mmWave Beam Selection in Dynamic Multi-Path Environments: A POMDP Approach

Ece Bingöl and Eylem Ekici
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH, USA
{bingol.2, ekici.2} @osu.edu

2024 International Conference on Computing, Networking and Communications (ICNC) | 979-8-3503-7099-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICNC59896.2024.10555922

***Abstract*—mmWave systems are integral parts of 5G+ wireless systems. Large bandwidths allocated at above 20GHz translate to large data rates otherwise unattainable at lower frequencies. However, signals in the mmWave bands require highly directional beams to overcome strong attenuation and do not propagate through objects along Line-of-Sight (LoS) paths. In this work, we study scenarios with direct LoS and reflected Non-Line-of-Sight (NLoS) paths, where the LoS paths are blocked temporarily. The so-called beam selection problem aims to choose beams to establish communication between two mmWave enabled devices and determine how long the communication with the chosen beam should last. Considering the system's state, defined as the LoS blockage, is observable in one choice of the beams but not others, we formulate the problem as a generalized case of Partially Observable Markov Decision Process (POMDP). The resulting policies result in the maximization of the reward (throughput) of the system, which are demonstrated through numerical examples.**

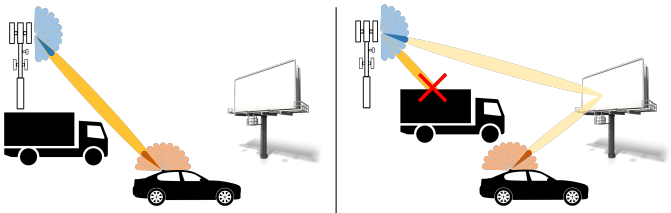
***Index Terms*—mmWave, beam selection, POMDP, 5G+**

I. Introduction

With the advent of 5G wireless systems, mmWave communication systems have found their first large-scale deployment [1]. At these high frequencies, large bandwidths allow high communication data rates. Line-of-Sight channels provide the strongest received signal strengths, therefore the highest data rates, while Non-Line-of-Sight (NLoS) channels utilizing reflections have smaller gains, hence lower rates. However, the attenuation in the mmWave bands is also very severe, which is countered by using highly focused beams, either through the use of horn antennas or using MIMO antenna arrays. The resulting directionality of communication is not only beneficial to overcome strong attenuation, but also isolates out most interference sources vis-a-vis sub-6GHz communication scenarios. On the other hand, this makes link maintenance more challenging as the signals are virtually blocked by any solid object along the path, requiring dynamic selection of beams to sustain seamless communication [2].

Over the years, numerous beam tracking and beam switching approaches have been proposed to overcome this problem [3]. [4] proposes switching to an NLoS link in the case of LoS blockages and evaluates two kinds of beam switching mechanisms for an indoor scenario. [5] scans all possible beam combinations and uses an Extended Kalman Filter (EKF) to track the Angle of Arrival (AoA) and Angle of Departure (AoD). [6] takes a similar approach but instead of a full scan, uses only a single measurement. [7] models the channels using a Markovian random walk and poses the problem as a POMDP. [8] models the evolution of AoD as a discrete Markov process and uses maximum aposteriori (MAP) estimation for tracking. [9] develops a model free beam tracking algorithm using Q-learning. The authors utilize auxiliary beam pairs to estimate AoA and AoD.

This work has been supported in part by NSF through grants 1955535, 2030141, 2112471.



(a) LoS (b) Blocked LoS

Fig. 1: LoS and NLoS Communication Examples

In this paper, we explore a case where the strong LoS path is subject to temporary blockages, requiring switching over to weaker NLoS paths. A sample scenario is depicted in Figure 1. The LoS path presents the highest throughput, but is blocked for random durations. In such cases, the reflective NLoS path is utilized, during which the availability of the LoS path remains unknown. The ON/OFF state of the LoS path can only be observed by switching to the LoS beam. Sensing the LoS path state is a potential loss of transmission opportunity over the NLoS path. However, if the transmitter waits too long before sensing the LoS channel, then the system stands to lose the opportunity of transmitting at a higher rate in case the blockage is lifted earlier. Hence, the problem is finding a policy of estimating the LoS channel availability, and if unavailable, determining the optimum time to sample it next.

The state of the LoS channel can be observed at every transmission attempt when the associated beam is used. When the alternate beam is used, the LoS channel state cannot be observed. The system is, therefore, partially observable. We also consider arbitrary sojourn time distributions for

state occupancy. For the sake of tractability, we consider two stationary endpoints with only one alternative NLoS path which is never blocked. Moreover, it is assumed that communication occurs over a single beam and multi-beam combining solutions are not utilized. The resulting system cannot be directly mapped to well-studied problems such as POMDP. To this end, we expand the definition of the system state from ON/OFF (corresponding to the availability/blockage of the LoS path, respectively) to one that includes both channel state as well as the time elapsed since the last state change (sojourn time), discretized and tracked by a counter that resets at every channel state change. Then, we formulate the problem as a POMDP on this expanded system state definition. The resulting policy can be pre-computed and stored for online usage.

II. Related Work

The idea of utilizing NLoS paths to resolve blockages was proposed in [4]. The authors consider an indoor scenario and utilize a random waypoint mobility model for blockages. Based on this they propose two switching mechanisms. However, regardless of the switching mechanism, they state that the devices should keep probing the LoS channel and switch back to it as soon as it becomes available. This means some resource must be allocated to monitoring the LoS channel which otherwise could be used for data transmission. In this work we address this by adjusting the probing instances using channel statistics.

[7] considers the problem of selecting pilot beam directions to detect the LoS and NLoS paths. The authors assume that the receiver and reflectors move according to some Markovian random walk and formulate the problem as a POMDP. They also provide a suboptimal greedy algorithm to reduce the computational complexity of finding an exact solution. Nonetheless, their main focus is on tracking the LoS and NLoS channels under user mobility. Therefore, the states (AoDs) are assumed to be more likely to shift to nearby states. Path movements with large AoD change is ignored. We, on the other hand, consider a static transmitter and receiver pair, but focus on mitigating the blockages of the LoS path. We also relax the Markovian assumption on state transitions.

When the transmitter probes the LoS channel, it observes the associated path gain, which is assumed to be continuous in this paper. POMDPs with continuous observation space are notorious for their difficulty. Thus, [10] proposes an observation aggregation method. The main idea behind this method is that, although the observation space is rich, most of the good policies select the same course of action for a range of observations. This enables the observation space to be discretized implicitly. We employed this method when finding the value function of the optimal policy.

III. System Model

In this work we consider a receiver and transmitter pair in a dynamic propagation environment with a strong LoS path and another weaker NLoS path. Time is slotted $t = 1, 2, \ldots$ and the transmitter selects a beam in each time slot to communicate with the receiver. We assume that the LoS channel is subject to temporary blockages due to the environment but the NLoS channel is always available. We denote the ON/OFF state of the LoS channel in slot t with $s_t \in \{0, 1\}$ where $s_t = 0$ corresponds to the OFF case and $s_t = 1$ to the ON case. Moreover, we denote the ON/OFF duration of the LoS channel with random variables T_{ON} and T_{OFF}, respectively. Their probability density functions (pdf) are given as $T_{\text{ON}} \sim g_{\text{ON}}(u)$ and $T_{\text{OFF}} \sim g_{\text{OFF}}(u)$. Since time is slotted, T_{ON} and T_{OFF} are discretized by sampling with some frequency f.

After selecting a beam x_t in time slot t from a set of beams $\mathcal{X}$, the transmitter observes a noisy reward $y_t = f_t(x_t) + n_t$, where $f_t(\cdot)$ is the reward function at time slot t and n_t is an independent and identically distributed (i.i.d.) noise term with $n_t \sim \mathcal{N}(0, \sigma^2)$. The reward functions conditioned on the channel state are known by the transmitter. Thus, given s_t, the transmitter deterministically selects beam $x_t^* = \arg\max_{x \in \mathcal{X}} f(x|s_t)$ and observes a reward $y_t = f(x_t^*|s_t) + n_t$. We denote the maximum reward conditioned on the channel state as $r_s \triangleq f(x^*|s)$. The reward of the LoS channel is higher than NLoS channel, i.e., $r_1 > r_0$. Then, the reward observations based on the channel state become

$$\begin{aligned} y_t^{\text{LoS}} &= n_t, & y^{\text{NLoS}} &= r_0 + n_t, \quad s_t = 0 \\ y_t^{\text{LoS}} &= r_1 + n_t, & y_t^{\text{NLoS}} &= r_0 + n_t, \quad s_t = 1 \end{aligned} \tag{1}$$

The transmitter cannot observe the state while using the NLoS channel as y_t^{NLoS} does not depend on the state. Sensing the LoS channel might cause a loss over the reward of NLoS channel, if the state is OFF. However, staying on the NLoS path too long might also lead to losing the opportunity for higher reward on the LoS channel if the state becomes ON in the meantime. Therefore, the transmitter senses the LoS channel after spending some time in the NLoS channel and determines the channel state and availability duration based on its past decisions and observations. We call the period between each decision an epoch. An epoch might encompass a single time slot or multiple time slots. However, we assume that at most one state transition can happen during an epoch.

As the state is not always observable, the transmitter retains a belief vector $\mathbf{b}$ regarding the channel state. If the true state of the system at the beginning of epoch k is denoted by s_k, the associated belief vector $\mathbf{b}_k$ denotes the state probabilities:

$$\mathbf{b}_k = \begin{bmatrix} P(s_k = 0) & P(s_k = 1) \end{bmatrix}. \tag{2}$$

It shows the confidence of the transmitter regarding the true state of the system. Moreover, since and $g_{\text{ON}}(u)$ and $g_{\text{OFF}}(u)$ are some arbitrary pdfs, the transmitter keeps a counter which shows how much time has passed since the last channel switch by the transmitter. Let $a_k = (\hat{s}_k, \tau_k)$ be the action taken at the beginning of epoch k, where $\hat{s}_k \in \{0, 1\}$ is the selected channel and τ_k is the epoch duration. Then the counter c_k at the beginning of epoch k is computed iteratively as

$$c_k = \tau_{k-1} + c_{k-1} \mathbb{I}\{\hat{s}_{k-1} = \hat{s}_{k-2}\}, \tag{3}$$

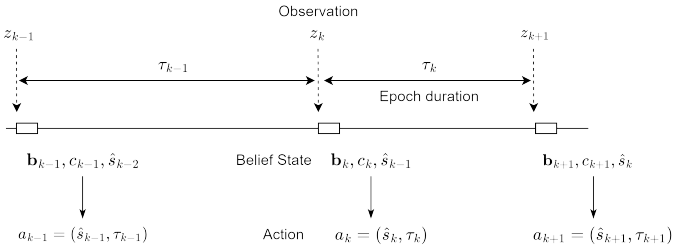


Fig. 2: Belief MDP evolution with time

where $\mathbb{I}\{\cdot\}$ is the indicator function. The update rule in (3) states that, given the previous channel decisions $\hat{s}_{k-2}$, $\hat{s}_{k-1}$, epoch duration τ_{k-1} and counter value c_{k-1}, the new counter value at the beginning of epoch k resets to τ_{k-1} if the transmitter switches channels in the previous epoch. Otherwise, the counter is incremented by τ_{k-1}. The counter value at the beginning of epoch k can be calculated as soon as the action $a_{k-1} = (\hat{s}_{k-1}, \tau_{k-1})$ is taken. Hence, the transmitter does not need to wait until the end of epoch $k-1$ to know c_k.

The transmitter computes a statistic z_k from a series of reward observations at the beginning of each epoch. If $\hat{s}_{k-1} = 0$, i.e., the NLoS path is chosen in the previous epoch, the transmitter samples the LoS channel for N slots at the beginning of the epoch and z_k is the average of those N reward observations. Otherwise, z_k is the moving average of the last N samples taken from the LoS channel. We assume that all samples come from the same distribution. In other words, we assume that the state doesn't change during the sampling period. The transmitter updates its belief $\mathbf{b}_k$ using $\mathbf{b}_{k-1}$, c_k, $\hat{s}_{k-1}$, τ_{k-1} and z_k according to some update rule $h(.)$

$$\mathbf{b}_k = h(\mathbf{b}_{k-1}, c_k, \hat{s}_{k-1}, \tau_{k-1}, z_k). \tag{4}$$

The goal of the transmitter is to find a policy $\pi \in \Pi$ that maps $(b_k, c_k, \hat{s}_{k-1})$ to an action $a_k = (\hat{s}_k, \tau_k)$ such that the discounted cumulative reward is maximized.

$$\pi^* = \arg\max_{\pi \in \Pi} \sum_{t=0}^{\infty} E\big[\gamma^t f_t(x_t)\big], \tag{5}$$

where $\gamma \in (0,1)$ is the discount factor, and Π is the set of policies. We consider stationary policies in this paper.

IV. Solution Approach

The POMDP framework [11] is suitable for this problem except that the state holding times T_{ON} and T_{OFF} have arbitrary pdfs, i.e., they might be non-Markovian. Therefore, we map this problem to a POMDP by extending the state definition to include the counter $\mathfrak{c}$ and previous channel selection $\hat{s}_p$. The state space of the resulting POMDP is $(s, \mathfrak{c}, \hat{s}_p) \in \{0,1\} \times \mathbb{N}^+ \times \{0,1\}$, where $\mathbb{N}^+$ is the set of positive integers. Similarly, the action space consists of tuples $a = (\hat{s}, \tau) \in \{0,1\} \times \mathbb{N}^+$. $\hat{s}$ is the current channel selection and τ is the epoch duration.

The state of the system is unveiled to the transmitter through rewards obtained from the LoS channel. Each observation is the average of N i.i.d. reward samples. Using the reward expressions given in (1) we have the conditional pdf of an observation z as

$$\begin{aligned} p_0(z) &\triangleq f_Z(z|s=0) = \mathcal{N}(0, \sigma^2/N) \\ p_1(z) &\triangleq f_Z(z|s=1) = \mathcal{N}(r_1, \sigma^2/N). \end{aligned} \tag{6}$$

TABLE I
Transition Probabilities for $P(s'|s, c, \hat{s}_p, a)$

s	$\hat{s}_p$	$P(0\|s,c,\hat{s}_p,a)$	$P(1\|s,c,\hat{s}_p,a)$
0	0	$P(T_{\text{OFF}} > c+\tau \mid T_{\text{OFF}} > c)$	$P(T_{\text{OFF}} \le c+\tau \mid T_{\text{OFF}} > c)$
0	1	$P(T_{\text{OFF}} > \tau)$	$P(T_{\text{OFF}} \le \tau)$
1	0	$P(T_{\text{ON}} \le \tau)$	$P(T_{\text{ON}} > \tau)$
1	1	$P(T_{\text{ON}} \le c+\tau \mid T_{\text{ON}} > c)$	$P(T_{\text{ON}} > c+\tau \mid T_{\text{ON}} > c)$

Neither the true state s of the system, nor the true counter $\mathfrak{c}$ is available to the transmitter. Therefore, the transmitter constructs a probability distribution called belief over the true state s by interacting with the environment. The belief vector $\mathbf{b}$ is a two-element vector with each element corresponding to $b(i) \triangleq P(s=i)$, $i \in \{0,1\}$. Since the true counter value $\mathfrak{c}$ is unobservable, the transmitter keeps its own counter c. It counts how many time slots have passed since the decision of the transmitter changed from 0 to 1 or vice versa. c reflects the true counter value $\mathfrak{c}$ to the best of the transmitter's knowledge under uncertainty. According to this formulation, given an action $a = (\hat{s}, \tau)$, c and $\hat{s}_p$ deterministically transition to $c' = \tau + c\,\mathbb{I}\{\hat{s} = \hat{s}_p\}$ and $\hat{s}'_p = \hat{s}$, respectively. The transition probabilities of the true state $P(s'|s, c, \hat{s}_p, a)$ for $s' \in \{0,1\}$ are given in Table I.

The transmitter obtains an instantaneous reward $R(s, c, \hat{s}_p, a)$ after executing action $a = (\hat{s}, \tau)$ when the true state is s, counter is c and previous decision is $\hat{s}_p$:

$$R(s,c,\hat{s}_p,a) = \begin{cases} (r_0 - r_1)\max\{\tau - E[V_{\text{OFF}}], 0\} & s=0,\ \hat{s}=0 \\ -r_0\tau & s=0,\ \hat{s}=1 \\ (r_0 - r_1)\min\{E[V_{\text{ON}}], \tau\} & s=1,\ \hat{s}=0 \\ 0 & s=1,\ \hat{s}=1 \end{cases} \tag{7}$$

We set the reward as the negative of the expected regret. Hence, each line of (7) is equal to the negative of the expected regret incurred for each case. V_{ON} and V_{OFF} denote the residual life duration of T_{ON} and T_{OFF}, respectively. Pdfs of V_{ON} and V_{OFF} can be found by conditioning T_{ON} and T_{OFF} on the counter c and the previous decision $\hat{s}_p$:

$$\begin{aligned} &P(V_{\text{ON}} \le v|c, \hat{s}_p = 0) = P(T_{\text{ON}} \le v) \\ &P(V_{\text{ON}} \le v|c, \hat{s}_p = 1) = P(T_{\text{ON}} \le c+v|T_{\text{ON}} > c) \\ &P(V_{\text{OFF}} \le v|c, \hat{s}_p = 0) = P(T_{\text{OFF}} \le v+c|T_{\text{OFF}} > c) \\ &P(V_{\text{OFF}} \le v|c, \hat{s}_p = 1) = P(T_{\text{OFF}} \le v) \end{aligned} \tag{8}$$

Using $\mathbf{b}$ and c, we can formulate the POMDP as a belief MDP in which the objective is to find a stationary policy π^* that maps $(\mathbf{b}, c, \hat{s}_p)$ to an action $a = (\hat{s}, \tau)$ such that the discounted cumulative reward (5) is maximized.

In order to understand the properties of this belief MDP, we need to investigate the evolution of the system with time as

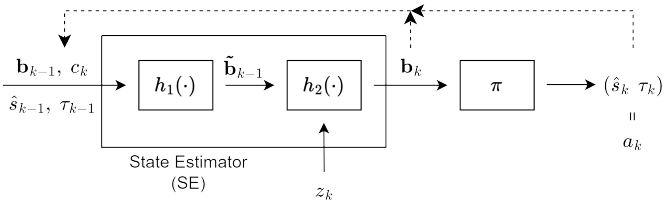


Fig. 3: Functional breakdown of the solution

shown in Figure 2. At the beginning of each slot, the belief state is updated according to some update rule (4). Resultant state $(\mathbf{b}_k, c_k\hat{s}_{k-1})$ is then mapped to an action $a_k = (\hat{s}_k, \tau_k)$ for that slot. Lastly, the counter c_k is updated according to (3). To solve this problem, we first need to determine the belief update rule in (4), and then find the optimal policy.

Figure 3 shows the two fundamental blocks of our solution. State estimator (SE) performs the belief update in (4). It can be further separated into two functions. The first one, $h_1(\cdot)$, handles the time aspect of belief evolution. We call its output *belief prior* and denote it with $\tilde{\mathbf{b}}$. Then, $\tilde{\mathbf{b}}_{k-1}$ corresponds to the belief vector at the beginning of epoch k, right before the observation z_k is received. Define $\tilde{b}_{k-1}(i) \triangleq P(s_k = i|\mathbf{b}_{k-1}, c_k, \hat{s}_{k-1}, \tau_{k-1})$ for $i \in \{0, 1\}$. Convert the counter and epoch duration to seconds as $\bar{c}_k = c_k/f$ and $\bar{\tau}_{k-1} = \tau_{k-1}/f$, respectively. Using Bayes' Theorem

$$\tilde{b}_{k-1}(0) = \begin{cases} \dfrac{P(T_{\text{OFF}} > \bar{c}_k)}{P(T_{\text{OFF}} > \bar{c}_k - \bar{\tau}_{k-1})} b_{k-1}(0) & \hat{s}_{k-1} = 0 \quad (9a) \\ 1 - \dfrac{P(T_{\text{ON}} > \bar{c}_k)}{P(T_{\text{ON}} > \bar{c}_k - \bar{\tau}_{k-1})} b_{k-1}(1) & \hat{s}_{k-1} = 1 \quad (9b) \end{cases}$$

The second function, $h_2(\cdot)$, updates $\tilde{\mathbf{b}}_{k-1}$, according to z_k. It can also be derived from Bayes' Theorem as

$$b_k(i) = \frac{\tilde{b}_{k-1}(i)p_i(z_k)}{\tilde{b}_{k-1}(0)p_0(z_k) + \tilde{b}_{k-1}(1)p_1(z_k)} \quad i \in \{0, 1\}. \quad (10)$$

(9) and (10) together make up the belief update rule in (4).

After the belief update, the policy π in Figure 3 maps $(\mathbf{b}_k, c_k, \hat{s}_{k-1})$ to an action $a_k = (\hat{s}_k, \tau_k)$. Since we consider stationary policies, the epoch index k can be dropped and the policy can be expressed as a mapping from $(\mathbf{b}, c, \hat{s}_p)$ to action $a = (\hat{s}, \tau)$.

In a standard POMDP problem, where the action and observation spaces are finite, the optimal policy can be found by solving a set of linear equations [11]. Even then, the computational complexity of the problem grows exponentially with the number of actions and observations. Both the action and observation spaces of this problem have continuous elements. Regarding the former, we focus on stationary deterministic policies such that the epoch duration τ is a deterministic function of the state $(\mathbf{b}, c, \hat{s}_p)$ and the current decision $\hat{s}$ for that epoch. While $\hat{s} = 1$, the LoS channel is constantly being monitored. The epoch duration is just 1 slot, or $1/f$ seconds in that case. However, when $\hat{s} = 0$, the transmitter waits for some time before sampling the LoS channel again. Had the exact transition instant from ON to OFF been known, the optimal (oracle) epoch duration could have been found by minimizing the expected loss (regret) incurred during that period. Say the transmitter waits τ_1^O seconds before probing the LoS channel again, under this scenario. If it probes early, while the channel is still OFF, it receives a penalty of $r_2 N/f$, where N/f is the sampling duration in seconds. On the other hand, if it probes the channel late, it misses the opportunity of higher reward. Thus, the expected penalty is $(r_1 - r_0)(\tau_1^O - E[T_{\text{OFF}}])$. We assume that if the channel becomes available during the sampling period, no penalty is received. Minimizing the expected penalty with respect to τ_1^O yields that the optimal waiting duration $\tau_1^{O^*}$ satisfies

$$\frac{P(T_{\text{OFF}} < \tau_1^{O^*})}{g_{\text{OFF}}(\tau_1^{O^*} + N/f)} = \frac{r_0 N}{(r_1 - r_0)f}. \quad (11)$$

The same method can be applied to any waiting duration τ_n^O given that $T_{\text{OFF}} > \sum_{i=1}^{n-1} \tau_i^O$. Thus we can generalize (11) as

$$\frac{P\big(\sum_{i=1}^{n-1} \tau_i^O < T_{\text{OFF}} < \sum_{i=1}^{n-1} \tau_i^O + \tau_n^{O^*}\big)}{g_{\text{OFF}}\big(\sum_{i=1}^{n-1} \tau_i^O + \tau_n^{O^*}\big)} = \frac{r_0 N}{(r_1 - r_0)f}, \quad (12)$$

where $\sum_{i=1}^{n-1} \tau_i^O = 0$ for $n = 1$. In the original problem, without the oracle, c counts the consecutive time slots during which the decision of the transmitter has been the same so far. In other words, it corresponds to how long the transmitter has been using a certain channel consecutively. Therefore, as soon as the decision $\hat{s} = 0$ has been made, it can be updated to an intermediate variable $\tilde{c} = \big(c/f\big)\mathbb{I}\{\hat{s} = \hat{s}_p\}$. Assuming the information at the transmitter is correct, we can use (12) and calculate an intermediate variable τ^O for epoch duration

$$\frac{P(\tilde{c} \le T_{\text{OFF}} < \tilde{c} + \tau^O)}{g_{\text{OFF}}(\tilde{c} + N/f + \tau^O)} = \frac{r_0 N}{(r_1 - r_0)f}. \quad (13)$$

Since $\mathbf{b}$ quantifies the uncertainty in the true channel state, we scale τ^O by $b(0)$ to get the epoch duration under uncertainty.

$$\tau = \begin{cases} \max\{\lfloor \tau^O f b(0) \rceil, 1\} & \hat{s} = 0 \quad (14a) \\ 1 & \hat{s} = 1 \quad (14b) \end{cases}$$

where $\lfloor \cdot \rceil$ represents rounding to nearest integer. This particular τ choice was motivated by the fact that as the transmitter's confidence about the channel state being OFF weakens, it should check the LoS channel more frequently.

To handle the continuity stemming from observations, we use the observation aggregation method proposed in [10]. The authors partition the observation space using the concept of conditional plans. A conditional plan $cp = \langle a, \nu(\cdot) \rangle$ consists of an action and an observation strategy. Together they specify which action to perform and which conditional plan to execute next, contingent on the observation. The value function of a conditional plan, V_{cp}, which is the expected cumulative reward to be obtained when starting from a state $(\mathbf{b}, c, \hat{s}_p)$, is linear with respect to the belief:

$$\begin{aligned} V_{cp}(\mathbf{b}, c, \hat{s}_p) &= \sum_s b(s) V_{cp}(s, c, \hat{s}_p) \\ &= \sum_s b(s) \alpha_{cp}(s, c, \hat{s}_p). \end{aligned} \quad (15)$$

Although the belief space is continuous, the value function of a conditional plan can be completely characterized by a finite set of parameters called an α-vector. Note that each α-vector is associated with an action. By definition, the optimal policy π^* achieves the highest return at all states. Thus, $V_{\pi^*}(\mathbf{b}, c, \hat{s}_p)$ is the upper surface of the α-vectors.

We define conditional plans 0 and 1 as using the NLoS and LoS channels, respectively. Since each conditional plan is associated with an action, $a_m = (m, \tau)$ for $m \in \{0, 1\}$, where τ is calculated according to (13) and (14). Let the updated state after executing an action $a = (\hat{s}, \tau)$ and observing z be $(\mathbf{b}_z^a, c^a, \hat{s}_p^a)$. Note that $\hat{s}_p^a = \hat{s}$, and c^a is given by (3). [10] partitions the observation space into regions $\mathcal{Z}_m$ in which the conditional plan m yields the highest return for $(\mathbf{b}_z^a, c^a, \hat{s}_p^a)$

$$\mathcal{Z}_m = \{z | m = \arg \max_{i \in \{0,1\}} \alpha_i(\mathbf{b}_z^a, c^a, \hat{s}_p^a)\}. \tag{16}$$

Define the probability $P(\mathcal{Z}_m | a, s')$ as an observation z such that $z \in \mathcal{Z}_m$ will be made if action a is taken and true state s' is reached as a result. In this problem, the observations only depend on the true state. Hence, it can be calculated by integrating (6) over the region $\mathcal{Z}_m$.

For a one-dimensional observation space, the regions in (16) are line segments. Two conditional plans have the same value at the segment boundaries. Hence, the segment boundaries can be found by solving $\alpha_0(\mathbf{b}_z^a, c^a, \hat{s}_p^a) - \alpha_1(\mathbf{b}_z^a, c^a, \hat{s}_p^a) = 0$. Notice that the initial state $(\mathbf{b}, c, \hat{s}_p)$ and action $a = (\hat{s}, \tau)$ are fixed, i.e., the only variable is z.

The optimal value function $V_{\pi^*}(\mathbf{b}, c, \hat{s}_p)$ can be computed by value iteration. At each iteration, the segment boundaries are calculated and α-vectors are updated by point-based dynamic programming backups [10].

$$\begin{aligned}\alpha_m(\mathbf{b}, c, \hat{s}_p) = &\sum_s b(s) R(s, c, \hat{s}_p, a_m) \\ &+ \gamma \sum_{i=0}^{1} P(\mathcal{Z}_i | \mathbf{b}, c, \hat{s}_p, a_m) \alpha_i(\mathbf{b}_{\mathcal{Z}_i}^{a_m}, c^{a_m}, m),\end{aligned} \tag{17}$$

where $m \in \{0, 1\}$ and $R(s, c, \hat{s}_p, a)$ is the immediate reward given in (7). $P(\mathcal{Z}_i | \mathbf{b}, c, \hat{s}_p, a_m)$ is the probability that the observation comes from region $\mathcal{Z}_i$ for some $i \in \{0, 1\}$ conditioned on the state $(\mathbf{b}, c, \hat{s}_p)$ and the action taken is a_m

$$P(\mathcal{Z}_i | \mathbf{b}, c, \hat{s}_p, a_m) = \sum_{s, s'} P(\mathcal{Z}_i | a_m, s') P(s' | s, c, \hat{s}_p, a_m) b(s). \tag{18}$$

The transition probabilities $P(s' | s, c, \hat{s}_p, a)$ are given in Table I. Similarly, $\mathbf{b}_{\mathcal{Z}_i}^{a_m}$ is the updated belief vector after taking action a_m and making an observation $z \in \mathcal{Z}_i$

$$b_{\mathcal{Z}_i}^{a_m}(s') = \frac{\sum_s P(\mathcal{Z}_i | a_m, s') P(s' | s, c, \hat{s}_p, a_m) b(s)}{\sum_{s, \tilde{s}} P(\mathcal{Z}_i | a_m, \tilde{s}) P(\tilde{s} | s, c, \hat{s}_p, a_m) b(s)}. \tag{19}$$

$V_{\pi^*}(\mathbf{b}, c, \hat{s}_p)$ is the highest return achieved by any of the conditional plans:

$$V_{\pi^*}(\mathbf{b}, c, \hat{s}_p) = \max\{\alpha_0(\mathbf{b}, c, \hat{s}_p),\ \alpha_1(\mathbf{b}, c, \hat{s}_p)\}. \tag{20}$$

Then, we can identify the best action for each state $(\mathbf{b}, c, \hat{s}_p)$ from corresponding α-vectors

$$\hat{s} = \arg \max_{i \in \{0,1\}} \alpha_i(\mathbf{b}, c, \hat{s}_p), \tag{21}$$

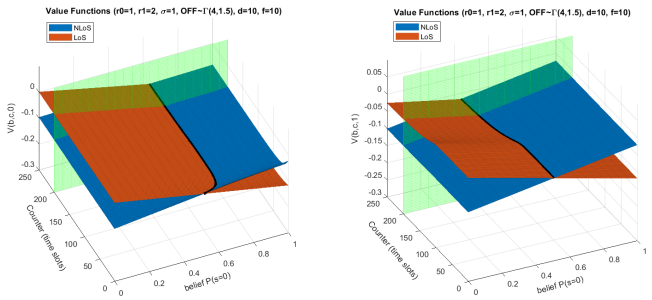
where $\alpha_i(\mathbf{b}, c, \hat{s}_p)$ for $i \in \{0, 1\}$ are obtained through (17). Once $\hat{s}$ is determined, τ is calculated using (13) and (14). Then we form the action $a = (\hat{s}, \tau)$.

V. Numerical Examples

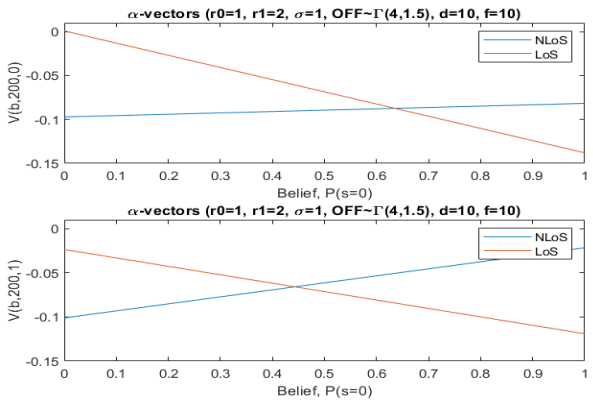
In this section, we present the α-vectors (17) and $V_{\pi^*}(\mathbf{b}, c, \hat{s}_p)$ (20) for two different sojourn time distributions. We take $r_0 = 1$ and $r_1 = 2$ for the rewards. The transmitter takes the average of $N = 2$ samples when computing z. The noise variance is set to $\sigma^2 = 1$, the sampling frequency $f = 10$ Hz and the discount factor $\gamma = 0.9$.

First, T_{OFF} is selected as a Gamma random variable with shape parameter 4 and scale parameter 1.5, i.e., $g_{\text{OFF}}(u) \sim \Gamma(4, 1.5),\ u \geq 0$. Similarly, T_{ON} is taken as a shifted Gamma random variable with the same shape and scale parameters and shift factor $d = 10$. The value functions of the conditional plans are shown in Figures 4a and 4b. The intersection between the two surfaces is highlighted with a black line in both figures. The value function of the optimal policy is the upper surface of the two curves. Thus, if the red curve is above the blue one for a given state, then the LoS channel is used as it yields a higher return. Otherwise, the NLoS channel is utilized. Figure 4c shows a cross-section of Figures 4a and 4b at $c = 200$. The first sub-figure is for $\hat{s}_p = 0$, hence it corresponds to the transmitter being in state 0 for 200 slots. In that case, α_1 is above α_0 until $b(0) \approx 0.65$, which means, unless the transmitter is confident that the state is 0, the optimal action is to use the LoS channel. This is because the average OFF duration is 60 time slots, which is much smaller than 200. Thus, the probability of true state being 1 is very high, unless the observations strongly suggest otherwise. The second sub-figure in Figure 4c shows the α-vectors for $\hat{s}_p = 1$. This time, the transmitter has been in state 1 for 200 time slots, which is slightly longer than the average ON duration, 160 slots. Therefore, if $b(0) \gtrapprox 0.4$, the optimal action is to use the NLoS channel.

Next, we perform the same analysis for uniform sojourn time distributions, $g_{\text{OFF}}(u) \sim U[0, 4]$ and $g_{\text{ON}}(u) \sim U[10, 30]$. The results are shown in Figure 5. The region of the NLoS channel in Figure 5a is significantly restricted compared to Figure 4a. The reason is, since T_{OFF} can last at most 40 time slots, the optimal action is to check the LoS channel after staying in the NLoS channel for more than 40 slots. In accordance with this, we see that α_1 completely dominates α_0 in the first sub-figure of Figure 5c. Meanwhile, the second sub-figure shows that NLoS channel is optimal for $b(0) \gtrapprox 0.5$ for $\hat{s}_p = 1$, because the average ON duration is also 200 slots. Moreover, r_1 and r_0 are close enough in these simulations, such that the risk of getting a 0 reward from the LoS channel when $c = 200$ is too high compared to a constant reward of 1 to be obtained from the NLoS channel.

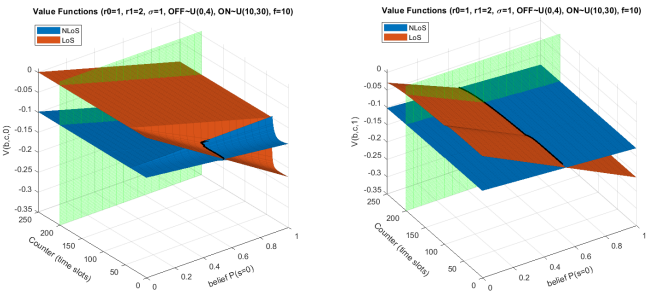


(a) Value functions, $V(\mathbf{b}, c, 0)$ (b) Value functions, $V(\mathbf{b}, c, 1)$

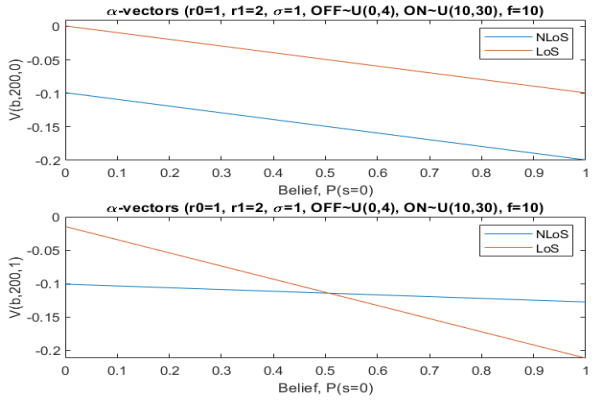


(c) α-vectors for $c = 200$

Fig. 4: Optimal policy for Gamma sojourn time distributions.



(a) Value functions, $V(\mathbf{b}, c, 0)$ (b) Value functions, $V(\mathbf{b}, c, 1)$



(c) α-vectors for $c = 200$

Fig. 5: Optimal policy for Uniform sojourn time distributions.

VI. Conclusions and Future Work

In this work, we investigated the beam selection problem in a mmWave system where a strong LoS path with time-varying availability and a weaker NLoS always-available path exist. The problem is then formulated as estimating the state of the system and the duration of the estimated state (called epoch). The resulting system is analyzed as a POMDP by expanding the simple channel state definition (ON/OFF) to include a discretized representation of the epoch duration. The policy can be precomputed and stored for online access. Numerical results reveal non-trivial relationships between prior state estimation, epoch duration, and sampling results in estimating the channel state and the duration of the upcoming epoch.

This work constitutes a first step towards the analysis of a more general system. In our future work, we will first generalize the solution to a larger number of NLoS paths and consider time-varying availability of the alternative paths. Moreover, we will explore unknown distributions of ON/OFF durations as well as reward functions.

References

[1] K. Sakaguchi, T. Haustein, S. Barbarossa, E. C. Strinati, A. Clemente, G. Destino, A. Pärssinen, I. Kim, H. Chung, J. Kim, W. Keusgen, R. J. Weiler, K. Takinami, E. Ceci, A. Sadri, L. Xain, A. Maltsev, G. K. Tran, H. Ogawa, K. Mahler, and R. W. H. J. au2, "Where, when, and how mmwave is used in 5g and beyond," 2017. [Online]. Available: https://arxiv.org/abs/1704.08131

[2] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave (mmwave) communications for 5g: Opportunities and challenges," 2015. [Online]. Available: https://arxiv.org/abs/1502.07228

[3] Y. Wang, Z. Wei, and Z. Feng, "Beam training and tracking in mmwave communication: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2205.10169

[4] X. An, C.-S. Sum, R. V. Prasad, J. Wang, Z. Lan, J. Wang, R. Hekmat, H. Harada, and I. Niemegeers, "Beam switching support to resolve link-blockage problem in 60 ghz wpans," in *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2009, pp. 390–394.

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

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

[7] J. Seo, Y. Sung, G. Lee, and D. Kim, "Training beam sequence design for millimeter-wave mimo systems: A pomdp framework," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1228–1242, 2016.

[8] D. Zhang, M. Xiao, and M. Skoglund, "Beam tracking for dynamic mmwave channels: A new training beam sequence design approach," in *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, 2022, pp. 276–282.

[9] S. Kim, G. Kwon, and H. Park, "Q-learning-based low complexity beam tracking for mmwave beamforming system," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 1451–1455.

[10] J. Hoey and P. Poupart, "Solving pomdps with continuous or large discrete observation spaces," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, ser. IJCAI'05. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005, p. 1332–1338.

[11] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.