

Second-best Beam-Alignment via Bayesian Multi-Armed Bandits

Muddassar Hussain, Nicolò Michelusi

Abstract—Millimeter-wave (mm-wave) systems rely on narrow-beams to cope with the severe signal attenuation in the mm-wave frequency band. However, susceptibility to beam misalignment due to mobility or blockage requires the use of beam-alignment schemes, with huge cost in terms of overhead and use of system resources. In this paper, a beam-alignment scheme is proposed based on Bayesian multi-armed bandits, with the goal to maximize the alignment probability and the data-communication throughput. A Bayesian approach is proposed, by considering the state as a posterior distribution over angles of arrival (AoA) and of departure (AoD), given the history of feedback signaling and of beam pairs scanned by the base-station (BS) and the user-end (UE). A simplified sufficient statistic for optimal control is identified, in the form of preference of BS-UE beam pairs. By bounding a value function, the *second-best preference* policy is formulated, which strikes an optimal balance between exploration and exploitation by selecting the beam pair with the current *second-best* preference. Through Monte-Carlo simulation with analog beamforming, the superior performance of the *second-best preference* policy is demonstrated in comparison to existing schemes based on *first-best preference*, linear Thompson sampling, and upper confidence bounds, with up to 7%, 10% and 30% improvements in alignment probability, respectively.

Index Terms—Millimeter-wave, beam-alignment, multi-armed bandits, Markov decision process

I. INTRODUCTION

Millimeter-wave (mm-wave) technology has emerged as a promising solution to meet the demands of future communication systems supporting high capacity and mobility, thanks to abundant bandwidth availability [1]. However, high isotropic path loss and sensitivity to blockages pose challenges in the design of these systems [2]. To overcome the severe signal attenuation, mm-wave systems leverage narrow-beam communications, by using large antenna arrays at base stations (BSs) and user-ends (UEs). However, narrow beams are highly susceptible to mis-alignment due to mobility and blockage, hence they require utilization of beam-alignment schemes, which may cause huge overhead.

Therefore, the design of beam-alignment schemes with minimal overhead is of paramount importance, and has been a subject of intense research. One of the earliest yet most popular schemes is *exhaustive search* [3], which scans sequentially through all possible BS-UE beam pairs and selects the one with maximum signal power for data communications. To reduce the delay of exhaustive search, *iterative search* is proposed in [4], where scanning is first performed using wider beams, followed by refinement using narrow beams. In the aforementioned heuristic schemes, the optimal design is not considered. To address this challenge, in our previous papers [5]–[8], we considered the optimal design of interactive beam-alignment protocols that utilize 1-bit feedback from UEs.

In [5], [6], we design a throughput-optimal beam-alignment scheme for a single UE and two UEs, respectively, and we prove the optimality of a *bisection search*; in [7], we optimize the trade-off between data communication and beam-sweeping in a mobile scenario where the BS widens its beam to mitigate the uncertainty on the UE position; in [8], we incorporate the energy cost of beam-alignment, and prove the optimality of a *fractional search* method. In our aforementioned papers [5]–[8], the optimal design is carried out under the restrictive assumption of error-free single-bit feedback. However, this assumption may not hold in the presence of significant side-lobe gain and/or low signal-to-noise ratio (SNR).

The case of erroneous or noisy feedback is considered in recent work [9], [10], and our work [11]. A coded beam-alignment scheme is proposed in [11] to correct these errors, but with no consideration of feedback to improve beam-selection. A multi-armed bandit (MAB) formulation based on upper confidence bound (UCB) is proposed in [9], by selecting the beam based on the empirical SNR distribution. A hierarchical beam-alignment scheme based on posterior matching is proposed in [10]: therein, a *first-best* policy is formulated, which selects the most likely beam pair based on the posterior distribution on the AoA-AoD pair. However, as we will see numerically, both UCB and first-best policies are prone to errors due to under-exploration of the beam space.

In this paper, we propose a beam-alignment design with the goal to maximize the alignment probability and the average throughput during the data communication phase. We pose the problem as a Markov decision process (MDP), where the beam pair is chosen based upon the *belief* over the AoA-AoD pair, given the history of scanned beams and the received signal power. We identify a simplified sufficient statistic in the form of preference of the AoA-AoD beam pairs. We derive lower and upper bounds to the value function, based on which we propose a heuristic policy which selects the beam pair with the *second-best* preference. We show numerically that this policy strikes a favorable trade-off between exploration and exploitation: instead of greedily choosing the beam corresponding to the most likely AoA-AoD pair (*first-best* [10]), it chooses the second most likely one, leading to better exploration; at the same time, it avoids wasting precious resources to scan unlikely beam pairs, leading to better exploitation than other MAB techniques, such as linear Thompson sampling (LTS) [12] and UCB [9]. The proposed *second-best* scheme is shown to outperform *first-best* [10], LTS-based [12] and UCB-based [9] schemes by up to 7%, 10% and 30% in alignment probability, respectively.

The rest of the paper is organized as follows. In Sec. II, we present the system model. In Sec. III, we formulate the problem and our proposed solution strategy. In Sec. IV, we present numerical results, followed by final remarks in Sec. V.

This research has been funded by NSF under grant CNS-1642982.

Authors are with the School of Electrical and Computer Engineering, Purdue University, email: {hussai13,michelusi}@purdue.edu.

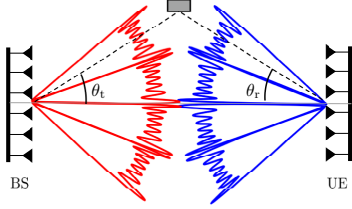


Fig. 1: System model; $M_t = M_r = 128$; beamforming algorithm in [13].

II. SYSTEM MODEL

We consider a downlink scenario with one BS and one UE, as depicted in Fig. 1. Time is divided into frames of duration $T_{fr} = T_s N$, each with N slots of duration T_s . The frame is partitioned into two phases: a *beam-alignment phase* of duration $L T_s$ ($L < N$ slots), followed by a *downlink data communication phase*, of duration $(N - L) T_s$. Each beam-alignment slot is further partitioned into a *pilot transmission phase*, of duration T_{pt} , followed by a *feedback phase*, of duration T_{fb} , with $T_s = T_{pt} + T_{fb}$. These are detailed next.

The BS and UE are equipped with uniform linear arrays (ULAs) with M_t and M_r antenna elements, respectively, and use analog beamforming. The signal received at the UE is

$$\mathbf{z}_k = \sqrt{P_{tx,k}} \mathbf{u}_k^H \mathbf{H}_k \mathbf{v}_k \mathbf{s} + \mathbf{w}_k, \quad \forall k \in \{0, 1, \dots, N - 1\}, \quad (1)$$

where $P_{tx,k}$ is the average transmit power of the BS; $\mathbf{s} \in \mathbb{C}^S$ is the transmitted signal with S symbols with $\mathbb{E}[\|\mathbf{s}\|_2^2] = S$; $\mathbf{H}_k \in \mathbb{C}^{M_r \times M_t}$ is the channel matrix; $\mathbf{v}_k \in \mathbb{C}^{M_t}$ is the BS beamforming vector with $\|\mathbf{v}_k\|_2^2 = 1$; $\mathbf{u}_k \in \mathbb{C}^{M_r}$ is the UE combining vector with $\|\mathbf{u}_k\|_2^2 = 1$; $\mathbf{w}_k \sim \mathcal{CN}(\mathbf{0}, N_0 W_{tot} \mathbf{I})$ is additive white Gaussian noise (AWGN), with one-sided power spectral density N_0 and system bandwidth W_{tot} .

Channel Model: We use the extended Saleh-Valenzuela geometric model with a single-cluster [14], as adopted in several previous works (e.g., see [8], [15], [16]). In fact, typical mm-wave channels have been shown to exhibit one dominant cluster containing most of the signal energy [17]. The single-cluster channel is modeled as

$$\mathbf{H}_k = \alpha_k \mathbf{a}_r(\theta_{r,k}) \mathbf{a}_t^H(\theta_{t,k}), \quad (2)$$

where $\theta_k \triangleq (\theta_{r,k}, \theta_{t,k}) \in [-\frac{\pi}{2}, \frac{\pi}{2}]^2$ is the angle of arrival (AoA) and angle of departure (AoD) pair associated to the dominant cluster, with complex fading gain α_k ; $\mathbf{a}_r$ and $\mathbf{a}_t$ are the UE and BS array response vectors, respectively, defined as

$$\mathbf{a}_x(\theta_x) = \frac{1}{\sqrt{M_x}} \left[1, e^{j \frac{2\pi d_x}{\lambda} \psi_x}, \dots, e^{j (M_x - 1) \frac{2\pi d_x}{\lambda} \psi_x} \right]^T, \quad x \in \{t, r\},$$

where $\psi_x = \sin \theta_x$, d_x is the antenna spacing, $\lambda = c/f_c$ is the wavelength at carrier frequency f_c , c denotes the speed of light. We assume that during the duration of one frame T_{fr} , θ_k remains unchanged, $\theta_k = \theta$, and α_k are i.i.d. Rayleigh fading in each slot with distribution $\alpha_k \sim \mathcal{CN}(0, \ell(d)^{-1})$, where $\ell(d)$ is the path loss at distance d from the BS. In fact, the AoA-AoD pair change much slower than the channel gain [18].

Codebook structure: In slot k , the BS uses the beamforming vector $\mathbf{v}_k \in \mathcal{V}$ and the UE uses the combining vector $\mathbf{u}_k \in \mathcal{U}$, from the codebooks $\mathcal{V}$ and $\mathcal{U}$, respectively. We assume a sectorized model [8], in which the AoA and AoD spaces are partitioned into sectors of equal beamwidth (as shown in Fig. 1 for the case of four sectors, this model approximates well analog beamforming). Accordingly, let $\mathcal{B}_r(\mathbf{u}) \subseteq [-\frac{\pi}{2}, \frac{\pi}{2}]$

and $\mathcal{B}_t(\mathbf{v}) \subseteq [-\frac{\pi}{2}, \frac{\pi}{2}]$ denote the AoA and AoD supports of the UE combiner and BS beamformer vectors $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$, respectively, with equal beamwidth $|\mathcal{B}_r(\mathbf{u})| = \frac{\pi}{|\mathcal{U}|}$, $\forall \mathbf{u} \in \mathcal{U}$ and $|\mathcal{B}_t(\mathbf{v})| = \frac{\pi}{|\mathcal{V}|}$, $\forall \mathbf{v} \in \mathcal{V}$, where $|\mathcal{B}|$ denotes the measure $|\mathcal{B}| \triangleq \int_{\mathcal{B}} dx$. We define $\mathcal{B}(\mathbf{u}, \mathbf{v}) \triangleq \mathcal{B}_r(\mathbf{u}) \times \mathcal{B}_t(\mathbf{v})$ as the joint AoA-AoD support of $(\mathbf{u}, \mathbf{v})$. We assume that the angular supports are mutually orthogonal and form a partition of the entire AoA-AoD space $[-\frac{\pi}{2}, \frac{\pi}{2}]^2$, i.e., $\mathcal{B}(\mathbf{u}, \mathbf{v}) \cap \mathcal{B}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) = \emptyset$, $\forall (\mathbf{u}, \mathbf{v}) \neq (\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ and $\cup_{\mathbf{u} \in \mathcal{U}} \mathcal{B}_r(\mathbf{u}) = \cup_{\mathbf{v} \in \mathcal{V}} \mathcal{B}_t(\mathbf{v}) = [-\frac{\pi}{2}, \frac{\pi}{2}]$. Let $(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})$, $i \in \mathcal{I} \triangleq \{1, 2, \dots, |\mathcal{U}||\mathcal{V}|\}$ be any ordering of combining and beamforming vectors, and $\mathcal{B}^{(i)} \triangleq \mathcal{B}(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})$ be their support. Let $A_k \in \mathcal{I}$ be the beam index of the combining and beamforming vectors scanned in slot k , so that $(\mathbf{u}_k, \mathbf{v}_k) = (\mathbf{u}^{(A_k)}, \mathbf{v}^{(A_k)})$. Let X be a discrete random variable denoting the index of the support that the AoA-AoD pair θ of the channel belongs to, so that $\theta \in \mathcal{B}^{(X)}$. Then, from (1)-(2), the received signal can be expressed as¹

$$\mathbf{z}_k \approx \sqrt{P_{tx,k}} \alpha_k \left[(\sqrt{G} - \sqrt{g}) \delta[A_k, X] + \sqrt{g} \right] \mathbf{s} + \mathbf{w}_k, \quad (3)$$

where $\delta[\cdot]$ is the Kronecker's delta function, equal to 1 if alignment is achieved ($A_k = X$), equal to 0 otherwise ($A_k \neq X$); G and g are, respectively, the main and side lobe gains of the sectorized model, expressed as

$$G = \min_{(\theta_r, \theta_t) \in \mathcal{B}(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})} |\mathbf{a}_r(\theta_r)^H \mathbf{u}^{(i)}|^2 |\mathbf{a}_t^H(\theta_t) \mathbf{v}^{(i)}|^2, \quad \forall i,$$

$$g = \max_{(\theta_r, \theta_t) \notin \mathcal{B}(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})} |\mathbf{a}_r(\theta_r)^H \mathbf{u}^{(i)}|^2 |\mathbf{a}_t^H(\theta_t) \mathbf{v}^{(i)}|^2, \quad \forall i.$$

In the following, we describe the beam-alignment and data communication procedures.

Beam-Alignment: In each slot k of the beam-alignment phase, the BS transmits a pilot sequence $\mathbf{s}$ using the beam index A_k , with transmit power $P_{tx,k} = P_{ba}$. Upon receiving $\mathbf{z}_k$ (based on the combining vector with index A_k), the UE uses a matched filter to compute the signal strength and sends the *normalized received power* feedback signal Y_k back to the BS, of the form

$$Y_k = \frac{|\mathbf{s}^H \mathbf{z}_k|^2}{\|\mathbf{s}\|^2 N_0 W_{tot} (1 + \Lambda g)}, \quad (4)$$

where $\Lambda \triangleq \frac{P_{ba} \|\mathbf{s}\|^2}{N_0 W_{tot} \ell(d)}$ is the pre-beamforming receive SNR during beam-alignment. Then, the probability density function (pdf) of Y_k conditional on $(X, A_k) = (x, a_s)$ is given by

$$f(Y_k = y | X = x; A_k = a_s) = [\nu e^{-\nu y}]^{\delta[a_s, x]} [e^{-y}]^{1 - \delta[a_s, x]}, \quad (5)$$

where $1/\nu$ is the mean signal power in case of alignment, with

$$\nu \triangleq \frac{1 + g\Lambda}{1 + G\Lambda}. \quad (6)$$

The BS uses a Bayesian approach to select A_k : starting from $\mathcal{H}_0 \triangleq \emptyset$ and given the history of feedback and scanned beam indices $\mathcal{H}_k \triangleq \{(A_j, Y_j)\}_{j=0}^{k-1}$, the next beam index A_k is selected. This procedure continues until the end of the beam-alignment phase.

Data communication: Upon completion of the beam-alignment phase, given the history of feedback and actions $\mathcal{H}_L$, the BS selects the data communication parameters: beam index for data communication $A_d \in \mathcal{I}$, transmission power

¹The phase of $\mathbf{u}_k^H \mathbf{a}_r(\theta_r) \mathbf{a}_t^H(\theta_t) \mathbf{v}_k$ is incorporated into α_k .

$P_d \in [0, P_{\max}]$, and data rate $R_d \geq 0$. These parameters are used until the end of the data communication phase.

Let $b_0[x]$ be the *prior belief* over $X=x$ (or equivalently over $\theta \in \mathcal{B}^{(x)}$) available at the beginning of the beam-alignment phase. We define the expected rate during the communication phase (normalized by the frame duration), as

$$\begin{aligned} \bar{R}(A_d, P_d, R_d | b_0, \mathcal{H}_L) \\ \triangleq \frac{T_{\text{fr}} - LT_s}{T_{\text{fr}}} \mathbb{P}(X = A_d | b_0, \mathcal{H}_L) \hat{R}(R_d, P_d), \end{aligned} \quad (7)$$

where we have defined

$$\hat{R}(R_d, P_d) \triangleq R_d \mathbb{P} \left[R_d \leq W_{\text{tot}} \log_2 \left(1 + \frac{|\alpha_k|^2 P_d G}{N_0 W_{\text{tot}}} \right) \right]. \quad (8)$$

The probability term in (7) is the probability of achieving correct alignment, given the prior b_0 and the history $\mathcal{H}_L$ of feedback and actions during the beam-alignment phase, whereas the probability term in (8) denotes the probability of non-outage with respect to the realization of the fading process (i.i.d. over time), given that correct alignment has been achieved (we assume that mis-alignment yields outage with probability one, since $g \ll G$).

III. PROBLEM FORMULATION AND SOLUTION

We now formulate the beam-alignment and data communication problem in the context of a decision process. We define a policy μ , part of our design, which operates as follows. At time k during beam-alignment, given the history of feedback and actions $\mathcal{H}_k$, the BS selects the *beam-alignment action* $A_k = a_s \in \mathcal{I}$ with probability $\mu_k(a_s | \mathcal{H}_k)$; given $\mathcal{H}_L$, the BS selects the data communication parameters as $(A_d, P_d, R_d) = \mu_d(\mathcal{H}_L)$. The goal is to design μ so as to maximize the expected communication rate, i.e.,

$$\mathbf{P0}: \max_{\mu} \mathbb{E}_{\mu} [\bar{R}(A_d, P_d, R_d | b_0, \mathcal{H}_L) | b_0],$$

where the expectation $\mathbb{E}_{\mu}$ is conditional on the prior belief b_0 and on the policy μ being executed during beam-alignment and data communication. Note that, using (7), we can rewrite the optimization problem as

$$\begin{aligned} \mathbf{P1}: \max_{\mu} \mathbb{E}_{\mu} \left[\mathbb{P}(X = A_d | b_0, \mathcal{H}_L) | b_0 \right] \\ \times \frac{T_{\text{fr}} - LT_s}{T_{\text{fr}}} \max_{R_d \geq 0, 0 \leq P_d \leq P_{\max}} \hat{R}(R_d, P_d), \end{aligned}$$

i.e., the problem can be decomposed into the following two independent problems: 1) find the optimal rate and power (R_d^*, P_d^*) that maximize the expected rate in the communication phase, conditional on correct alignment being achieved ($X = A_d$); 2) find the optimal beam-alignment policy and the beam index for communication A_d so as to maximize the probability of correct alignment. The first problem can be solved efficiently by maximizing (8). In the sequel, we consider the latter problem.

Let $b_k[x] \triangleq \mathbb{P}(X = x | \mathcal{H}_k, b_0)$ be the belief over $X=x$ given the history of actions and feedback and prior belief b_0 . It serves as a sufficient statistic for optimal control for problem **P1**. In the following lemma, we present an equivalent simplified sufficient statistic along with its dynamics.

Lemma 1. Let $m_0[x] \triangleq \ln b_0[x]$ denote the prior preference of $X = x$. Given the action and feedback pair (A_k, Y_k) , the belief at $k+1$ is updated as

$$b_{k+1}[x] = \frac{\exp\{m_{k+1}[x]\}}{\sum_{l \in \mathcal{I}} \exp\{m_{k+1}[l]\}}, \quad (9)$$

where

$$m_{k+1}[x] = m_k[x] + J(Y_k) \delta[A_k, x], \quad \forall x \in \mathcal{I}, \quad (10)$$

and we have defined

$$J(y) \triangleq (1 - \nu)y + \ln \nu. \quad (11)$$

Proof. Given the belief b_k and $(A_k, Y_k) = (a_s, y)$, we have

$$\begin{aligned} b_{k+1}[x] &\stackrel{(a)}{=} \mathbb{P}(X = x | \mathcal{H}_{k+1}) \\ &\stackrel{(b)}{\propto} f(Y_k = y | X = x, A_k = a_s, \mathcal{H}_k) \mathbb{P}(X = x | A_k = a_s, \mathcal{H}_k) \\ &\stackrel{(c)}{=} f(Y_k = y | X = x, A_k = a_s) b_k[x] \\ &\stackrel{(d)}{=} [\nu \exp\{-\nu y\}]^{\delta[a_s, x]} [\exp\{-y\}]^{1 - \delta[a_s, x]} b_k[x] \\ &\stackrel{(e)}{=} \exp\{-y + J(y) \delta[a_s, x]\} b_k[x], \end{aligned} \quad (12)$$

where (a) follows from the definition of belief; (b) follows from Bayes' rule and $\propto$ denotes proportionality up to a normalization factor independent of x ; (c) follows from the facts that Y_k is independent of history $\mathcal{H}_k$ given (X, A_k) , and X is independent of action A_k given $\mathcal{H}_k$, and by the definition of belief b_k ; (d-e) follow by substitution of the pdf of Y_k given in (5) and by definition of $J(y)$. We prove the lemma using induction. The lemma holds for b_0 by definition of m_0 . Let $0 \leq k \leq L-1$ and b_k be given by (9), then using (12)(e) normalized to sum to one, we get

$$\begin{aligned} b_{k+1}[x] &= \frac{\exp\{J(y) \delta[a_s, x]\} \exp\{m_k[x]\}}{\sum_{l \in \mathcal{I}} \exp\{J(y) \delta[a_s, l]\} \exp\{m_k[l]\}} \\ &= \frac{\exp\{m_{k+1}[x]\}}{\sum_{l \in \mathcal{I}} \exp\{m_{k+1}[l]\}}, \end{aligned} \quad (13)$$

where $m_{k+1}[x]$ is given by (10). ■

Let $\mathbf{m}_k \triangleq [m_k[1], \dots, m_k[L]]$. Then, the previous lemma demonstrates that $\mathbf{m}_k$ is a sufficient statistic for control decisions, since it is sufficient for computing the belief b_k at time k . Therefore, μ can be expressed as $A_k = \mu_k(\mathbf{m}_k)$, $\forall 0 \leq k \leq L$, which maps the current preference vector $\mathbf{m}_k$ to beam index $A_k \in \mathcal{I}$. This result makes it possible to achieve an efficient implementation, since the belief can be updated according to simple preference update rules as in (10), rather than via complex Bayesian belief updates. In the subsequent analysis, we will use $\mathbf{m}_k$ rather than b_k as the state.

A. MDP Formulation

Thanks to the identification of the sufficient statistic $\mathbf{m}_k$, we model the optimization problem **P1** as a Markov decision process (MDP) and optimize the decision variables to maximize the alignment probability in the data-communication phase. The MDP is a 5-tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{I}, f(\mathbf{m}_{k+1} | \mathbf{m}_k, a_k), r_k(\mathbf{m}_k, a_k), \forall k \in \mathcal{T} \rangle$, with elements described as follows.

Time Horizon: given as $\mathcal{T} = \{0, 1, \dots, L\}$ where $\mathcal{T}_{\text{BA}} \triangleq \mathcal{T} \setminus \{L\}$ denote the slot indices associated with the

beam-alignment phase, whereas at $k=L$, the communication parameters are selected and used until the end of the frame. **State space:** given as $\mathcal{S} = \mathbb{R}^{|\mathcal{I}|}$, i.e., all possible values of preference vectors $\mathbf{m}_k$.

Action space: the set containing all the beam indices, $\mathcal{I}$.

State transition distribution: Given state $\mathbf{m}_k = \mathbf{m}$ and action $A_k = a_s$ used in the k th stage of the beam-alignment phase, the feedback $Y_k = y$ is generated with pdf

$$f(y|\mathbf{m}, a_s) \triangleq \sum_{x \in \mathcal{I}} f(Y_k = y|X = x, A_k = a_s) b_k[x] \quad (14)$$

$$= \frac{\exp\{m[a_s]\}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}} \nu e^{-\nu y} + \left[1 - \frac{\exp\{m[a_s]\}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}}\right] e^{-y},$$

leading to the new state

$$\mathbf{m}_{k+1} = \mathbf{m} + J(y)\delta[a_s], \quad (15)$$

where $\delta[a_s] = [\delta[a_s, x]]_{\forall x \in \mathcal{I}}$ is the vector with entries $\delta[a_s, x]$.

Reward function: the reward is the probability of choosing a beam index such that $A_d = X$ in the data communication phase, so that correct alignment is achieved, yielding

$$r_k(\mathbf{m}, a) = \begin{cases} 0, & k \in \mathcal{T}_{\text{BA}}, \\ \frac{\exp\{m[a]\}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}}, & k = L. \end{cases} \quad (16)$$

We now formulate the value function iteration for the MDP.

B. Value Function

The value function under the optimal policy is given as

$$V_k^*(\mathbf{m}) = \max_{a_s \in \mathcal{I}} q_k(\mathbf{m}, a_s), \quad (17)$$

where q_k is the Q-function under the state-action pair $(\mathbf{m}, a)$, defined recursively as

$$q_L(\mathbf{m}, A_d) = r_L(\mathbf{m}, A_d) = \frac{\exp\{m[A_d]\}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}},$$

and for $k \in \mathcal{T}_{\text{BA}}$, using (14),

$$q_k(\mathbf{m}, a_s) = \int_{\mathbb{R}^{|\mathcal{I}|}} V_{k+1}^*(\mathbf{m}') f(\mathbf{m}_{k+1} = \mathbf{m}' | \mathbf{m}_k = \mathbf{m}, A_k = a_s) d\mathbf{m}'$$

$$= \int_0^\infty V_{k+1}^*(\mathbf{m} + J(y)\delta[a_s]) f(y|\mathbf{m}, a_s) dy. \quad (18)$$

This yields the optimal value function in the data communication phase, by choosing the beam index with maximum preference $A_d^* = \arg \max_{A_d \in \mathcal{I}} m[A_d]$,

$$V_L^*(\mathbf{m}) = \max_{A_d \in \mathcal{I}} q_L(\mathbf{m}, A_d) = \frac{\exp\{m[A_d^*]\}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}}. \quad (19)$$

In the beam-alignment phase ($k \in \mathcal{T}_{\text{BA}}$), combining (17) and (18), we obtain iteratively the value function as

$$V_k^*(\mathbf{m}) = \max_{a_s \in \mathcal{I}} \int_0^\infty V_{k+1}^*(\mathbf{m} + J(y)\delta[a_s]) f(y|\mathbf{m}, a_s) dy.$$

In the following theorem, whose proof is provided in the Appendix, we unveil structural properties of $V_k^*(\mathbf{m})$. We find a lower-bound and an upper-bound to the Q-function and show that these bounds are optimized by a policy which, in each stage of the beam-alignment phase, selects the beam index with the *second-best* preference. This result will be the basis for our proposed policy evaluated numerically in Sec. IV.

Theorem 1. For $k \in \mathcal{T}_{\text{BA}}$, the Q-function is bounded as

$$q_k(\mathbf{m}, a_s) \geq q_k^{LB}(\mathbf{m}, a_s) \triangleq \frac{1}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}} \left[\xi(a_s; \mathbf{m}) \right. \\ \left. + \exp \left\{ \frac{\min_{x_i \neq x_j} m[x_i] - \nu m[x_j]}{1 - \nu} \right\} h(\nu) \frac{g(\nu) - [g(\nu)]^{L-k}}{1 - g(\nu)} \right], \quad (20)$$

$$q_k(\mathbf{m}, a_s) \leq q_k^{UB}(\mathbf{m}, a_s) \triangleq \frac{[1 + h(\nu)]^{L-k-1}}{\sum_{l \in \mathcal{I}} \exp\{m[l]\}} \xi(a_s; \mathbf{m}), \quad (21)$$

where we have defined $\xi(a_s; \mathbf{m})$

$$\triangleq \begin{cases} \exp\{m[a_s]\}, & \text{if } \max_{\hat{a} \neq a_s} m[\hat{a}] - m[a_s] < \ln \nu, \\ \exp\{\max_{\hat{a} \neq a_s} m[\hat{a}]\} \\ + h(\nu) \exp \left\{ \frac{m[a_s] - \nu \max_{\hat{a} \neq a_s} m[\hat{a}]}{1 - \nu} \right\}, & \text{otherwise,} \end{cases} \quad (22)$$

where

$$h(\nu) \triangleq \exp \left\{ \frac{\nu}{1 - \nu} \ln \nu \right\} - \exp \left\{ \frac{\ln \nu}{1 - \nu} \right\} > 0, \quad (23)$$

$$g(\nu) \triangleq \exp \left\{ \frac{\ln \nu}{1 - \nu} \right\} \left[\frac{1}{\nu + 1} - \frac{\ln \nu}{1 - \nu} \right] > 0. \quad (24)$$

Let $x_{[1]}, x_{[2]}, \dots, x_{[|\mathcal{I}|]}$ be an ordering of beam indices in decreasing order of preference, i.e., $m[x_{[1]}] \geq m[x_{[2]}] \geq \dots, m[x_{[|\mathcal{I}|]}]$, then the optimal value function is bounded as

$$V_k^*(\mathbf{m}) \geq \max_{a_s \in \mathcal{I}} q_k^{LB}(\mathbf{m}, a_s) = q_k^{LB}(\mathbf{m}, x_{[2]}), \quad \forall k \in \mathcal{T}_{\text{BA}}, \quad (25)$$

$$V_k^*(\mathbf{m}) \leq \max_{a_s \in \mathcal{I}} q_k^{UB}(\mathbf{m}, a_s) = q_k^{UB}(\mathbf{m}, x_{[2]}), \quad \forall k \in \mathcal{T}_{\text{BA}}, \quad (26)$$

with the maximizer of q_k^{UB} and q_k^{LB} given by the second-best beam index $x_{[2]}$.

Proof. The proof is provided in the Appendix. ■

As a result of this Theorem, both the upper and lower bounds of the Q-function are maximized by the *second-best* beam index policy, which selects the beam index with the second-best preference during the beam-alignment phase. This policy will be evaluated numerically in the next section, against other MAB-based schemes proposed in the literature.

IV. NUMERICAL RESULTS

In this section, we evaluate the performance of the *second-best* beam index selection scheme ($a_s = x_{[2]}$) with analog beamforming, and compare it with three other schemes. The first one is based on LTS, a popular MAB scheme [12]. In LTS, at each slot the action is chosen according to the belief distribution, i.e., $a_s \sim b[x]$. The second scheme is based on scanning the most-likely beam index ($a_s = x_{[1]}$) as proposed in [10] (first-best). The third scheme is based on UCB as proposed in [9]. We evaluate the performance of these three schemes in terms of the probability of alignment and spectral efficiency using Monte-Carlo simulation with 10^5 iterations for each simulated point, with parameters as follows: $M_t = 128$, $M_r = 1$, $N_0 = -174$ dBm/Hz, $W_{\text{tot}} = 200$ MHz, $T_{\text{fr}} = 20$ ms, $T_s = 0.1$ ms, $f_c = 30$ GHz, $d = 10$ m, [path loss exponent] = 2. The BS uses $M_t = 128$ antennas and partitions the AoD space into 32 sectors, each with a beamwidth of $\pi/32$ rad and with uniform prior $b_0[x] = 1/32$, $\forall x \in \mathcal{I}$; the UE is isotropic, hence it uses $M_r = 1$ antenna with a single sector. We use the beamforming design proposed in [13] for ULAs with antenna spacing

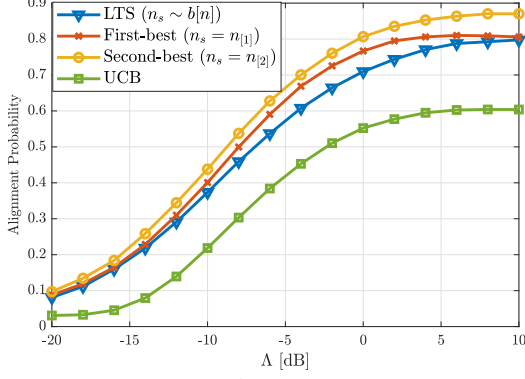


Fig. 2: Alignment Probability vs Λ ; $L = 32$ (beam-alignment takes 16% of frame duration).

$d_t = \lambda/2$. With this configuration, the main-lobe and side-lobe gains are best approximated by $G \approx 14\text{dB}$, $g \approx -11\text{dB}$.

In Fig. 2, we depict the probability of alignment achieved by the aforementioned schemes versus the pre-beamforming SNR Λ . It can be observed that second-best has better performance than the other three schemes, with up to 7%, 10%, and 30% performance gains compared to first-best, LTS-based and UCB-based schemes. The performance gain of second-best is attributed to a better exploration-exploitation trade-off. The first-best scheme suffers from poor exploration since it "greedily" chooses the beam index most likely to succeed, but fails to test other beams that may be under-explored, and is thus prone to make alignment errors. On the other hand, LTS-based scheme suffers from poor exploitation since it may scan least likely beams. The proposed *second-best* scheme, on the other hand, strikes a favorable trade-off between exploration and exploitation: instead of greedily choosing the most likely beam, it chooses the second most likely one, leading to better exploration than *first-best*; simultaneously, by not choosing beam pairs that are unlikely to succeed, it leads to a better exploitation compared to the LTS-based and UCB-based schemes. Finally, compared to UCB, second-best is better tailored to the structure of the model, since it aims to maximize the alignment probability *at the end* of the beam-alignment phase (see (16)), rather than the surrogate metric of UCB – the cumulative SNR accrued *during* beam-alignment.

In Fig. 3, we depict the spectral efficiency against the fraction of T_{fr} used for BA LT_s/T_{fr} . We fix the SNR for beam-alignment as $\Lambda = 0\text{dB}$ and the data-communication power as $P_d^* = 22\text{dBm}$. Similar to Fig. 2, *second-best* outperforms the three other schemes, owing to improved alignment. The spectral efficiency is maximized at a unique maximizer L^* : it increases initially with $L \leq L^*$ as the beam-alignment probability improves with L . However, as L increases beyond L^* , this gain is offset by the increased overhead and reduced duration of the data communication phase.

V. CONCLUSIONS

In this paper, we have formulated the beam-alignment problem as a Bayesian MAB problem. For the optimal control design, we have identified a simplified sufficient statistic referred to as the preference of beam pairs. Based on the preference and bounding of the value function, we have proposed a heuristic policy, which selects the beam pair with the second best-preference to scan. We have shown numerically that the

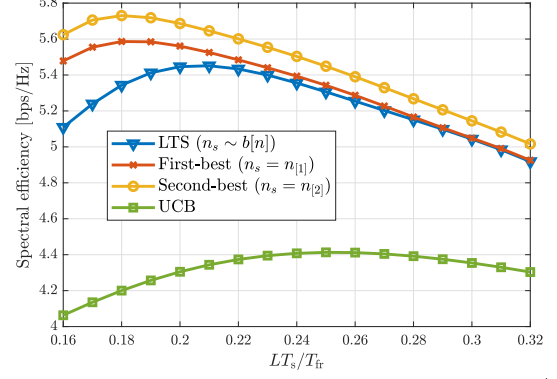


Fig. 3: Spectral efficiency vs fraction of T_{fr} used for BA LT_s/T_{fr} .

proposed scheme outperforms the first-best, LTS, and UCB based beam-alignment schemes proposed in the literature.

APPENDIX: PROOF OF THEOREM 1

Proof. We prove the theorem using induction. Notice that from the definition of Q-function (18) and the optimal value function expression (19) for $k = L$, we get

$$q_{L-1}(\mathbf{m}, a_s) = \int_0^\infty \frac{e^{\max_{\hat{a}} m'[\hat{a}|\mathbf{m}, a_s, y]}}{\sum_{l \in \mathcal{I}} e^{m'[l|\mathbf{m}, a_s, y]}} f(y|\mathbf{m}, a_s) dy,$$

where we have defined the preference update (15) as

$$m'[x|\mathbf{m}, a_s, y] = m[x] + J(y)\delta[a_s, x].$$

Moreover, using (14) and (11) we note that

$$\sum_{l \in \mathcal{I}} e^{m'[l|\mathbf{m}, a_s, y]} = \sum_{l \in \mathcal{I}} e^{m[l] + J(y)\delta[a_s, l]} = e^y f(y|\mathbf{m}, a_s) \sum_{l \in \mathcal{I}} e^{m[l]}. \quad (27)$$

This yields

$$\begin{aligned} q_{L-1}(\mathbf{m}, a_s) &\stackrel{(a)}{=} \frac{1}{\sum_{l \in \mathcal{I}} e^{m[l]}} \int_0^\infty e^{\max_{\hat{a}} m[\hat{a}] + J(y)\delta[a_s, \hat{a}]} e^{-y} dy \\ &\stackrel{(b)}{=} \frac{1}{\sum_{l \in \mathcal{I}} e^{m[l]}} \xi(a_s; \mathbf{m}) \end{aligned} \quad (28)$$

where (b) follows by evaluating the integral in (a) for the two cases in (22), and noting that it is given by $\xi(a_s; \mathbf{m})$. Using Lemma 2 and (28)(b), the optimal value function becomes

$$V_{L-1}^*(\mathbf{m}) = \frac{1}{\sum_{l \in \mathcal{I}} e^{m[l]}} \left[e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[x_{[2]}] - \nu m[x_{[1]}]}{1-\nu}} \right].$$

Thus, the theorem statement holds for $k=L-1$ with equality. Assume it holds for $k+1$. Using Lemma 2, we can bound

$$\begin{aligned} \max_{\hat{a}} \xi(\hat{a}; \mathbf{m}'[n|\mathbf{m}, a_s, y]) &\geq \exp\{\max_{\hat{a}} m'[\hat{a}|\mathbf{m}, a_s, y]\} \\ &\quad + h(\nu) e^{\frac{\min_{x_i \neq x_j} m'[x_i|\mathbf{m}, a_s, y] - \nu m'[x_j|\mathbf{m}, a_s, y]}{1-\nu}}. \end{aligned}$$

Using (18), the induction hypothesis (25) for $k+1$ and the above bound, we obtain $q_k(\mathbf{m}, a_s)$

$$\begin{aligned} &\geq \int_0^\infty \left\{ \frac{e^{\max_{\hat{a}} m'[\hat{a}|\mathbf{m}, a_s, y]}}{\sum_{l \in \mathcal{I}} e^{m'[l|\mathbf{m}, a_s, y]}} + \frac{e^{\frac{\min_{x_i \neq x_j} m'[x_i|\mathbf{m}, a_s, y] - \nu m'[x_j|\mathbf{m}, a_s, y]}{1-\nu}}}{\sum_{l \in \mathcal{I}} e^{m'[l|\mathbf{m}, a_s, y]}} \right. \\ &\quad \times h(\nu) \frac{1 - [g(\nu)]^{L-k-1}}{1 - g(\nu)} \Big\} f(y|\mathbf{m}, a_s) dy. \end{aligned} \quad (29)$$

Moreover, we note that

$$\begin{aligned} & \min_{x_i \neq x_j} m'[x_i | \mathbf{m}, a_s, y] - \nu m'[x_j | \mathbf{m}, a_s, y] \\ & \geq \min_{x_i \neq x_j} [m[x_i] - \nu m[x_j]] + \min_{x_i \neq x_j} J(y) \{ \delta[a_s, x_i] - \nu \delta[a_s, x_j] \} \\ & = \min_{x_i \neq x_j} [m[x_i] - \nu m[x_j]] + \min \{ J(y), -\nu J(y) \}. \end{aligned} \quad (30)$$

By substituting (30) and (27) into (29), yields

$$\begin{aligned} q_k(\mathbf{m}, a_s) & \geq \frac{1}{\sum_{l \in \mathcal{I}} e^{m[l]}} \left[\int_0^\infty e^{\max_{\hat{a}} m[\hat{a}] + J(y) \delta[a_s, \hat{a}]} e^{-y} dy \right. \\ & \quad \left. + e^{\frac{\min_{x_i \neq x_j} m[x_i] - \nu m[x_j]}{1-\nu}} \int_0^\infty e^{\frac{\min \{ J(y), -\nu J(y) \}}{1-\nu}} e^{-y} dy \right. \\ & \quad \left. \times h(\nu) \frac{1 - [g(\nu)]^{L-k-1}}{1 - g(\nu)} \right]. \end{aligned} \quad (31)$$

The first integral in (31) is equal to $\xi(a_s; \mathbf{m})$ and the second integral is found to be equal to

$$\int_0^\infty e^{\frac{\min \{ J(y), -\nu J(y) \}}{1-\nu}} e^{-y} dy = e^{\frac{\ln \nu}{1-\nu}} \left[\frac{1}{\nu+1} - \frac{\ln \nu}{1-\nu} \right] = g(\nu) > 0.$$

Upon substituting these integrals into (31) yields the following lower-bound to the Q-function,

$$q_k(\mathbf{m}, a_s) \geq \frac{\xi(a_s; \mathbf{m}) + e^{\frac{\min_{x_i \neq x_j} m[x_i] - \nu m[x_j]}{1-\nu}} h(\nu) \frac{g(\nu) - [g(\nu)]^{L-k}}{1 - g(\nu)}}{\sum_{l \in \mathcal{I}} e^{m[l]}},$$

which proves the induction step (20), and whose maximization (see Lemma 2) yields (25).

Similarly, using the induction hypothesis (26) for $k+1$ and the upper-bound

$$\max_{\hat{a}} \xi(\hat{a}; \mathbf{m}'[\mathbf{m}, a_s, y]) \leq (1 + h(\nu) \exp\{\max_{\hat{a}} m'[\hat{a} | \mathbf{m}, a_s, y]\},$$

we obtain the following upper-bound to the Q-function,

$$q_k(\mathbf{m}, a_s) \leq \frac{[1 + h(\nu)]^{L-k-1}}{\sum_{l \in \mathcal{I}} e^{m[l]}} \int_0^\infty e^{\max_{\hat{a}} m[\hat{a}] + J(y) \delta[a_s, \hat{a}]} e^{-y} dy.$$

The integral above is equal to $\xi(a_s; \mathbf{m})$, which proves the induction step (21), hence

$$V_k^*(\mathbf{m}) = \max_{a_s \in \mathcal{I}} q_k(\mathbf{m}, a_s) \leq \frac{[1 + h(\nu)]^{L-k-1}}{\sum_{l \in \mathcal{I}} e^{m[l]}} \max_{a_s \in \mathcal{I}} \xi(a_s; \mathbf{m}). \quad (32)$$

Noting that $\max_{a_s} \xi(a_s; \mathbf{m}) = \xi(x_{[2]}; \mathbf{m})$ (see Lemma 2), and upon substitution in (32) yields (26). ■

Lemma 2. We have that $\arg \max_{a_s \in \mathcal{I}} \xi(a_s; \mathbf{m}) = x_{[2]}$ and

$$\max_{a_s \in \mathcal{I}} \xi(a_s; \mathbf{m}) = e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[x_{[2]}] - \nu m[x_{[1]}]}{1-\nu}}. \quad (33)$$

Proof. To show that $\arg \max_{a_s \in \mathcal{I}} \xi(a_s; \mathbf{m}) = x_{[2]}$, we proceed as follows. Clearly, if $a_s \in \{x_{[2]}, x_{[3]}, \dots, x_{[|\mathcal{I}|]}\}$, then $\max_{\hat{a} \neq a_s} m[\hat{a}] - m[a_s] = m[x_{[1]}] - m[a_s] \geq 0 > \ln(\nu)$, hence

$$\xi(a_s; \mathbf{m}) = e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[a_s] - \nu m[x_{[1]}]}{1-\nu}},$$

maximized at $a_s = x_{[2]}$. Therefore, we restrict $a_s \in \{x_{[1]}, x_{[2]}\}$ without loss in performance. Next, we show that $\xi(x_{[2]}; \mathbf{m}) \geq \xi(x_{[1]}; \mathbf{m})$. Let $\Delta \triangleq m[x_{[1]}] - m[x_{[2]}]$. If $\Delta > -\ln \nu$, then

$\xi(x_{[1]}; \mathbf{m}) = e^{m[x_{[1]}]}$ and $\xi(x_{[2]}; \mathbf{m}) > \xi(x_{[1]}; \mathbf{m})$. Otherwise,

$$\xi(x_{[2]}; \mathbf{m}) - \xi(x_{[1]}; \mathbf{m}) \propto \frac{e^{\Delta-1}}{e^{\frac{\Delta}{1-\nu}} - e^{-\nu \frac{\Delta}{1-\nu}}} - h(\nu) \triangleq u(\Delta, \nu).$$

Note that $u(\Delta, \nu)$ is decreasing in $\Delta \in (0, -\ln \nu]$, $\forall \nu \in (0, 1)$, minimized at $\Delta = -\ln \nu$, yielding, after algebraic steps,

$$\xi(x_{[2]}; \mathbf{m}) - \xi(x_{[1]}; \mathbf{m}) \propto u(\Delta, \nu) \geq \frac{h(\nu)}{e^{-\frac{1+\nu}{1-\nu} \ln \nu} - 1} > 0.$$

In both cases, $\max_{a_s} \xi(a_s) = \xi(x_{[2]})$. Upon substitution of $a_s = x_{[2]}$ in (22), yields (33). ■

REFERENCES

- [1] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.
- [2] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Prentice Hall, 2015.
- [3] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: issues and approaches," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 180–185, January 2015.
- [4] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmWave communications," in *48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014.
- [5] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *2017 Information Theory and Applications Workshop (ITA)*, Feb 2017, pp. 1–6.
- [6] R. A. Hassan and N. Michelusi, "Multi-user beam-alignment for millimeter-wave networks," in *2018 Information Theory and Applications Workshop (ITA)*, Feb 2018, pp. 1–6.
- [7] N. Michelusi and M. Hussain, "Optimal beam-sweeping and communication in mobile millimeter-wave networks," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.
- [8] M. Hussain and N. Michelusi, "Energy-efficient interactive beam alignment for millimeter-wave networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 838–851, Feb 2019.
- [9] M. Hashemi, A. Sabharwal, C. Emre Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE INFOCOM 2018*, April 2018, pp. 2393–2401.
- [10] S. Chiu, N. Ronquillo, and T. Javidi, "Active learning and CSI acquisition for mmwave initial alignment," *IEEE Journal on Selected Areas in Communications*, 2019, to appear. [Online]. Available: <http://arxiv.org/abs/1812.07722>
- [11] M. Hussain and N. Michelusi, "Coded Energy-Efficient Beam-Alignment for Millimeter-Wave Networks," in *56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2018, pp. 407–412.
- [12] A. G. Sutton, Richard S. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.
- [13] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-Resolution Codebook and Adaptive Beamforming Sequence Design for Millimeter Wave Beam Alignment," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5689–5701, Sep. 2017.
- [14] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, February 1987.
- [15] 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.
- [16] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and C. J. Zhang, "Design and analysis of initial access in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6409–6425, Oct 2017.
- [17] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.
- [18] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, June 2017.