MAMBA: A Multi-armed Bandit Framework for Beam Tracking in Millimeter-wave Systems

Irmak Aykin, Berk Akgun, Mingjie Feng, and Marwan Krunz
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85721
Email: {aykin, berkakgun, mingjiefeng, krunz}@email.arizona.edu

***Abstract*—Millimeter-wave (mmW) spectrum is a major candidate to support the high data rates of 5G systems. However, due to directionality of mmW communication systems, misalignments between the transmit and receive beams occur frequently, making link maintenance particularly challenging and motivating the need for fast and efficient beam tracking. In this paper, we propose a multi-armed bandit framework, called MAMBA, for beam tracking in mmW systems. We develop a reinforcement learning algorithm, called adaptive Thompson sampling (ATS), that MAMBA embodies for the selection of appropriate beams and transmission rates along these beams. ATS uses prior beam-quality information collected through the initial access and updates it whenever an ACK/NACK feedback is obtained from the user. The beam and the rate to be used during next downlink transmission are then selected based on the updated posterior distributions. Due to its model-free nature, ATS can accurately estimate the best beam/rate pair, without making assumptions regarding the temporal channel and/or user mobility. We conduct extensive experiments over the 28 GHz band using a 4×8 phased-array antenna to validate the efficiency of ATS, and show that it improves the link throughput by up to 182%, compared to the beam management scheme proposed for 5G.**

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

I. Introduction

Millimeter-wave (mmW) communications are one of the frontiers of next-generation wireless systems, including 5G New Radio (NR) [1] and WiGig [2]. The abundant spectrum available in the mmW bands 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 [3]. Traditionally, mmW bands have not been utilized for terrestrial communications due to their harsh propagation characteristics. Nevertheless, even though mmW signals are attenuated more heavily compared to their sub-6 GHz counterparts, their small wavelengths allow large antenna arrays to be implemented into small form-factor radios. By using high-dimensional phased-array antennas, transmissions/receptions can be beamed towards desired directions. The resulting beamforming gain makes it possible to achieve high data rates, despite the unfavorable characteristics of the channel [4].

While beamforming allows for high gains, establishing and maintaining a directional link can be quite challenging [5]–[7]. Due to the limited scattering at mmW frequencies, the channel between the BS and the user equipment (UE) is typically sparse [8], [9]. Specifically, 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 that takes place before a BS-UE link can be established.

Although the IA problem has been well studied (e.g., [10]–[16]), tracking a mobile UE efficiently and reliably is still a research challenge [17]. After the initial directional link between a BS and a UE is established, significant beam misalignments occur frequently due to UE mobility, environmental changes, or even wind [18]. These misalignments incur a large beamforming loss, resulting in reduced data rate or link outage. Consequently, tracking UEs and maintaining the quality of their directional links are quite critical.

In this paper, we propose a multi-armed bandit framework, called MAMBA, for beam tracking in mmW systems. In MAMBA, each beam is modeled as an arm of the multi-armed bandit, and the BS acts as the agent interacting with these arms to learn the underlying system dynamics. To quantify a beam's quality, we consider the modulation and coding schemes (MCSs) that can be supported by a beam. We then develop a reinforcement learning (RL) algorithm, called adaptive Thompson sampling (ATS), to be used in MAMBA for determining the optimal beam/MCS pair for each downlink transmission. Specifically, ATS aims at maximizing the expected transmission rate, taking into account the estimated reward distributions associated with each beam. However, due to the time-varying nature of the environment, keeping track of these reward distributions is nontrivial. To address this issue, ATS uses apriori beam-quality information 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. Due to its model-free nature, ATS can accurately estimate the best beam/MCS pair, without making unrealistic assumptions regarding the temporal channel and/or the UE mobility. Because all RL algorithms require some time to learn the optimal strategy, we derive an upper bound on regret, which is the loss in reward resulting from deviating from the optimal strategy.

The main contributions of this paper are as follows:

- We introduce a multi-armed bandit framework, MAMBA,

to model a joint beam tracking and adaptive rate selection problem in mmW systems. MAMBA does not incur extra control overhead to the system. It utilizes the ACK/NACK feedback obtained from the UE to select the best action, i.e., the best beam/MCS pair.
- We develop an RL algorithm called ATS to be used within MAMBA. ATS selects the optimal beam/MCS pair so as to maximize the data rate of the underlying transmission. In ATS, the prior reward distribution of each beam is set based on its quality measured during IA. These priors are updated whenever an ACK/NACK packet is received, and the beam/MCS pair to be used in the next transmission is determined using the updated posterior distributions of the rewards. 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.
- Because the goal of ATS is to minimize the cumulative regret, 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 at 28 GHz frequency using a 4×8 phased-array antenna, we verify the efficiency of ATS in terms of throughput, average data rate, and outage duration in both indoor and outdoor scenarios. Our experiments show that ATS can improve the throughput by up to 182%, compared to a static beam management scheme that is proposed for 5G.

II. Related Work

Multi-armed bandit (MAB) modeling framework has been extensively applied in the literature to various online optimization problems [19]. Its goal is to capture the exploration versus exploitation tradeoff and to minimize the cumulative regret of deviating from the optimal strategy. RL is the most common technique for solving MAB problems, and there are three widely applied RL algorithms in the literature to achieve minimum regret in *stationary* MAB problems: ϵ-greedy, upper confidence bound (UCB), and Thompson sampling (TS) [20]. ϵ-greedy is a simplistic approach in which the algorithm selects the action with the highest empirical mean with probability $1-\epsilon$, or a random action with probability ϵ. UCB, on the other hand, maintains a confidence interval for each arm, in addition to the empirical means. Then, in each round, the algorithm greedily picks the action with the highest upper confidence bound. Finally, in TS, the rates of exploration and exploitation are dynamically updated with respect to the posterior distribution of each beam. Specifically, the beams with higher estimated rewards are exploited more frequently, but the beams with lower estimated rewards are still occasionally explored. Recently, TS was empirically shown to outperform the two other approaches for a wide variety of problems [21], [22]. Accordingly, in this paper, we adopt the TS approach, but adapt it to nonstationary scenarios. This adaptation is necessary due to UE mobility and/or environmental changes.

RL and Kalman-filter-based estimation techniques have also been used for solving various problems in wireless communications. In [23], the authors proposed a variation of TS for optimal rate selection over time-varying wireless channels with unknown channel statistics, without considering directional communications. The authors in [24] and [25] used RL for beam tracking. However, their methods utilize location information, which may not always be available at the BS. The authors in [26] used extended Kalman filters (EKF) for angle-of-arrival (AoA) and angle-of-departure (AoD) tracking. However, since their method tracks the currently utilized channel cluster, it can only track one AoA/AoD pair at a time. Similarly, the authors in [27] used Kalman filters to track the AoA and AoDs at the transmitter (Tx) and the receiver (Rx). However, both [26] and [27] assume that the angles are randomly perturbed according to a zero-mean Gaussian distribution, which may not hold in a real wireless system.

Our proposed ATS algorithm is model-free, and hence, does not make assumptions regarding the underlying channel and/or user mobility. The online decision-making process in each round is done by solving a system of linear equations, which is easy to parallelize. In addition, to our knowledge, our paper is the first to study joint beam/rate selection for mmW systems.

III. System Model

Without loss of generality, we consider the beam tracking problem for a single UE. We first briefly describe how beamforming is typically applied on a mmW channel, then explain the MAMBA framework, and finally formulate the reward-maximization problem.

A. Codebook-based Beamforming

Consider a link between a BS and a UE that communicate using uniform planar arrays (UPAs). Let the total number of antennas at the BS and the UE be A_{BS} and A_{UE}, respectively. Also, let $\mathbf{H}$ denote 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 [4]. Denote the 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

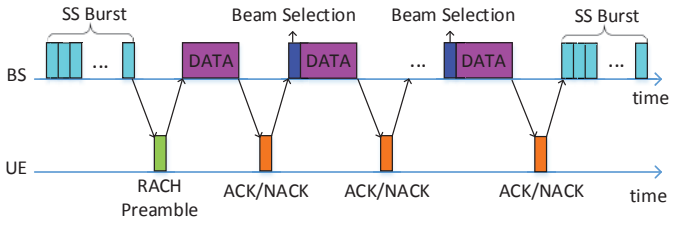


Fig. 1. Timeline of the proposed downlink communication scheme between a BS and a UE.

t, where $P_{ij}(t) = |y_{ij}|^2$. Because $\mathbf{H}$ is time-varying, the distribution of $P_{ij}(t)$ is nonstationary. Our goal here is to find the best strategy to achieve the highest long-term throughput.

B. MAMBA Framework

A naive tracking strategy is to exploit the current best beam pair, say $(\mathbf{f}_i, \mathbf{q}_j)$, for a relatively long time. In the current 5G standard [28], when a new UE joins the network, it waits for the BS to execute the IA procedure, through which they can discover initial communication directions. During IA, the BS transmits synchronization signals (SS), allowing a listening UE to measure beam qualities and report them back. BS periodically reruns the IA to discover new UEs and update the best beams for already discovered ones. In between IA cycles, other periodic control signals, called channel state information-reference signals (CSI-RS), are transmitted by the BS to maintain communication. CSI-RS is 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 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 [29]. 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. Note that ACK/NACK mechanism is already a part of 5G physical layer [30]. We assume that the UE communicates using relatively wide beams so the tracking problem is only applicable to the BS side. This is a reasonable assumption, considering the smaller form-factor and fewer antenna elements on a UE. Given our focus on the BS side only, in the subsequent sections, the subscript 'BS' will be dropped from related variables.

A reasonable choice for modeling the beam tracking problem is to use Markov decision processes (MDPs). MDPs have been extensively used in the literature to model a wide variety of problems in which an agent continually interacts with the environment to achieve a goal [31]. The agent selects actions and the environment responds to these actions, presenting new states to the agent. While very powerful, MDPs can be computationally expensive to solve. Thankfully, in our beam tracking problem, selecting a new action (beam) does not change the state of the BS, i.e., selecting a beam at time t does not limit (or increase) the beam selection possibilities at time $t+1$. Therefore, our problem can be modeled as a single-state MDP, i.e., an MAB problem.

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 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 $\mathbf{r}_t \in \mathcal{R}$ is observed. This $\mathbf{r}_t$ is a random sample drawn from the selected beam's underlying reward distribution. Let $\boldsymbol{\Theta}_{i,t}$ denote the reward distribution associated with beam i at time t, and let $\boldsymbol{\theta}_{i,t}$ denote the mean of $\boldsymbol{\Theta}_{i,t}$, i.e., $\mathbb{E}[\boldsymbol{\Theta}_{i,t}] = \boldsymbol{\theta}_{i,t}$ where $\boldsymbol{\theta}_{i,t}$ is unknown. Note that there are D distributions in total associated with various BS beams. Letting $a_t = \mathbf{f}_i$ means that beamformer $\mathbf{f}_i$ is selected at time t, and hence the BS receives a reward $\mathbf{r}_t \sim \boldsymbol{\Theta}_{i,t}$. In MAMBA, the BS obtains the reward by measuring the received signal strength (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 do beam/MCS selection for the subsequent downlink data transmission. Fig. 1 shows the proposed downlink communication timeline.

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.

C. Problem Formulation

In MAMBA, the BS has some prior "belief" about the reward distribution of each beam, thanks to 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 $\boldsymbol{\theta}$ after taking into account the observed data $\mathbf{x}$, can be computed as:

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

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.

Here we model the rewards as M-dimensional variables. For each transmission, the BS chooses a beam and 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,

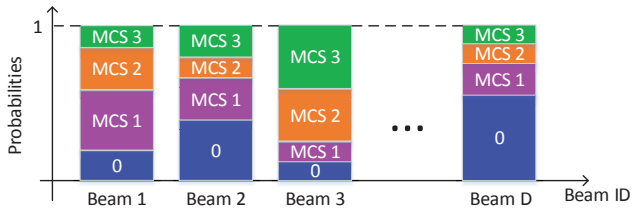


Fig. 2. Illustration of qualities of different BS beams in terms of achievable MCS indices ($M = 4$). Probabilities refer to $\theta_{i,t}^{(m)}$, $\forall i \in \mathcal{A}$ and $\forall m \in \{1, \cdots, M-1\}$ for a given t.

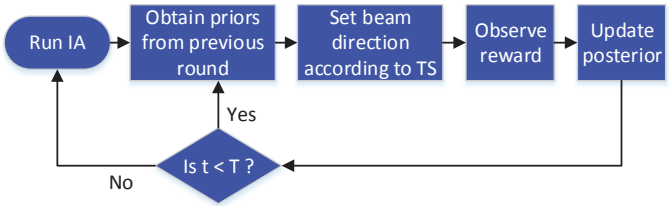


Fig. 3. Flowchart of the proposed beam tracking method at the BS.

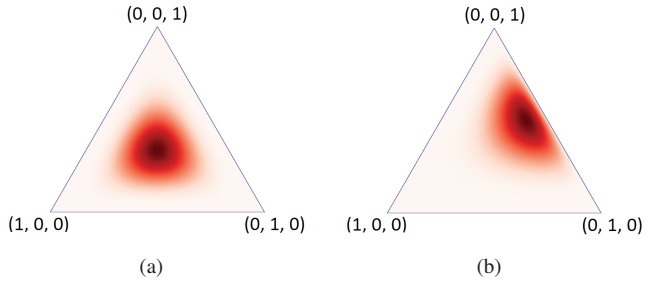


Fig. 4. Visualization of 3D Dirichlet distributions as a heatmap, where darker areas denote higher probabilities and lighter areas denote lower probabilities. (a) $\boldsymbol{\alpha}_{i,t} = [5, 5, 5]$, (b) $\boldsymbol{\alpha}_{i,t} = [2, 5, 6]$.

each of which has an associated rate v_m, $m \in \{1, \cdots, M-1\}$, or 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 on that beam. If neither an ACK nor a NACK is received, the reward is set to 0.

The 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. 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 of beam qualities is shown in Fig. 2 using $M = 4$.

At any time t, the observed reward vector $\mathbf{r}_t = [r_t^{(0)}, \cdots, r_t^{(M-1)}]$ contains a single 1 at the highest attainable MCS index (based on the RSS of ACK/NACK packets) and 0s elsewhere. For convenience, we assign $r_t^{(0)} = 1$ for an unsuccessful communication and $r_t^{(m)} = 1$ 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 $\mathbf{r}_t \mathbf{v}^T$, 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.

With the above preliminaries, 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:

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

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 Algorithm

In this section, we explain our TS-based algorithm used by a BS in MAMBA for beam tracking. 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 the proposed method is shown in Fig. 3. We first consider a stationary system where the expected rewards of the arms do not change in time, and 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 categorical distribution and the Dirichlet distribution is the conjugate prior of it, 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).

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 arbitrary 3D Dirichlet distributions are shown in Fig. 4.

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

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 $\mathbf{r}_t$
  Update distributions:
    for $i \in \mathcal{A}$ **do**
      if $a_t = i$ **then**
        $\boldsymbol{\alpha}_{i,t+1} \leftarrow \boldsymbol{\alpha}_{i,t} + \mathbf{r}_t$
      else
        $\boldsymbol{\alpha}_{i,t+1} \leftarrow \boldsymbol{\alpha}_{i,t}$

$\text{Cat}(\boldsymbol{\theta}_{i,t})$ rewards and $\text{Dir}(\boldsymbol{\alpha}_{i,t})$ priors $\forall i \in \mathcal{A}$, the update rule for the posterior is as follows:

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

First condition refers to the case when beam i is selected for transmission at time t and a reward $\mathbf{r}_t$ is observed. The posterior distribution of beam i is then updated accordingly. Second condition refers to the case when beam i is not selected for transmission at time t, and thus, its posterior is not changed.

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 *virtually* 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 an effective mechanism to cope with time-varying channel characteristics.

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 the nonstationarity, 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. In practice, this involves implementing a "forget" factor γ_1 that slowly alters the posterior distributions and a "boost" factor γ_2. For $i \in \mathcal{A}$, the update rule can then be written as:

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

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 $\boldsymbol{\alpha}_{i,t}$ with a constant γ_1 effectively increases the variance (given that $0 < \gamma_1 < 1$), but does not alter the mean of the Dirichlet distribution. To show that, let us first calculate $\boldsymbol{\mu}_{i,t+1} \triangleq \mathbb{E}[\text{Dir}(\boldsymbol{\alpha}_{i,t+1})]$:

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

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

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

given that $0 < \gamma_1 < 1$. 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 the multi-dimensional uniform, i.e., $\text{Dir}(\mathbf{1})$. We incorporate this new update rule into an algorithm called ATS (see Algorithm 2).

Prior Selection: In general, a uniform prior works well with most TS algorithms. For our problem formulation, this would correspond to the multi-dimensional uniform distribution, i.e., $\text{Dir}(\mathbf{1})$. 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, convergence time can be significantly reduced.

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. The MCS selection is particularly important, as the effective data rate of a given transmission would be

Algorithm 2 Adaptive Thompson Sampling

for $t = 1, 2, \cdots, T$ **do**
Take Samples:
  for $i \in \mathcal{A}$ **do**
    Sample $\mathbf{s}_{i,t} \sim \text{Dir}(\boldsymbol{\alpha}_{i,t})$
Choose and apply action:
  $a_t = \text{argmax}_{i \in \mathcal{A}} \, \mathbf{s}_{i,t} \mathbf{v}^T$
  Select a_t and observe $\mathbf{r}_t$
Update distributions:
  for $i \in \mathcal{A}$ **do**
    if $a_t = i$ **then**
      $\boldsymbol{\alpha}_{i,t+1} \leftarrow \gamma_1 \boldsymbol{\alpha}_{i,t} + \gamma_2 \mathbf{r}_t$
    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
      $\boldsymbol{\alpha}_{i,t+1} \leftarrow \mathbf{1}$

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 next propose two techniques for the MCS selection.

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

$$m^* = \underset{m \in \{0, \cdots, M-1\}}{\text{argmax}} \; s_{a_t,t}^{(m)} v^{(m)}. \tag{5}$$

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 the conservative selection scheme, the MCS index that is most likely to be attained and can achieve a non-zero rate on 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 that action a_t has been selected, the rate selection problem can be written as:

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

V. Regret Analysis

In this section, we compute an upper bound on the Bayesian regret for our proposed method. Let $\mathcal{I}$ denote an instance of the MAB problem drawn initially from some known prior $\mathbb{P}$ 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 until time T is defined as:

$$\text{BR}(T) = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{I} \sim \mathbb{P}} \left[\mathbb{E} \left[\boldsymbol{\theta}_{a_t^*,t} \mathbf{v}^T - \boldsymbol{\theta}_{a_t,t} \mathbf{v}^T \mid \mathcal{I} \right] \right] \tag{7}$$

where $\boldsymbol{\theta}_{a_t,t}$ denotes the expected reward vector of the action selected by our algorithm at time t and $a_t^* = \text{argmax}_{i \in \mathcal{A}} \, \boldsymbol{\theta}_{i,t} \mathbf{v}^T$. The inner expectation in (7) 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., $\text{BR}_t = \mathbb{E}[\boldsymbol{\theta}_{a_t^*,t} \mathbf{v}^T - \boldsymbol{\theta}_{a_t,t} \mathbf{v}^T]$, where inner and outer expectations in (7) are merged into a single expectation. Then, BR_t can be also written as:

$$\text{BR}_t = \sum_{m=0}^{M-1} \mathbb{E} \left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)} \right] v_m. \tag{8}$$

Now, we focus on bounding $\mathbb{E}[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}]$, $\forall m \in \mathcal{M}$.

We use a random walk process to model the nonstationarity of the rewards obtained from various beams [32]. 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 [32]. The direction of the walk is also determined by a uniform distribution within all the possible directions in $(M-1)$ dimensions. See Fig. 4 for an illustration of this model. 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. 4, 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 from the beam i and MCS index m 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 $t \leq T$,

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

Proof: To simplify the proof, we drop 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 [32]. 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,

$$\Pr\left(\left|\hat{\theta}_{i,t}^{(m)} - \theta_{i,t}^{(m)}\right| \geq \delta_{i,t}\right) \leq \mathcal{O}(T^{-4}) \tag{9}$$

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 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 Hoeffding inequality, we obtain (10). In the next step, each term of the inequality inside the probability expression is divided by γ_2. (12) follows from Lemma 1. Finally, by dividing the terms of the inequality inside the probability expression by $n_{i,t}$, we obtain (9). ∎

In the rest of the analysis, we exploit similar techniques to bound the Bayesian regret as in [33]. 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:

$$U_t^{(m)}(i) = \hat{\theta}_{i,t}^{(m)} + \delta_{i,t} \quad \text{and} \quad L_t^{(m)}(i) = \hat{\theta}_{i,t}^{(m)} - \delta_{i,t}. \tag{13}$$

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

$$\mathrm{BR}_t \leq 2M v_{M-1} \mathbb{E}\left[\delta_{a_t,t}\right] + \mathcal{O}(T^{-4}). \tag{14}$$

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 [33] for further details). Hence,

$$\begin{aligned}
&\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)}\right] = \mathbb{E}_{H_t}\left[\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)} | H_t\right]\right] \\
&= \mathbb{E}_{H_t}\left[\mathbb{E}\left[U_t^{(m)}(a_t) - U_t^{(m)}(a_t^*) + \theta_{a_t^*,t}^{(m)} - \theta_{a_t,t}^{(m)} | H_t\right]\right] \\
&= \mathbb{E}_{H_t}\left[\mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)} | H_t\right] + \mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*) | H_t\right]\right] \\
&= \mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] + \mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right].
\end{aligned} \tag{15}$$

We separately investigate the two terms in (15). Let $(a)^+ \triangleq \max\{0, a\}$ for any real number a. First, consider the second term in (15):

$$\mathbb{E}\left[\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right] \leq \mathbb{E}\left[\left(\theta_{a_t^*,t}^{(m)} - U_t^{(m)}(a_t^*)\right)^+\right] \tag{16}$$

$$\leq \Pr\left(\theta_{a_t^*,t}^{(m)} \geq U_t^{(m)}(a_t^*)\right) \leq \mathcal{O}(T^{-4}) \tag{17}$$

The first inequality in (17) 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 (17) is due to Lemma 2. Now, consider the first term in (15):

$$\begin{aligned}
\mathbb{E}\left[U_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] &= \mathbb{E}\left[2\delta_{a_t,t} + L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right] \\
&= 2\mathbb{E}\left[\delta_{a_t,t}\right] + \mathbb{E}\left[L_t^{(m)}(a_t) - \theta_{a_t,t}^{(m)}\right]
\end{aligned} \tag{18}$$

Similar to (16) and (17):

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

Combining (8) and (15):

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

Theorem 4: Given D beams, each of which has a volatility σ, the Bayesian regret of the ATS within 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:

$$\mathrm{BR}(T) \leq \left(2M v_{M-1}\mathbb{E}\left[\delta_{a_t,t}\right] + \mathcal{O}\left(T^{-4}\right)\right) \tag{19}$$

$$= \mathcal{O}\left(M\sqrt{\log T}\right)\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{1/n_{a_t,t}}\right] + \mathcal{O}\left(M\min\{1,\sigma\}T\sqrt{8T\log T}\right) \tag{20}$$

Lemma 1 in [33] 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 (20) 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 [33] 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

In this section, we evaluate the performance of MAMBA through extensive experiments. We compare ATS with two other schemes: a dynamic oracle that always selects the best beam and MCS index at each location, and a static oracle that runs an exhaustive scan once and uses the best beam/MCS pair it finds until the next scan. We implemented the dynamic oracle by running an exhaustive scan at each location, without taking into account its search time overhead. Therefore, the performance of the dynamic oracle shows the upper bound for all tracking algorithms. Note that the static oracle is similar to what is proposed for 5G NR, i.e., using the same beam found during IA, until the transmission of the next SS burst.

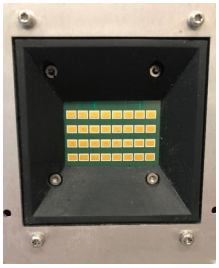
In the experiment setup (see Fig. 5), a 4×8 UPA is used at the Tx side to transmit a continuous wave (CW) signal with 0 dBm amplitude at 28 GHz frequency. Keysight E8267D PSG signal generator is used to generate the signal. At the Rx side,

$$\Pr\left(\left|S_{i,t}^{(m)} - \mathbb{E}[S_{i,t}^{(m)}]\right| \geq \gamma_2\sqrt{2n_{i,t}\log T}\right) \leq \mathcal{O}(T^{-4}) \tag{10}$$

$$\Pr\left(\left|\sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)} - \sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} \theta_{i,\tau_{i,k}}^{(m)}\right| \geq \sqrt{2n_{i,t}\log T}\right) \leq \mathcal{O}(T^{-4}) \tag{11}$$

$$\Pr\left(\left|\sum_{k=1}^{n_{i,t}} \gamma_1^{t-\tau_{i,k}} r_{\tau_{i,k}}^{(m)} - n_{i,t}\theta_{i,t}^{(m)}\right| \geq \sqrt{2n_{i,t}\log T} + n_{i,t}\min\{1,\sigma\}\sqrt{8T\log T}\right) \leq \mathcal{O}(T^{-4}) \tag{12}$$

(a)



(b)

(c)

Fig. 5. Experimental setup used for performance evaluation. (a) Outdoor scenario with 7 m Tx-Rx separation, (b) 4×8 UPA at the Tx side, (c) 20 dBi gain horn antenna at the Rx side.

a 20 dBi gain horn antenna is connected to Keysight 9038A MXE EMI receiver to obtain RSS measurements. The PSG and the EMI receiver are connected to a host PC, and the RSS results are obtained via a TCP 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 variable and obtain a reward by determining the most appropriate MCS through the lookup table of WiGig [2].

We conducted the experiments in two different scenarios: an indoor office (Tx-Rx separation of 3.5 m) and an outdoor environment (Tx-Rx separation of 7 m). Recall that ATS only runs at the Tx side, while the Rx keeps pointing towards the Tx. To ensure that, we moved the Rx along an arc whose center is the Tx. RSS measurements were taken at equally spaced locations on that arc, with 3.5 cm minimum spacing. The number of measurement points depends on the Tx-Rx separation, and hence, it is different for different scenarios. As the Tx and Rx antennas were at the same elevation, beam sweeping was done only in the azimuthal domain. The UPA was steered $\pm 45^\circ$ from broadside with 5° steps.

In Fig. 6(a) and 6(b), we depict the throughput performance of ATS against dynamic and static oracles. Recall that, if the MCS selected by the BS cannot be supported at the UE, the throughput gain from the given transmission will be 0. Throughout the experiments, the slot duration is 1 ms and the Rx speed is 14 cm/slot, unless otherwise specified. In the outdoor scenario with 80 measurement points, a line-of-sight (LOS) path is available at each point. Hence, the dynamic oracle can always achieve the highest rate (see dashed line in Fig. 6(a)). On the other hand, in the indoor office environment with 40 measurement points, the LOS path is blocked at certain locations. Therefore, even the dynamic oracle cannot satisfy the highest MCS index at all points, as seen from the last 10 slots in Fig. 6(b). Notice that in both figures, the performance of all three algorithms are identical in the beginning. This is because when the displacement of the Rx 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. 6(a), the throughput of ATS is 182% higher than that of the static oracle in the outdoor scenario, and only 21% lower than that of the dynamic oracle. For the indoor measurements, Fig. 6(b) shows that the throughput of ATS is 75% higher than that of the static oracle and 27% lower than that of the dynamic oracle. Therefore, in both scenarios, ATS significantly outperforms the static oracle, and performs reasonably close to the dynamic oracle (i.e., low regret).

Fig. 6(c) depicts the evolution of throughput over time for ATS with 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 the exploration and the exploitation, and can achieve the highest throughput.

Next, we compute the CDF of the outage duration (considering both indoor and outdoor scenarios) and display the result in Fig. 6(d). The figure shows how long ATS stays in outage after it loses communication. We observe that with probability exceeding 0.9, the outage will last less than or equal to 5 slots. This result is particularly important for URLLC communications, where latency is of utmost importance. URLLC systems cannot tolerate the lengthy outages resulting from static beam management (see Fig. 6(a) and 6(b)).

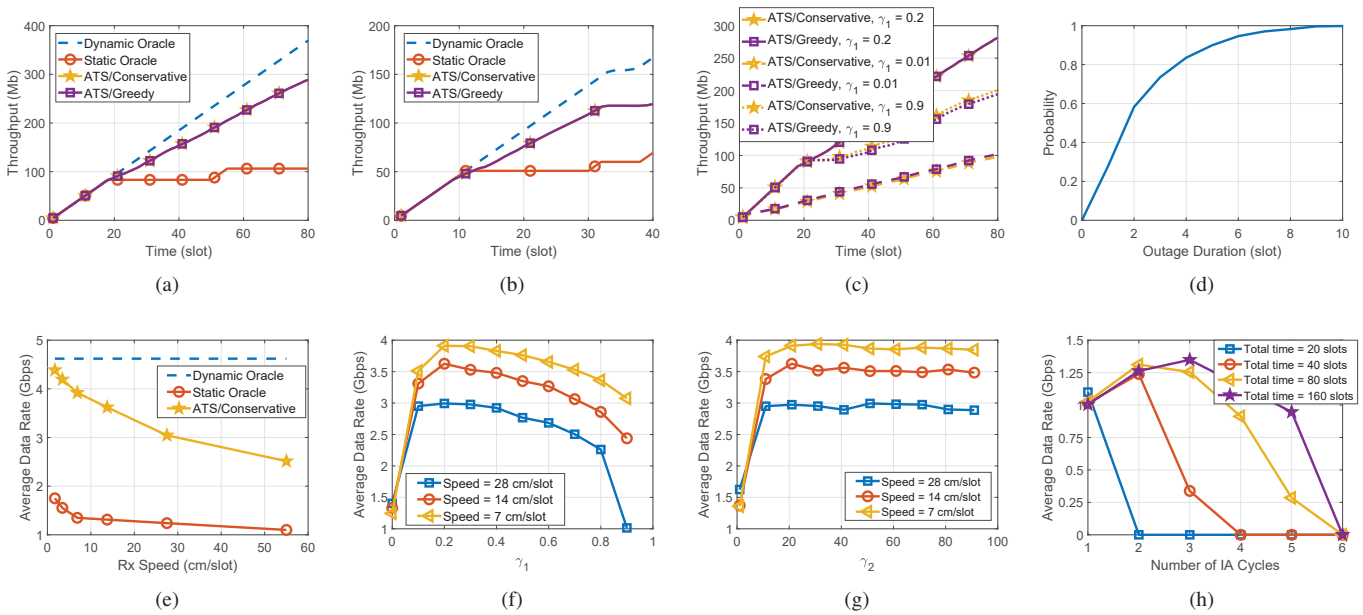


Fig. 6. Performance evaluation results. (a) Throughput vs. time for the outdoor scenario, (b) throughput vs. time for the indoor office scenario, (c) throughput vs. time for ATS with different γ_1 parameters, (d) CDF of the outage duration, (e) average data rate vs. Rx speed, (f) average data rate vs. γ_1 ($\gamma_2 = 20$), (g) average data rate vs. γ_2 ($\gamma_1 = 0.2$), (h) average data rate vs. number of IA cycles for the static oracle.

The effect of Rx speed on the average data rate is studied in Fig. 6(e). 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 w.r.t. 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 speeds will always be less than 10 cm/slot, which translates into 360 km/hour. Therefore, the operating point of a BS will always be on the left end of the figure. Higher speeds are illustrated here to show the trend in the average data rate.

Next, we study the impact of γ_1 and γ_2 on the performance of ATS. Fig. 6(f) 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 previous information is almost instantly forgotten. On the other hand, when γ_1 is too large, the algorithm loses its reactiveness. Fig. 6(f) also shows that when the Rx moves more slowly, a larger γ_1 selection does not reduce the average data rate as much. That is, when the environment changes more slowly, ATS does not need to forget the old information very quickly. The effect of γ_2 is different, as seen in Fig. 6(g). Except for 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.

Finally, in Fig. 6(h), we depict the average data rate under 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 to the system. When the total measurement duration (IA+data) is short, the overhead of rerunning IA becomes more prominent. For that reason, when the total duration is 20 slots, the static oracle can only run one IA cycle, or else, 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.

VII. Conclusions

In this paper, we proposed a multi-armed bandit framework, MAMBA, for beam tracking in mmW systems. We developed a reinforcement learning algorithm, called adaptive Thompson sampling (ATS), used in MAMBA for the selection of beam direction and MCS. ATS uses prior information collected through IA and updates it at each iteration based on the ACK/NACK feedback obtained from the UE. The beam and the MCS used during next transmission are then selected based on the updated posterior distributions. Our experimental results validated the efficiency of ATS and showed that it can improve the throughput by up to 182% compared to a 5G-like beam tracking scheme. Our future work will focus on the dynamic selection of γ_1 and γ_2 for different environments.

Acknowledgments

The authors would like to thank Keysight Technologies for making their mmW experimental setup available. We would also like to thank Dean Gienger and Andrew Smail from Keysight for providing technical expertise regarding RF measurements. This research was supported by NSF (grants IIP-1822071, CNS-1513649, CNS-1563655, and CNS-1731164). 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.

REFERENCES

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE JSAC*, vol. 32, no. 6, pp. 1065–1082, 2014.

[2] 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. [Online]. Available: https://standards.ieee.org/findstds/standard/802.11ad-2012.html

[3] 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.

[4] 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.

[5] 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.

[6] 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, Oct. 2018, pp. 109–117.

[7] 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.

[8] 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 JSAC*, vol. 32, no. 6, pp. 1164–1179, 2014.

[9] 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.

[10] 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, no. 12, pp. 7926–7940, Dec. 2016.

[11] 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'18*, Honolulu, HI, Apr. 2018, pp. 2393–2401.

[12] 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, Aug. 2018, pp. 432–445.

[13] I. Aykin, B. Akgun, and M. Krunz, "Smartlink: Exploiting channel clustering effects for reliable millimeter wave communications," in *Proc. of the IEEE INFOCOM'19*, Paris, France, Apr. 2019, pp. 1117–1125.

[14] 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 (TON)*, vol. 26, no. 5, pp. 2282–2295, 2018.

[15] 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. PP, no. 99, 2019. [Online]. Available: doi.org/10.1109/JSAC.2019.2947938

[16] I. Aykin and M. Krunz, "Efficient beam sweeping algorithms and initial access protocols for millimeter-wave networks," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, 2020. [Online]. Available: doi.org/10.1109/TWC.2020.2965926

[17] 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.

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

[20] 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.

[21] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Proc. of the Advances in Neural Information Processing Systems*, 2011, pp. 2249–2257.

[22] I. Osband and B. Van Roy, "Why is posterior sampling better than optimism for reinforcement learning?" in *Proc. of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, Aug. 2017, pp. 2701–2710.

[23] H. Gupta, A. Eryilmaz, and R. Srikant, "Link rate selection using constrained thompson sampling," in *Proc. of the IEEE INFOCOM'19*, Paris, France, Apr. 2019.

[24] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30 507–30 526, 2019.

[25] 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'18*, Honolulu, HI, Apr. 2018, pp. 1961–1969.

[26] 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, Dec. 2016, pp. 743–747.

[27] 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, May 2016, pp. 1–6.

[28] 3GPP TR 38.802 v14.2.0, "Study on new radio access technology-physical layer aspects (release 14)," Sep. 2017. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/38series/38.802/38802e20.zip

[29] 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.

[30] 3GPP TS 38.213 v15.5.0, "Physical layer procedures for control (release 15)," May 2019.

[31] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

[33] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, Nov. 2014.