality in (18) is due to Lemma 2. Now, consider the first term in (16):

$$\begin{aligned}
\mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] &= \mathbb{E}\left[2\delta_{a_t,t} + L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] \\
&= 2\mathbb{E}\left[\delta_{a_t,t}\right] + \mathbb{E}\left[L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right]. \qquad (19)
\end{aligned}$$

Similar to (17) and (18):

$$\begin{aligned}
\mathbb{E}\left[L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] &\leq \mathbb{E}\left[\left(L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right)^+\right] \\
&\leq \Pr\left(\theta_{a_t,t}^{(m)} \leq L_t^{(m)}(a_t)\right) = \mathcal{O}(T^{-4}).
\end{aligned}$$

Combining (9) and (16):

$$\begin{aligned}
\mathrm{BR}_t &= \sum_{m=0}^{M-1}\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}\right]v_m \leq M\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}\right]v_{M-1} \\
&\leq 2Mv_{M-1}\mathbb{E}\left[\delta_{a_t,t}\right] + \mathcal{O}(T^{-4}).
\end{aligned}$$

■

Theorem 4: Given D beams, each of which has a volatility σ, the Bayesian regret of the ATS algorithm over a time horizon T is bounded by:

$$\mathrm{BR}(T) = \mathcal{O}\left(M\sqrt{DT\log T} + M\min\left\{T, \sigma T\sqrt{8T\log T}\right\}\right).$$

Proof: We know that $\mathrm{BR}(T) = \sum_{t=1}^{T}\mathrm{BR}_t$. Hence, following Lemma 3:

$$\begin{aligned}
\mathrm{BR}(T) &\leq \left(2Mv_{M-1}\mathbb{E}\left[\delta_{a_t,t}\right] + \mathcal{O}\left(T^{-4}\right)\right) \qquad (20) \\
&= \mathcal{O}\left(M\sqrt{\log T}\right)\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{1/n_{a_t,t}}\right] + \\
&\qquad \mathcal{O}\left(M\min\{1,\sigma\}T\sqrt{8T\log T}\right). \qquad (21)
\end{aligned}$$

$$\Pr\left(\left|S_{i,t}^{(m)} - \mathbb{E}[S_{i,t}^{(m)}]\right| \geq \gamma_2\sqrt{2n_{i,t}\log T}\right) = \mathcal{O}(T^{-4}) \tag{11}$$

$$\Pr\left(\left|\sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)} - \sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} \theta_{i,\tau_{i,k}}^{(m)}\right| \geq \sqrt{2n_{i,t}\log T}\right) = \mathcal{O}(T^{-4}) \tag{12}$$

$$\Pr\left(\left|\sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)} - n_{i,t}\theta_{i,t}^{(m)}\right| \geq \sqrt{2n_{i,t}\log T} + n_{i,t}\min\{1,\sigma\}\sqrt{8T\log T}\right) = \mathcal{O}(T^{-4}) \tag{13}$$

Lemma 1 in [51] states that $\mathbb{E}\left[\sqrt{1/n_{a_t,t}}\right] = \mathcal{O}(\sqrt{DT})$. Furthermore, when the rewards are bounded by the interval $[0,1]$, the maximum possible regret within a time horizon T is T. Thus, the second term in (21) is bounded by MT. ■

Theorem 4 states that if σ is relatively low, the regret scales with $\sqrt{T\log T}$. Note that the authors in [51] prove that the regret of a *stationary* system also scales with $\sqrt{T\log T}$. Therefore, when σ is low, ATS can alleviate the affect of nonstationarity. On the other hand, if σ is large, the worst-case regret scales linearly with T.

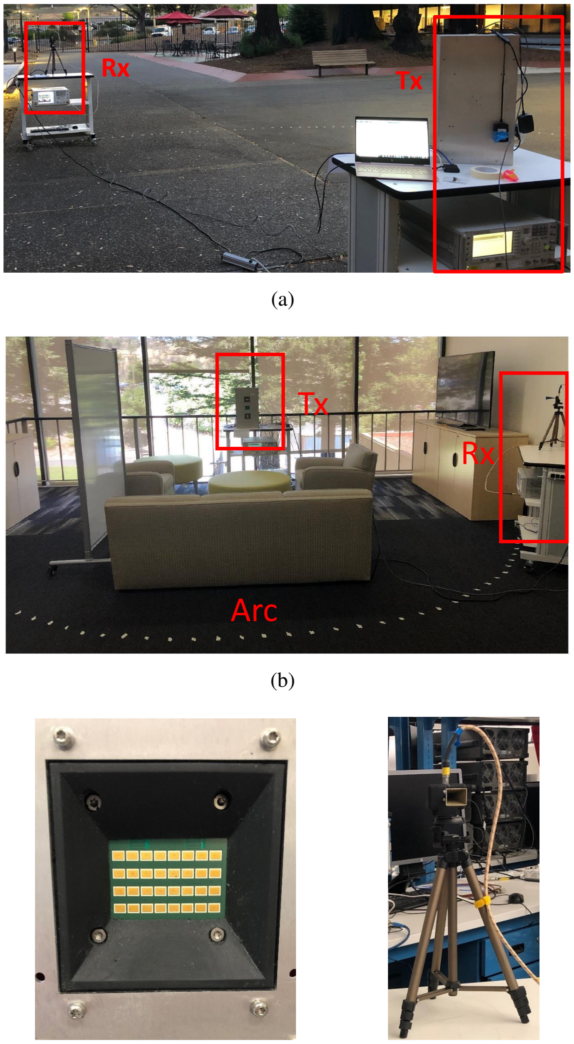
VI. Performance Evaluation

We evaluate the performance of MAMBA through OTA experiments and simulations, considering a single phase of beam selection. We study both fixed- and adaptive-rate variants of MAMBA, and contrast them with four other beam tracking strategies: *static oracle*, *dynamic oracle*, the ϵ-greedy algorithm [1], and the UBA algorithm [2]. The static oracle is essentially the same as Phase 1 of the 5G NR beam selection scheme. It runs an exhaustive beam search every time an outage occurs. The dynamic oracle is a theoretically optimal but practically infeasible beam tracking scheme. It always selects the best beam/MCS combination at each slot, obtained through exhaustive search. The performance of this scheme is meant to provide an upper bound on practical tracking algorithms.

A. Experimental Results

For the experimental setup, we consider a mmW link (see Fig. 8) in which the Tx is comprised of a Keysight E8267D PSG signal generator that connects to a 15 dBi 4-by-8 UPA. The Tx transmits a continuous wave (CW) 28 GHz signal at 0 dBm transmit power. The Rx consists of a Keysight 9038A MXE EMI receiver, connected to a 20 dBi horn antenna. Both the PSG and the EMI receiver are connected to the host PC through USB ports, allowing them to send/receive SCPI commands through a serial connection.

To simulate the effect of ACK/NACK, the Rx measures the RSS and stores it in a variable in the host PC. The Tx can then read this RSS value and obtain a reward by determining the most appropriate MCS through table lookup. The mapping from an RSS (or SNR) value to an MCS index (or sometimes, to a Channel Quality Indicator (CQI)) depends on the specific implementation of the receiver, e.g., the iterative decoder. Such mapping, typically provided by vendors, can be used to obtain the spectral efficiency (in bits/sec/Hz) for each MCS index, and consequently determine the data rate that can be supported by that MCS index; see, for example, the mappings in [4] for WiGig and in [52] (page 20) for 5G NR PDSCH messages. For our experiments, we use the mapping in [4].



(a)

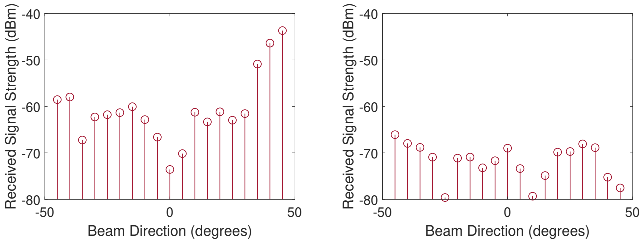
(b)

(c)

(d)

Fig. 8. Experimental setup used for performance evaluation. (a) Outdoor scenario with 7m Tx-Rx separation, (b) indoor office scenario with 3.5m Tx-Rx seperation, (c) 4 × 8 UPA at the Tx side, (d) 20 dBi gain horn antenna at the Rx side.

We conduct the experiments in two scenarios: an outdoor scenario (Fig. 8(a)) with Tx-Rx separation of 7 meters, and an indoor office environment (Fig. 8(b)) with Tx-Rx separation of 3.5 meters. For reproducibility of the experiments, RSS measurements are taken at discrete, equally spaced points on an arc that is centered at the Tx. For the outdoor (indoor)



(a)

(b)

Fig. 9. RSS for different beam directions in the indoor office scenario when: (a) Rx is located at the rightmost corner of the office with direct LOS path, and (b) Rx is located at the left side of the office, blocked by the whiteboard.

scenario, the distance between two consecutive points on the arc is 15 cm (30 cm), resulting in 72 (41) measurement points. A LOS path is always available at each measurement point in the outdoor scenario. Beam selection is performed only at the Tx side, while the Rx is always pointed optimally towards the Tx. In practice, a UE is equipped with a small array, limiting its ability to achieve precise beam pointing, i.e., the antenna's boresight will not be perfectly aligned with the line-of-sight (LOS) between the Rx and Tx. Nonetheless, considering that UE beams are generally much wider than the beams at the BS, beam selection is less of an issue at the UE side. As the Tx and Rx antennas are at the same elevation, beam sweeping at the Tx is done only in the azimuthal plane by steering the UPA $\pm 45^\circ$ around the broadside of the UPA in 5° steps. This results in 19 possible Tx beams at each Rx location. We assume saturated downlink traffic, i.e., the Tx always has packets to send. A detailed description of our experimental setups along with the datasets are available at [53].

The indoor office environment is more complex than the ourdoor scenario, as it includes one large blocker (a whiteboard) and several metal objects (a TV, cabinet door handles, a fence with several bars, and metallic parts of other pieces of furniture). It is hard to accurately determine the number of metal objects in this setup and, more importantly, how many of them act as reflectors. To give an idea of the resulting LOS and Non-LOS (NLOS) paths in the indoor office scenario, we plot in Figs. 9(a) and Fig. 9(b) the RSS values corresponding to various Tx beam directions at two different Rx locations (the Rx locations for the indoor scenario are shown as white dots in Fig. 8(b)). In Fig. 9(a), the Rx is located at the rightmost corner of the room. We can easily distinguish the LOS path from the NLOS paths based on the RSS values. Moreover, we observe multiple "peaks" in the figure, indicating reflections from nearby objects. In Fig. 9(b), the Rx is located at the left side of the office, blocked by the whiteboard. The RSS values, in this case, are much weaker than those in Fig. 9(a), and there is no distinguishable maximum. This is expected for a NLOS scenario.

In Fig. 10(a)–10(c), we depict the total delivered traffic vs. time under ATS, dynamic oracle, and static oracle. Throughout the experiments, the slot duration is set to 1 ms. Fig. 10(a) and 10(b) depict the performance for the outdoor scenario under high mobility (UE moves at a fixed speed of $\beta = 14$ cm/slot $= 504$ km/hr) and moderate mobility ($\beta = 3.5$ cm/slot $= 126$ km/hr), respectively. Similarly, Fig. 10(c) depicts the performance for the indoor scenario under low mobility ($\beta = 0.5$ cm/slot $= 18$ km/hr).

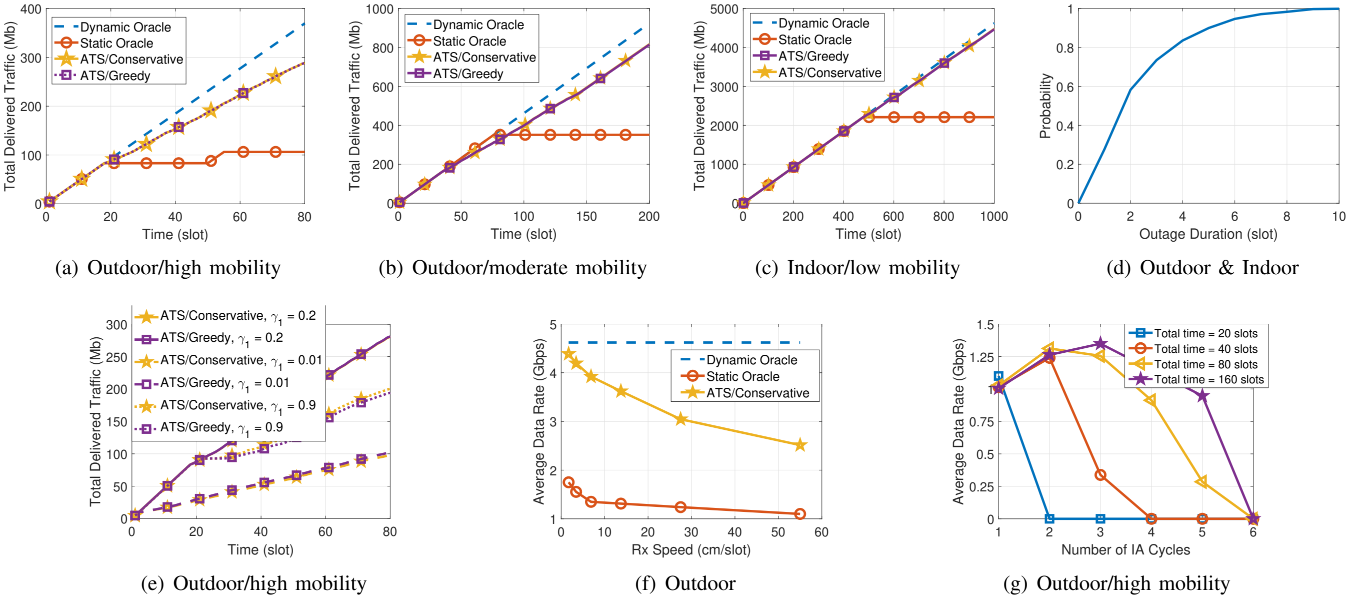
All three algorithms perform similarly in the beginning. This is because when the change in the Rx location is small, the Tx can keep using the best beam that was identified during IA. Also note that the ATS/greedy and ATS/conservative exhibit the same performance for the selected design parameters ($P = 100$, $\gamma_1 = 0.2$, and $\gamma_2 = 20$). As seen from Fig. 10(a), the total delivered traffic using ATS is 182% higher than that of the static oracle in the outdoor scenario, and is only 21% lower than that of the dynamic oracle. For the indoor scenario, Fig. 10(c) shows that the total delivered traffic using ATS is 102.14% higher than that of the static oracle and 3.28% lower than that of the dynamic oracle. In both scenarios, ATS provides significant gains in terms of total delivered traffic compared to a static oracle, and performs reasonably close to the dynamic oracle (i.e., low regret). Similar trends are observed under moderate mobility (see Fig. 10(b)).

Fig. 10(d) depicts the CDF for the outage duration, considering both indoor and outdoor scenarios. The outage duration refers to the time from the onset of an outage until communications are restored. This metric is especially important for real-time traffic (e.g., voice/video), which can tolerate occasional, non-persistent packet losses. It can be observed that with probability greater than 0.9, outage durations will not exceed 5 slots. Outage can be reduced further by limiting MAMBA's operation to low-order modulation schemes, as shown later.

Fig. 10(e) depicts the total delivered traffic versus time under ATS for different γ_1 values. The worst performance is seen when $\gamma_1 = 0.01$, i.e., when the information obtained during IA is almost instantly forgotten. In this case, ATS cannot exploit the useful prior information, and hence, the dashed curves do not follow others even during the first 20 slots. On the other hand, when $\gamma_1 = 0.9$, ATS cannot adapt to the changing environment fast enough. Specifically, it keeps using the previous beam even after its quality has degraded. When $\gamma_1 = 0.2$, ATS can balance exploration and exploitation, and can achieve the highest total delivered traffic.

The effect of Rx speed on the average data rate is studied in Fig. 10(f). When the Rx is slow, ATS performs quite close to the dynamic oracle and achieves only 4% lower average rate. In addition, the regret, i.e., the gap between the dynamic oracle and ATS, scales logarithmically with the Rx speed. As the Rx moves faster, the performance of ATS drops because it cannot learn the environment fast enough and adapt its behavior. Note that in practice, the Rx speed is unlikely to exceed 10 cm/slot, which translates into 360 km/hour. Therefore, the operating point of a BS will always be on the left side of the figure. Higher speeds are illustrated here to show the trend in the average data rate.

In Fig. 10(g), we depict the average data rate using the static oracle for different numbers of IA cycles. Running IA more frequently makes the static oracle more reactive, but also adds a significant search overhead. When the total measurement duration (IA+data) is short, the overhead of rerunning IA becomes more pronounced. For this reason, when the total duration is 20 slots, the static oracle can only run



(a) Outdoor/high mobility

(b) Outdoor/moderate mobility

(c) Indoor/low mobility

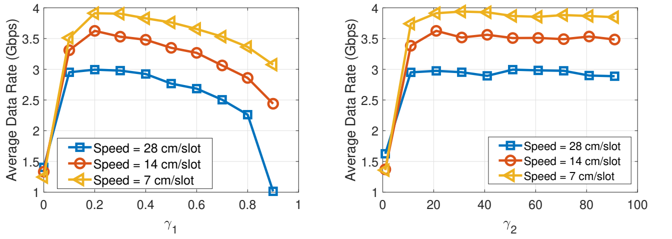
(d) Outdoor & Indoor

(e) Outdoor/high mobility

(f) Outdoor

(g) Outdoor/high mobility

Fig. 10. Experimental evaluation results. Total delivered traffic vs. time for outdoor scenario at speeds up to (a) 14 cm/slot and (b) 3.5 cm/slot, and in indoor at speeds up to (c) 0.5 cm/slot. (d) CDF of the outage duration for ATS considering both indoor and outdoor scenarios. For the outdoor experiments, (e) depicts the total delivered traffic vs. time for ATS with different γ_1 values, (f) average data rate vs. Rx speed, and (g) depicts the average data rate vs. number of IA cycles for the static oracle.



(a)

(b)

Fig. 11. Average data rate versus: (a) γ_1 ($\gamma_2 = 20$), and (b) γ_2 ($\gamma_1 = 0.2$). Outdoor scenario.

one IA cycle; otherwise, its average rate drops to 0. For longer measurement durations, the optimum number of IA cycles increases progressively, as the overhead of rerunning IA becomes less noticeable.

Finally, we study the impact of γ_1 and γ_2 on the performance of ATS. Fig. 11(a) shows that the selection of a very small γ_1 can significantly degrade the performance of ATS. Specifically, when γ_1 is too small, it cannot learn the environment, as the past information is almost instantly forgotten. On the other hand, when γ_1 is too large, the algorithm loses its reactiveness. Fig. 11(a) also shows that when the Rx moves more slowly, a larger γ_1 does not reduce the average data rate as much. That is, when the environment changes more slowly, ATS does not need to forget old information very quickly. The effect of γ_2 is different, as seen in Fig. 11(b). Except when $\gamma_2 < 10$, the selection of γ_2 does not change the performance of ATS significantly. Recall that the effect of γ_1 is multiplicative, whereas the effect of γ_2 is additive. Therefore, γ_1 affects the average data rate more substantially.

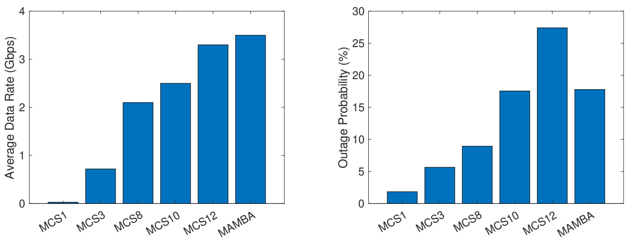
B. Simulation Results

Due to the limitations of our experimental setup, we also consider computer simulations to study the effects of different Tx-Rx distances, UE speeds, and mobility patterns. We set the Tx power to $P_{\text{Tx}} = 30$ dBm, $A_{\text{BS}} = 16$ with half-wavelength element spacing, and $A_{\text{UE}} = 2$ with a spacing of 0.1 wavelengths between the elements. The Tx and Rx use uniform linear arrays (ULA) and are placed on parallel lines facing each other. The BS performs azimuthal beam tracking in the range $\pm 30°$ around the broadside, with a scanning resolution of $5°$. This results in 13 possible Tx beams. The Rx beam is fixed, with beam width of $\sim 60°$. No tracking is performed by the UE. In the simulations, the location of the BS is always fixed, whereas the UE moves according to a random waypoint model, i.e., it first selects a random destination in the simulation area and a random speed between 0.1 km/h and a maximum speed of $\beta = 4$ km/h for pedestrian speeds and $\beta = 80$ km/h for vehicular speeds. The UE then moves to this destination and pauses for a random time before selecting another random location and speed. Tx-Rx distance varies between 10m and 250m. Results are averaged over 1000 runs. The simulations are obtained for a center frequency of 28 GHz.

Channel Model: We simulate a channel with path loss, shadowing, and small-scale effects. We also consider both LOS and NLOS paths. To synthesize the channel, we first determine whether or not a LOS path exists based on the probabilities in Equations (8a), (8b), and (8c) of [11] (these probabilities were obtained following extensive measurements). Next, we determine the large-scale effects, including path loss, using the channel parameters in Table I of [11]. To create multi-path components (MPCs), we randomly place three *point scatterers* on an ellipsoid between the Tx and Rx, thereby introducing small-scale effects as well as three NLOS clusters (along with the LOS cluster). At the Rx, each of the four clusters contains 32 rays (16 transmitted rays times two Rx antenna elements). Each ray that arrives at a scattering point has its own AoA and AoD, as dictated by the location of the Tx element, scattering point, and Rx element.

Next, we compute the coefficients of the MIMO channel, given the position of the Tx, Rx, and the three scatterers. The small-scale channel gain for each path (ray) between a Tx antenna element and an Rx antenna element is sampled from a complex normal distribution of zero mean and unit variance. We then calculate the Tx and Rx beamforming vectors based on their respective azimuth angles (only azimuthal tracking is performed). The Tx azimuth angle is determined by a given beam tracking algorithm. The beamforming vectors and MIMO channel matrix are used to calculate the beamforming gain. Along with the large-scale path loss, this gain can be used to determine the received signal strength.

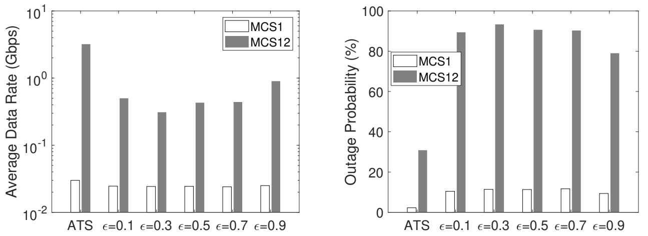
First, we evaluate the performance gain of MAMBA due to beam tracking, separately from the combined gain of beam tracking and rate adaptation. To do that, we run MAMBA at a fixed MCS, selected from the following set: MCS1 (BPSK with 1/2 code rate), MCS3 (BPSK with 5/8 code rate), MCS8 (QPSK with 3/4 code rate), MCS10 (16-QAM with 1/2 code rate), and MCS12 (16-QAM with 3/4 code rate). The results are shown in Fig. 12 (where MAMBA refers to the adaptive-rate version). Fig. 12(a) shows that under a fixed-rate setup, MCS1 and MCS12 achieve the lowest and the highest throughput, respectively. This is expected, as selecting a lower MCS index will result in lower data rate. We also observe the opposite trend for the outage probability in Fig. 12(b), where MCS1 results in the lowest outage probability and MCS12 results in the highest outage probability. Under MCS1, outages occur because the MAMBA algorithm occasionally explores inadequate beam directions. As we increase the MCS index, the outage probability increases because there is a higher likelihood that the link between the BS and UE cannot support the selected MCS. This motivates the use of a rate adaptation scheme. As can be seen in Fig. 12, adaptive-rate MAMBA achieves a higher data rate than any fixed-rate scheme. For example, compared with MCS12, adaptive-rate MAMBA achieves a slightly higher throughput but reduces the outage probability by 35%. Comparing it with MCS10 (which has almost the same outage probability), adaptive-rate MAMBA provides 37% throughput gain.



(a) (b)

Fig. 12. Data rate and outage performance for fixed-rate and adaptive-rate variants of MAMBA.

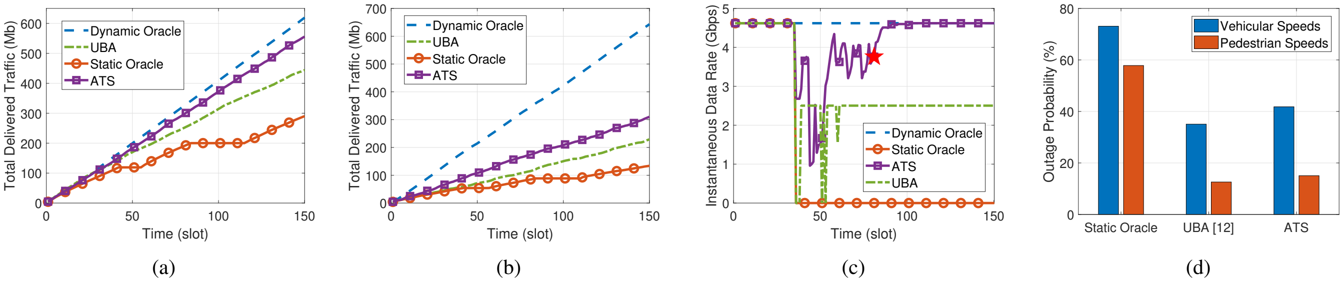
Next, we compare ATS used within a fixed-rate variant of MAMBA against the ϵ-greedy algorithm [1] (which targets beam tracking but does not perform adaptive rate selection). Without loss of generality, we consider pedestrian speeds. Fig. 13 depicts the throughput and outage probability for both algorithms with $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. A smaller ϵ pushes the ϵ-greedy algorithm to exploit the current beam more often, and vice versa. At the lowest MCS (MCS1), the data rate achieved by ATS is 21% higher than the data rate achieved by the best-possible realization of the ϵ-greedy algorithm (at $\epsilon = 0.9$). Such a throughput advantage is combined with 78% reduction in the outage probability. At the highest available MCS (MCS12), ATS achieves about 255% improvement in throughput over the best possible ϵ-greedy algorithm, along with 61% reduction in the outage probability.



(a) (b)

Fig. 13. Comparison between ATS and the ϵ-greedy algorithm under a fixed MCS: (a) average data rate; (b) outage probability.

In the next set of simulations, we compare ATS against dynamic oracle, static oracle, and the UBA algorithm. Fig. 14(a) depicts the total delivered traffic for the four techniques as a function of time, considering pedestrian speeds. The figure shows that ATS outperforms both UBA and the static oracle, and is only 10% below the dynamic oracle. At vehicular speeds, tracking UEs becomes much harder, and so the performance gap between the dynamic oracle and the other algorithms increases significantly. Nevertheless, ATS still performs better than UBA and the static oracle, as seen in Fig. 14(b). The instantaneous data rate is shown in Fig. 14(c) as a function of time. Here, we let the UE roam freely for 80 slots and then bring it to a stop (the slot after which the UE stops is marked with the red star). We observe that the performance of ATS converges to that of the dynamic oracle within only 10 slots, whereas UBA converges to a suboptimal data rate. Therefore, we conclude that for stationary scenarios, ATS quickly converges to the optimal beam and MCS index. Finally, in Fig. 14(d), we study the outage probabilities of different algorithms. The static oracle is significantly outperformed by UBA and ATS, with UBA performing slightly better than ATS at both vehicular and pedestrian speeds. However, as observed from Fig. 14(b), even though ATS is slightly more prone to outages, its total delivered traffic is still higher than UBA, implying that UBA often settles on suboptimal data rates, while ATS is more likely to perform optimally.



(a) (b) (c) (d)

Fig. 14. Performance comparison based on simulation results. (a) Total delivered traffic for pedestrian speeds up to 4 km/h, (b) total delivered traffic for vehicular speeds up to 80 km/h, (c) instantaneous data rates (slot after which the UE stops is marked with the red star), (d) outage probabilities for the pedestrian and vehicular speeds.

VII. Conclusions

In this paper, we proposed MAMBA, an ATS-based multi-armed bandit scheme for joint beam tracking and MCS selection in mmW systems. MAMBA relies on prior information collected during IA and updates such information based on the RSS of ACK/NACK messages received from the UE. We derived an upper bound on the regret of the ATS algorithm used in MAMBA, and used OTA experiments and computer simulations to study its performance and contrast it with several other beam tracking techniques. Our results indicate that MAMBA improves the link throughput by up to 182% compared to a static oracle scheme (the default approach for 5G NR systems) and performs reasonably close to an optimal but practically infeasible dynamic oracle scheme. MAMBA was also compared with two other state-of-the-art beam tracking schemes (ϵ-greedy and UBA algorithms). Compared with UBA, MAMBA has a 25-35% throughput advantage. Considering only its beam tracking part (i.e., fixing the MCS), MAMBA achieves a 21% throughput gain over the ϵ-greedy algorithm at the lowest MCS index, and 255% gain at the highest MCS index. Finally, we verified the effectiveness of MAMBA's rate adaptation by comparing it with a fixed-rate counterpart. For roughly the same outage probability, rate adaptation results in about 37% throughput gain over a fixed-rate variant.

Future Directions: ATS relies on frequent HARQ feedback at the BS for timely update of the reward distributions. However, when the downlink traffic is light, the feedback will be sparse and MAMBA will not learn fast enough changes in beam quality. The same is true when block ACKs are used, which result in fewer HARQ transmissions and longer times between reward updates. The slow learning rate translates into slower convergence and degradation in the beam tracking performance. Likewise, as was shown in Fig. 14, the performance of ATS degrades at high speeds due to rapid fluctuations in the mmW channel. More specifically, in ATS the time duration between two successive instances of beam selection is much larger than the "beam coherence time" (as defined in [54]). This makes past observations less relevant and yields poor performance. As a future work, we plan to extend MAMBA to account for varying traffic load, block ACKs, and very high mobility. The latter aspect will involve determining (online) a suitable beam selection instance based on the beam coherence time. We will also exploit the inherent correlations between adjacent beams to enable simultaneous updates of the reward distributions of multiple beams (following the arrival of an ACK/NACK message). This facilitates fast recovery in scenarios where the best beam changes frequently due to UE mobility (as opposed to blockage). We will also investigate the effectiveness of ATS in a multi-UE setting, where the BS tracks multiple users that operate over exclusive frequency/time resource blocks (hence, they do not interfere with each other) or share common blocks.

References

[1] E. M. Mohamed, "Millimeter wave beamforming training: A reinforcement learning approach," *International Journal of Electronics and Telecommunications*, vol. 67, no. No 1, pp. 95–102, 2021.

[2] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. of the IEEE INFOCOM 2018 Conference*, (Honolulu, HI), pp. 2393–2401, Apr. 2018.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[4] IEEE Computer Society, "IEEE Standard-part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 3: Enhancements for very high throughput in the 60 GHz band (adoption of IEEE std 802.11ad-2012)," 2014.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

[6] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[7] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, "Directional cell discovery in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6664–6678, 2015.

[8] I. Aykin and M. Krunz, "FastLink: an efficient initial access protocol for millimeter wave systems," in *Proc. of the 21st ACM MSWiM Conference*, (Montreal, CA), pp. 109–117, Oct. 2018.

[9] I. Aykin, B. Akgun, and M. Krunz, "Multi-beam transmissions for blockage resilience and reliability in millimeter-wave systems," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 37, Dec. 2019.

[10] B. Akgun, M. Krunz, and D. Manzi, "Impact of beamforming on delay spread in wideband millimeter-wave systems," in *Proc. of the IEEE ICNC 2020 Conference*, (Big Island, Hawaii), Feb. 2020.

[11] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[12] T. Bai, A. Alkhateeb, and R. W. Heath, "Coverage and capacity of millimeter-wave cellular networks," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 70–77, 2014.

[13] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial access in millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 7926–7940, Dec. 2016.

[14] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proc. of the 2018 Conference of the ACM Special Interest Group on Data Communication*, (Budapest, Hungary), pp. 432–445, Aug. 2018.

[15] I. Aykin, B. Akgun, and M. Krunz, "Smartlink: Exploiting channel clustering effects for reliable millimeter wave communications," in *Proc. of the IEEE INFOCOM 2019 Conference*, (Paris, France), pp. 1117–1125, Apr. 2019.

[16] I. Aykin and M. Krunz, "Efficient beam sweeping algorithms and initial access protocols for millimeter-wave networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2504–2514, 2020.

[17] A. Zhou, L. Wu, S. Xu, H. Ma, T. Wei, and X. Zhang, "Following the shadow: Agile 3-D beam-steering for 60 GHz wireless networks," in *Proc. of the IEEE INFOCOM 2018 Conference*, (Honolulu, HI), Apr. 2018.

[18] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4391–4403, 2013.

[19] C. Liu, M. Li, S. V. Hanly, P. Whiting, and I. B. Collings, "Millimeter-wave small cells: Base station discovery, beam alignment, and system design challenges," *IEEE Wireless Communications*, vol. 25, no. 4, pp. 40–46, 2018.

[20] I. Aykin, B. Akgun, M. Feng, and M. Krunz, "MAMBA: a multi-armed bandit framework for beam tracking in millimeter-wave systems," in *Proc. of the IEEE INFOCOM 2020 Conference*, (Toronto, Canada), July 2020.

[21] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 1861–1870, 10–15 July 2018.

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[23] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 1582–1591, July 2018.

[24] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 1928–1937, June 2016.

[25] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 1889–1897, July 2015.

[26] Y. Wang, Z. Wei, and Z. Feng, "Beam training and tracking in mmwave communication: A survey," 2022.

[27] V. Va, H. Vikalo, and R. W. Heath, "Beam tracking for mobile millimeter wave communication systems," in *Proc. of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, (Washington, DC), pp. 743–747, Dec. 2016.

[28] C. Zhang, D. Guo, and P. Fan, "Tracking angles of departure and arrival in a mobile millimeter wave channel," in *Proc. of the IEEE International Conference on Communications (ICC)*, (Kuala Lumpur, Malaysia), pp. 1–6, May 2016.

[29] S. Sur, X. Zhang, P. Ramanathan, and R. Chandra, "BeamSpy: Enabling robust 60 GHz links under blockage," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, (Santa Clara, CA), pp. 193–206, USENIX Association, Mar. 2016.

[30] A. Zhou, X. Zhang, and H. Ma, "Beam-forecast: Facilitating mobile 60 GHz networks via model-driven beam steering," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, 2017.

[31] A. Patra, L. Simić, and M. Petrova, "Experimental evaluation of a novel fast beamsteering algorithm for link re-establishment in mm-wave indoor WLANs," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–7, 2016.

[32] J. Palacios, D. De Donno, and J. Widmer, "Tracking mm-Wave channel dynamics: Fast beam training strategies under mobility," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, 2017.

[33] A. Zhou, T. Wei, X. Zhang, and H. Ma, "FastND: Accelerating directional neighbor discovery for 60-GHz millimeter-wave wireless networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 5, pp. 2282–2295, 2018.

[34] D. Burghal, N. A. Abbasi, and A. F. Molisch, "A machine learning solution for beam tracking in mmWave systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 173–177, 2019.

[35] Y. Guo, Z. Wang, M. Li, and Q. Liu, "Machine learning based mmWave channel tracking in vehicular scenario," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2019.

[36] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30507–30526, 2019.

[37] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. of the IEEE INFOCOM 2018 Conference*, (Honolulu, HI), pp. 1961–1969, Apr. 2018.

[38] Y. Koda, M. Shinzaki, K. Yamamoto, T. Nishio, M. Morikura, Y. Shirato, D. Uchida, and N. Kita, "Millimeter wave communications on overhead messenger wire: Deep reinforcement learning-based predictive beam tracking," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, pp. 1216–1232, 2021.

[39] Y. Liu, Z. Jiang, S. Zhang, and S. Xu, "Deep reinforcement learning-based beam tracking for low-latency services in vehicular networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2020.

[40] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[41] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[42] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained thompson sampling," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 739–747, 2019.

[43] 3GPP TR 38.802 v14.2.0, "Study on new radio access technology-physical layer aspects (release 14)," Sept. 2017.

[44] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[45] 3GPP TS 38.213 v15.5.0, "Physical layer procedures for control (release 15)," May 2019.

[46] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS long term evolution: From theory to practice*. John Wiley & Sons, 2011.

[47] K. Benkic, M. Malajner, P. Planinsic, and Z. Cucej, "Using RSSI value for distance estimation in wireless sensor networks based on ZigBee," in *Proc. of the 15th International Conference on Systems, Signals and Image Processing*, (Bratislava, Slovakia), pp. 303–306, June 2008.

[48] O. G. Adewumi, K. Djouani, and A. M. Kurien, "RSSI based indoor and outdoor distance estimation for localization in WSN," in *Proc. of the IEEE International Conference on Industrial Technology (ICIT)*, (Cape Town, South Africa), pp. 1534–1539, Feb. 2013.

[49] Z. Xu, R. Wang, X. Yue, T. Liu, C. Chen, and S.-H. Fang, "FaceME: Face-to-machine proximity estimation based on RSSI difference for mobile industrial human-machine interaction," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3547–3558, 2018.

[50] A. Slivkins and E. Upfal, "Adapting to a changing environment: the Brownian restless bandits," in *Proc. of the 21st Conference on Learning Theory (COLT)*, pp. 343–354, July 2008.

[51] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, pp. 1221–1243, Nov. 2014.

[52] 3GPP TS 38.214 version 15.7.0 Release 15, "NR: Physical layer procedures for data," Jan. 2018. (available online).

[53] WICON, 2022. https://wireless.ece.arizona.edu/software-and-datasets.

[54] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, 2017.

Acknowledgments

The authors would like to thank Keysight Technologies for providing the experimental setup used in this paper. This research was supported in part by NSF and by the Broadband

Wireless Access & Applications Center (BWAC). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of NSF.

Marwan Krunz (S'93-M'95-SM'04-F'10) is a is a Regents Professor at the University of Arizona. He holds the Kenneth VonBehren Endowed Professorship in ECE and is also a professor of computer science. He directs the Broadband Wireless Access and Applications Center (BWAC), a multi-university NSF/industry center that focuses on next-generation wireless technologies. He also holds a courtesy appointment as a professor at the University of Technology Sydney. Previously, he served as the site director for the Connection One center. Dr. Krunz's research is on resource management, network protocols, and security for wireless systems. He has published more than 300 journal articles and peer-reviewed conference papers, and is a named inventor on 12 patents. His latest h-index is 60. He is an IEEE Fellow, an Arizona Engineering Faculty Fellow, and an IEEE Communications Society Distinguished Lecturer (2013-2015). He received the NSF CAREER award. He served as the Editor-in-Chief for the IEEE Transactions on Mobile Computing. He also served as editor for numerous IEEE journals. He was the TPC chair for INFOCOM'04, SECON'05, WoWMoM'06, and Hot Interconnects 9. He was the general vice-chair for WiOpt 2016 and general co-chair for WiSec'12. Dr. Krunz served as chief scientist for two startup companies that focus on 5G and beyond systems and machine learning for wireless communications.

Irmak Aykin received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Turkey, in 2012; the M.S. degree in EE from Bilkent University, Turkey, in 2014; and the Ph.D. degree in ECE from the University of Arizona in 2020. She is currently a Senior Engineer at Qualcomm. Her research interests include millimeter-wave systems, wireless networking, algorithm design and machine learning. She is a recipient of the ACM MSWiM 2018 Best Paper Award.

Sopan Sarkar received the B.S. degree in applied physics, electronics, and communication engineering from University of Chittagong, Bangladesh, in 2016; and the M.S. degree in computer engineering from University of Alabama in Huntsville, Huntsville, AL, in 2020. His research interests include robust millimeter-wave and sub-terahertz system design, with emphasis on artificial intelligence for wireless communication and networking.

Berk Akgun received the B.S. and M.S. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2012 and in 2014, respectively, and the Ph.D. degree in ECE from the University of Arizona (Tucson, AZ) in 2020. He is currently a Senior Engineer at Qualcomm. From 2012 to 2014, he was a software design engineer at the Communication and Information Technologies Division of Aselsan, Ankara, Turkey. His research interests lie in the areas of mmWave channel characterization, robust mmWave system design, wireless communications and networking, with emphasis on designing secure multiuser MIMO systems.