Robust Beam Tracking and Data Communication in Millimeter Wave Mobile Networks

Shahram Shahsavari[†], Mohammad A. (Amir) Khojastepour[◇], Elza Erkip[†]
[†]NYU Tandon School of Engineering, [◇]NEC Laboratories America, Inc.
Emails: [†]{shahram.shahsavari,elza}@nyu.edu, [◇] amir@nec-labs.com

***Abstract*— Millimeter-wave (mmWave) bands have shown the potential to enable high data rates for next generation mobile networks. In order to cope with high path loss and severe shadowing in mmWave frequencies, it is essential to employ massive antenna arrays and generate narrow transmission patterns (beams). When narrow beams are used, mobile user tracking is indispensable for reliable communication. In this paper, a joint beam tracking and data communication strategy is proposed in which, the base station (BS) increases the beamwidth during data transmission to compensate for location uncertainty caused by user mobility. In order to evade low beamforming gains due to widening the beam pattern, a probing scheme is proposed in which the BS transmits a number of probing packets to refine the estimation of angle of arrival based on the user feedback, which enables reliable data transmission through narrow beams again. In the proposed scheme, time is divided into similar frames each consisting of a probing phase followed by a data communication phase. A steady state analysis is provided based on which, the duration of data transmission and probing phases are optimized. Furthermore, the results are generalized to consider practical constraints such as minimum feasible beamwidth. Simulation results reveal that the proposed method outperforms well-known approaches such as optimized beam sweeping.**

I. Introduction

Millimeter wave spectrum (30 GHz-300 GHz) offers an order of magnitude greater bandwidth for wireless communications which can be utilized to provide multi-Gbps data rates and meet the growing demand for speed in wireless networks [1]. Although high path loss and severe shadowing attenuate signal power intensely in mmWave frequencies, various beamforming (BF) techniques have been proposed to overcome these effects by forming directional radiation patterns using massive antenna arrays [2].

Signal space BF (also referred to as digital BF) requires high quality channel estimation and sophisticated hardware [2]. However, most of the practical systems support a limited number of RF chains and have limited capability to adapt the BF coefficients. In this paper, physical BF (also referred to as analog BF) is adopted which is one of the most popular choices for mmWave systems. The properties of the propagation channel in mmWave systems such as having a limited number of spatial clusters [3] make analog BF an appealing choice.

Misalignment between transmitter and receiver beam patterns can diminish the BF gain required for a high data rate mmWave links [4]. A variety of beam alignment (BA) techniques have been proposed for static scenarios where the user is stationary [5]. *Exhaustive search* (ES) algorithms scan different beams sequentially, and pick the beam with the highest received power, which leads to a large overhead [6], [7]. *Hierarchical search* (HS) (or fractional search) algorithms are proposed to reduce BA overhead and delay [8], [9], [10], and envisioned for standards such as IEEE 802.11ad [11]. These algorithms search wider sectors first using coarser beams, and then refine the search within the best sector.

Device mobility (rotational and/or linear) breaks the alignment, demanding constant retraining which can increase BA overhead significantly [4]. To avoid this overhead, reference [4] suggests to track the user in lower frequencies (i.e. sub 6 GHz) while transmitting data in mmWave frequencies. However, this approach is costly in terms of operating two sets of radios at higher and lower frequencies. An optimized beam sweeping approach is proposed in [12] which extends the exhaustive search algorithms, previously used for BA in static scenarios, to support user mobility. In this paper, we propose a robust method for extending hierarchical search algorithms to support user mobility which leads to a higher performance than beam sweeping methods due to lower tracking overhead.

We consider a single-cell scenario, where the BS transmits data to a user over a mmWave link. In order to enable reliable data communication through narrow beams, the BS is required to allocate a fraction of the time slots to estimate the angle of arrival (AoA), and the remaining time to data communication. To refine the estimation of AoA, the BS scans regions of uncertainty of AoA by transmitting probing packets and analyzing the user response. In this paper, one of the goals is *to find the optimal probing strategy leading to a narrower region of uncertainty for AoA*. We use dynamic programming to show that *onward bisection* is optimal among all fractional search strategies, which scan a fraction of uncertainty region at each probing time slot.

When the user is mobile, the uncertainty of AoA increases during data communication. Consequently, the BS needs to expand the beamwidth to compensate for location uncertainty caused by user mobility. To avoid low BF gains caused by widening the beam in the proposed scheme, each data communication phase is followed by a probing phase reducing the reliable beamwidth for the next data communication phase. On the one hand, spending more time on the probing leads to narrower beams which increases the BF gain during the next data communication phase, but on the other hand less

This work is supported by NEC Laboratories America and National Science Foundation grants 1547332 and 1527750.

time remains for data communication. Consequently, there is a trade-off between the BF gain during data communication and the duration of data transmission. A similar trade-off has been identified for the duration of channel estimation in multiple-antenna systems [13]. In order to balance this trade-off, we provide a steady state analysis based on which an average throughput optimization problem is formulated to answer the following questions: *i) For how long should we probe?* and *ii) How long is the probing good for (how long to communicate data after each probing phase)?* We prove that the optimal scheme spends only one time slot on probing at each probing phase. Furthermore, we prove that the average throughput is an strictly quasi-concave function of the duration of data communication. Hence, the optimal duration of data transmission can be obtained efficiently using quasi-concave programming.

In practice, there are various constraints such as minimum feasible beamwidth and imperfect beam patterns. Moreover, the communication channel can include multiple spatial clusters. We generalize our proposed approach to satisfy these practical constraints. More precisely, we find the optimal solution to the average throughput maximization problem when there is a minimum beamwidth constraint. Furthermore, we generalize the proposed method to support imperfect beam patterns and multi-path channels.

The rest of the paper is organized as follows. In Section II, we provide the system model. We describe the proposed beam tracking approach in Section III. Section IV provides the steady state analysis, frame optimization problem, and the solution. We present the simulation results in Section V and conclude in Section VI.

II. System Model

We consider the downlink of a single-cell system where the BS communicates with a mobile user through a mmWave link which is already established by an initial access algorithm [5]. We assume that the propagation channel between the BS and the user consists of a single path (spatial cluster). This assumption has been adopted in several previous studies such as [9], [12], and is supported by channel measurements in mmWave frequencies. For instance, it is shown in [3] that with a high probability the channel incorporates one or two clusters in mmWave frequencies. In Section IV-C, we show that our proposed method also works well when the channel includes multiple paths.

Let $\phi(t)$ be the angle of arrival (AoA) of the signal received from the user at the BS. We note that $\phi(t)$ is time varying due to the user mobility. We define the angular velocity $\omega(t)$ as the rate of change of $\phi(t)$, i.e. $\omega(t) = \frac{d}{dt}\phi(t)$. We also note that $\omega(t)$ depends on the magnitude and direction of the user velocity as well as the environment. Furthermore, we define ω_{max} as the maximum angular velocity that the system can tolerate. In other words, a necessary condition for the proposed method to work properly is $-\omega_{max} \leq \omega(t) \leq \omega_{max}$. It should be noted that ω_{max} is an algorithm parameter which can be set sufficiently high by the BS such that this condition is satisfied. It will be shown in Section V that higher ω_{max} leads to a lower average throughput. Furthermore, note that if the necessary condition is not satisfied, the link may be broken and the BS should re-establish the link by performing another round of initial access.

We assume that the BS contains a massive antenna array as envisioned for mmWave communications [1]. To model the directionality of the BS transmission pattern due to BF, we adopt a *sectored antenna pattern* model from [14], characterized by three parameters: main-lobe gain G, beamwidth θ, and the angular coverage region Θ which is the angular region covered by the main-lobe of the transmission pattern. Clearly we have $\theta = |\Theta|$. Furthermore, we neglect the effect of the side-lobes for tractability. In this model, energy conservation implies that $G = \frac{2\pi}{\theta}$. We discuss more practical beam patterns where the roll-off is not sharp in Section IV-C. Let $G(t)$, $\theta(t)$, and $\Theta(t)$ be the BF gain, beamwidth, and the angular coverage region at time t, respectively, which are controlled by the BS to track and communicate with the mobile user as will be discussed later. We assume that the user has an omni-directional transmission and reception pattern. Investigation of directional pattern for user is left for future research.

In order to maintain the connection while using narrow beams, the BS is required to track the AoA while the user moves. We define the concept of uncertainty region as follows.

Definition 1 (AUR). *The angular uncertainty region (AUR) associated with AoA at time t, denoted by $\Phi(t)$, is the shortest angular interval $[a(t), b(t)]$, such that the BS knows $\phi(t) \in \Phi(t)$ with probability 1. Furthermore, the length of uncertainty region is defined as $u(t) \triangleq |\Phi(t)| = b(t) - a(t)$.*

According to this definition, if the BS matches its transmission pattern to the AUR (i.e. if $\Theta(t) = \Phi(t)$) during data transmission, the connection will be maintained with probability 1, leading to a reliable data communication. To investigate the evolution of AUR over time, assume that the BS transmits data to the user over the time interval $[0, t']$. Without tracking, AUR expands over time as AoA may change due to the user mobility. Using condition $-\omega_{max} \leq \omega(t) \leq \omega_{max}$ and Definition 1, it is straightforward to show that if $\Phi(0) = [a(0), b(0)]$, then we have $\Phi(t') = [a(t'), b(t')]$, where

$$a(t') = a(0) - \omega_{max}t', \quad b(t') = b(0) + \omega_{max}t', \tag{1}$$

$$u(t') = u(0) + 2\omega_{max}t'. \tag{2}$$

Fig. 1 illustrates the expansion of AUR without beam tracking over time. Note that AUR expands from both sides with the rate of ω_{max} as $\omega(t)$ can be positive or negative. Also, note that $\Phi(0)$ is the initial AUR which can be obtained from the initial access algorithm used to establish the connection. However, we will show in Section IV that the steady state parameters do not depend on $\Phi(0)$.

III. Beam Tracking and Data Communication

Equation (2) suggests that the BS should expand its transmission beamwidth $\theta(t)$ with the rate of $2\omega_{max}$ during data transmission so as to maintain the connection with probability

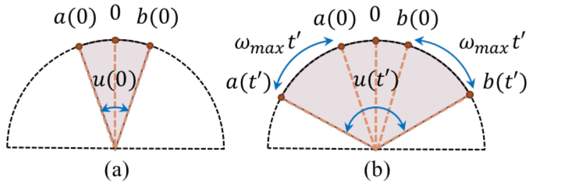


Fig. 1: The expansion of angular uncertainty region over The shaded region shows the AUR at (a) $t = 0$, and (b) $t = t'$. The vertical line is the angular origin, i.e. $\phi = 0$.

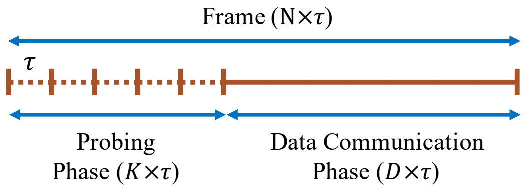


Fig. 2: Frame structure in the proposed beam tracking method.

1. However, this leads lower BF gains as time proceeds. In order to avoid low BF gains (wide beams) while maintaining connection with the mobile user, it is necessary for the BS to reduce the uncertainty of AoA intermittently. We assume that time is divided into similar *frames*. Fig. 2 illustrates the structure of a frame consisting of two phases: *i) probing phase (PP)* and *ii) data communication phase (DP)*. While the PP consists of $K \in \mathbb{N}$ time slots each with duration τ, we assume that the DP is not slotted as will be explained later. Furthermore, the duration of DP is assumed to be $D \times \tau$, where $D \in \mathbb{R}^+$. We conclude that the frame length is $N \times \tau$, where $N = K + D$.

In the PP, the BS probes the AUR while considering that AUR expands due to user mobility. At the beginning of each time slot, the BS transmits a probing packet while matching its angular coverage region to a fraction of AUR. Next, the AUR is updated based on the user feedback which is either an ACK or NACK. We consider the following assumptions:

i) The transmission and reception patterns of the BS are equivalent during each time slot.
ii) The length of the ACK packet and the processing time required for probing and ACK packets are negligible.

It is straightforward to relax the second assumption and generalize the results with minor modification. Furthermore, as we consider a single path for the propagation channel, receiving an ACK from the user implies that AoA has been in the angular coverage region used for probing packet transmission during that time slot, and NACK means otherwise. After PP, in the second phase (DP), the BS matches its angular coverage region to the most updated AUR and starts data communication while expanding the beamwidth continuously in time according to (2). Consequently, the AoA is in the coverage region constantly in DP. We assume that the communication between the BS and user is error free, that is, the user receives the packet correctly as long as AoA is in the coverage region of the BS. As a result, there is no need for user feedback in DP. Therefore, we assume that the DP is not time slotted.

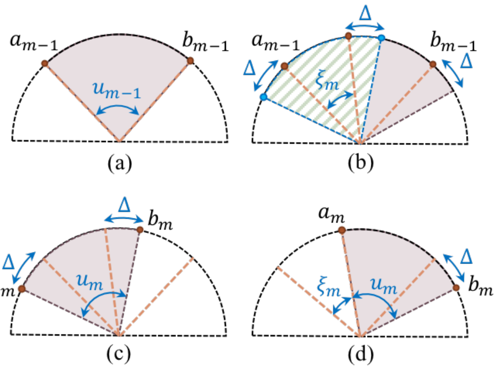


Fig. 3: Probing procedure at a generic time slot m.

Next, we describe each phase in details and analyze the variation of AUR over time, which will be used to perform a steady state analysis and optimize the frame structure later.

A. Probing Phase (PP)

In the PP, the duration of each time slot, τ, is the time required for transmitting the probing packet and receiving user response. Since PP is time slotted, we define $\Phi_m \triangleq \Phi(m\tau)$, $a_m \triangleq a(m\tau)$, $b_m \triangleq b(m\tau)$, $u_m \triangleq u(m\tau)$, and $\Theta_m \triangleq \Theta(m\tau)$ for notation brevity, where $m = 0, 1, \dots, K$. Note that $m = 0$ corresponds to the initial value at the beginning of the PP.

We focus on a family of probing strategies called *'onward fractional search (On-FS)'*. Algorithm 1 provides a step-by-step description of On-FS. Furthermore, the operation of On-FS at a generic probing time slot is depicted in Fig. 3. We note that the AUR is Φ_{m-1} at the beginning of time slot $m = 1, 2, \dots, K$ (Fig. 3a). The BS transmits a probing packet with size τ' while adjusting its transmission pattern to cover angular region $\Theta_m(\xi_m) = [a_{m-1} - \Delta, a_{m-1} + \xi_m + \Delta]$ where $\xi_m \in (-\Delta, u_{m-1}), m = 1, 2, \dots, K$ is a design parameter discussed later, and $\Delta = \tau\omega_{max}$ is the maximum change of AoA in the duration of one time slot (Fig. 3b). Note that $\Theta_m(\xi_m)$ is a fraction of the AUR at the end of time slot m if no probing occurs, i.e. $\Theta_m(\xi_m) \subset [a_{m-1} - \Delta, b_{m-1} + \Delta]$. The user transmits back an ACK if it receives the entire probing packet. Consequently, an ACK will be received by the BS if $\phi(t) \in \Theta_m(\xi_m)$ for $(m-1)\tau \leq t \leq (m-1)\tau + \tau'$. The BS needs to update the AUR based on the user response at the end of the time slot, hence we should have $\tau' \leq \tau$. We assume that $\tau' = \tau$ for the simplicity of expressions. However, the results can be generalized for $\tau' < \tau$ with minor modifications. If the BS does not receive an ACK, it is interpreted as NACK.

If the BS receives an ACK, it implies that the AoA has been in angular region $\Theta_m(\xi_m)$ for the entire time slot, hence AUR is updated as $\Phi_m = \Theta_m(\xi_m)$ (Fig. 3c). Otherwise, AoA has been outside of $\Theta_m(\xi_m)$ at least for a fraction of the time slot, hence the AUR is updated as $\Phi_m = \Theta'_m(\xi_m) \triangleq [a_{m-1} + \xi_m, b_{m-1} + \Delta]$ (Fig. 3d). There are two important observations: first, the AoA at the end of time slot m, i.e., ϕ_m belongs to region $\Theta_m(\xi_m) \cup \Theta'_m(\xi_m)$ with probability 1, if condition $-\omega_{max} \leq \omega(t) \leq \omega_{max}$ is satisfied. This is because $\Theta_m(\xi_m) \cup \Theta'_m(\xi_m) = [a_{m-1} - \Delta, b_{m-1} + \Delta]$, which is the AUR at the end of time slot m without probing which includes the AoA with probability 1, if condition $-\omega_{max} \leq \omega(t) \leq \omega_{max}$ is satisfied. Therefore, the BS will not lose the track of AoA regardless of the user response. Second, if a NACK happens, the updated AUR still includes a part of the probed angular region, i.e. $\Theta_m(\xi_m) \cap \Theta'_m(\xi_m) = [a_{m-1} + \xi_m, a_{m-1} + \xi_m + \Delta]$. The reason is that in the worst case, the AoA can be $\phi_{m-1} = a_{m-1} + \xi_m + \Delta + \epsilon$ with an arbitrary small $\epsilon > 0$ at the beginning of the time slot, and decrease with maximum rate ω_{max} during the time slot. In that case we have $\phi_m = a_{m-1} + \xi_m + \epsilon$ at the end of time slot m while the user does not send back an ACK since it has not received the entire probing packet (it misses the first ϵ portion of it). Therefore, $\Theta'_m(\xi_m)$ should also include a part of $\Theta_m(\xi_m)$ with length Δ to ensure that the BS keeps track of AoA correctly. Note that this overlap should be larger if the roll-off of the beam pattern is not sharp as will be studied in Section IV-C. We call this method onward fractional search since it probes a fraction of $[a_{m-1} - \Delta, b_{m-1} + \Delta]$ which is the AUR at the end of time slot m if no probing occurs.

Algorithm 1 Onward Fractional Search (On-FS)

```
Input: Initial AUR in the probing phase $\Phi_0 = [a_0, b_0]$
Output: Final AUR in the probing phase $\Phi_K = [a_K, b_K]$
for $m = 1$ to $K$ do
    The BS transmit a probing packet with a beam-pattern covering the angular region $[a_{m-1} - \Delta, a_{m-1} + \xi_m + \Delta]$ as illustrated by Fig. 3(b).
    if The BS receives an ACK from the user then
        the AUR, illustrated by Fig. 3(c), is updated as
            $\Phi_m = [a_m, b_m] = [a_{m-1} - \Delta, a_{m-1} + \xi_m + \Delta]$,   (3)
            $u_m = b_m - a_m = \xi_m + 2\Delta$.   (4)
    else
        the AUR, illustrated by Fig. 3(d), is updated as
            $\Phi_m = [a_m, b_m] = [a_{m-1} + \xi_m, b_{m-1} + \Delta]$,   (5)
            $u_m = b_m - a_m = u_{m-1} - \xi_m + \Delta$.   (6)
    end if
end for
```

To characterize On-FS, we need to determine the design vector $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_K]$. The goal of PP is to reduce the length of AUR at the beginning of DP so as to start data transmission with a narrower beam. Therefore, we want to reduce u_K as much as possible during PP. However, we note that u_K is a random variable whose distribution depends on $\boldsymbol{\xi}$ as well as the user mobility model which affects the probability distribution of $\phi(t)$ and $\omega(t)$ introduced in Section II. This is because AUR is updated based on the user feedback directly affected by the user mobility model. As the details of the mobility model is assumed to be arbitrary for robustness, we consider a deterministic approach to find $\boldsymbol{\xi}$ such that the worst case (maximum) of u_K over all mobility models satisfying $-\omega_{max} \leq \omega(t) \leq \omega_{max}$ is minimized. Therefore, we formulate the problem as

$$\begin{aligned} \mathbf{\Pi_1}: \quad & \boldsymbol{\xi}^* = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} \max_{\mathbb{M}} \big(u_K(\boldsymbol{\xi})\big), \\ & \text{subject to: } -\Delta < \xi_m < u_{m-1}, \quad m = 1, 2, \ldots, K. \end{aligned}$$

where $\mathbb{M}$ is the set of all mobility models satisfying $-\omega_{max} \leq \omega(t) \leq \omega_{max}$.

Lemma 1. *The solution of problem* $\mathbf{\Pi_1}$ *is*

$$\xi^*_m = (u_{m-1} - \Delta)/2, \ m = 1, 2, \ldots, K.$$

The sketch of proof is provided in Appendix A. The proof of Lemma 1 reveals that using the optimal vector $\boldsymbol{\xi}^*$, the length of AUR at the end of each probing time slot is independent of the user feedback, i.e. $|\Theta_m(\xi^*_m)| = |\Theta'_m(\xi^*_m)|$, $m = 1, 2, \ldots, K$. As a result, we call On-FS, *onward bisection (On-Bi)* when we use $\boldsymbol{\xi}^*$ since the length of AUR is equal given ACK and NACK. This implies that $u_m, m = 1, 2, \ldots, K$ is independent of user mobility model when $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ as expected. Furthermore, using $\boldsymbol{\xi}^*$ we have

$$u_m = \frac{u_{m-1}}{2} + \frac{3\Delta}{2}, \quad 1 \leq m \leq K. \tag{7}$$

It is straightforward to show that (7) leads to

$$u_m = \frac{u_0}{2^m} + 3\Delta\left(1 - \frac{1}{2^m}\right), \quad 1 \leq m \leq K, \tag{8}$$

where u_0 is the length of AUR at the beginning of PP.

B. Data Communication Phase (DP)

In this phase, the BS matches its angular coverage region to the updated AUR from PP, i.e. Φ_K, and starts data transmission. To avoid connection loss during data transmission due to user mobility, the BS expands the beamwidth to constantly cover AUR, i.e. $\Theta(t) = \Phi(t)$ for $K\tau \leq t \leq N\tau$. So we have

$$\theta(t) = u_K + 2\omega_{max}(t - K\tau), \quad K\tau \leq t \leq N\tau. \tag{9}$$

We define the average throughput of the frame as follows

$$\bar{R}(K, D, u_0) = \frac{1}{N\tau} \int_{K\tau}^{N\tau} \log_2 \Big(1 + \text{SNR}(K, D, u_0, t)\Big) dt, \tag{10}$$

where, the signal to noise ratio (SNR) is defined as

$$\text{SNR}(K, D, u_0, t) = \frac{PG(t)}{N_0 W L}, \tag{11}$$

where, P is the BS transmit power, $G(t) = 2\pi/\theta(t)$ is the BF gain, N_0 is the noise spectral density, W is the communication bandwidth, and L is the path loss between the BS and the user. We note that $\bar{R}$ is a function of K, D, and u_0. The dependency of $\bar{R}$ on u_0 originates from $\theta(t)$ in which u_K depends on u_0 according to (8). Later, we will use average throughput as the objective function to optimize the frame structure.

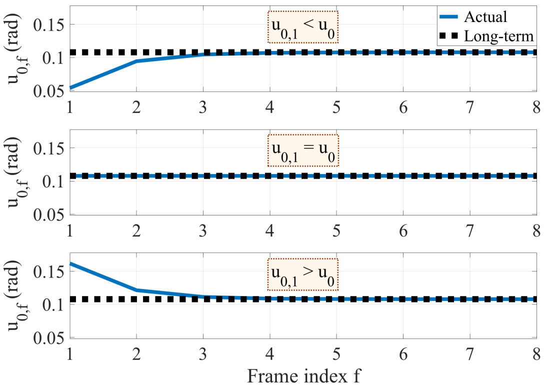


Fig. 4: The convergence of $u_{0,f}$ (solid curve) to u_0 (dotted line) when $K = 2$ and $D = 100$.

IV. Steady State Analysis and Frame Optimization

In this section, we first analyze the steady state performance of the proposed approach when many similar frames are concatenated. Next, we formulate an optimization problem to find the optimal duration of PP and DP maximizing steady state average throughput.

A. Steady State Analysis

In this section, we assume that a large number of similar frames are concatenated. Let $u_{m,f}$ denote the length of AUR at the end of time slot $m \in \{1, 2, \dots K\}$ during PP in frame $f \in \mathbb{N}$. Also, let $u_{0,f}$ be the length of AUR at the beginning of frame f. We note that AUR at the end of frame f is equal to the AUR at the beginning of frame $f+1$. Thus according to (8) and (9) we have

$$u_{0,f+1} = \frac{u_{0,f}}{2^K} + 3\Delta\left(1 - \frac{1}{2^K}\right) + 2\Delta D, \quad f \in \mathbb{N}. \quad (12)$$

Lemma 2. *For any $u_{0,1}$, sequence $\{u_{0,f}\}_{f\in\mathbb{N}}$ converges to*

$$u_0 \triangleq 2\Delta\left(\frac{3}{2} + \frac{2^K}{2^K - 1}D\right). \quad (13)$$

The proof is omitted due to the lack of space. The main idea of the proof is that the sequence is either constant or monotonic and bounded. Consequently, there exists a finite limit which can be obtained by assuming that $u_{0,f+1}$ is equal to $u_{0,f}$ in (12). Lemma 2 reveals an important property of the multi-frame structure. As time proceeds, the length of initial AUR of the frames converges to u_0, introduced in (13), which is independent of the initial AUR of the first frame, i.e. $u_{0,1}$. In practice, after some initial frames, the transient effect of $u_{0,1}$ fades away and the system converges to a steady state. Fig. 4 illustrates this property for three cases: *i)* $u_{0,1} > u_0$, *ii)* $u_{0,1} = u_0$, and *iii)* $u_{0,1} < u_0$ when $K = 2$ and $D = 100$. We observe that $u_{0,f}$ (solid curve) converges to its limit u_0 (dotted line) fast in every case.

Remark 1. *Using the same approach, it can be shown that sequence $\{u_{m,f}\}_{f\in\mathbb{N}}$ converges for every $m \in \{1, 2, \dots K\}$.*

Hereafter, we assume $u_{0,1} = u_0$. As all of the frames are similar in the steady state, the analysis is focused on a single frame. Considering (8) with $m = K$ and substituting u_0 with its value provided in (13) lead to

$$u_K = 2\Delta\left(\frac{3}{2} + \frac{D}{2^K - 1}\right), \quad (14)$$

where u_K is the AUR at the end of PP and at the beginning of DP in the steady state. We note that u_K is a function of K and D. Let $\widehat{R}(K, D)$ denote the average throughput in the steady state. Using (9)-(11) we have

$$\widehat{R}(K, D) = \frac{1}{N\tau}\int_{K\tau}^{N\tau} \log_2\left(1 + \frac{A}{u_K + 2\Delta(\frac{t}{\tau} - K)}\right) dt, \quad (15)$$

where, $A = \frac{2\pi P}{N_0 W L}$. Taking the integral in (15) and using steady state value of u_K provided in (14) lead to

$$\begin{aligned}\widehat{R}(K, D) = \frac{1}{(K+D)}\Bigg[&\widehat{P}\log_2\left(1 + \frac{D}{\frac{D}{2^K-1} + \alpha + \widehat{P}}\right)\\ &+ \left(\frac{2^K}{2^K-1}D + \alpha\right)\log_2\left(1 + \frac{\widehat{P}}{\frac{2^K}{2^K-1}D + \alpha}\right)\\ &- \left(\frac{D}{2^K-1} + \alpha\right)\log_2\left(1 + \frac{\widehat{P}}{\frac{D}{2^K-1} + \alpha}\right)\Bigg], \quad (16)\end{aligned}$$

where $\widehat{P} = \frac{A}{2\Delta}$ and $\alpha = 3/2$.

B. Frame Optimization

Increasing number of probing time slots K leads to a narrower AUR during data transmission according to (9) and (14); hence a higher SNR can be achieved in DP due to the higher BF gain which can potentially increase the average throughput. On the other hand, increasing K reduces the fraction of the time spent on data communication (i.e. $\frac{D}{K+D}$) which can reduce the average throughput. In this section, we study this trade-off by formulating an optimization problem to find the optimal values of K and D maximizing steady state average throughput $\widehat{R}(K, D)$ provided in (16). The problem can be formulated as

$$\mathbf{\Pi_2}: \quad (K^*, D^*) = \underset{K\in\mathbb{N}, D\in\mathbb{R}^+}{\operatorname{argmax}} \quad \widehat{R}(K, D).$$

We solve optimization problem $\mathbf{\Pi_2}$ in two steps: i) we prove that $K^* = 1$, and ii) we find D^* while assuming $K = K^* = 1$.

Theorem 1. *The optimal value of K in $\mathbf{\Pi_2}$ is $K^* = 1$.*

The proof is provided in Appendix B. The idea of the proof is to compare the case (K, D) with the case $(1, D/K)$ and show that the latter outperforms the former, i.e. $\widehat{R}(K, D) \leq \widehat{R}(1, D/K)$, $\forall K \in \mathbb{N}, D \in \mathbb{R}^+$. Consequently we have $K^* = 1$, which means that the probing slots should be distributed in time to avoid large AURs. Using Theorem 1, problem $\mathbf{\Pi_2}$ is reduced to

$$\mathbf{\Pi_3}: \quad D^* = \underset{D\in\mathbb{R}^+}{\operatorname{argmax}} \quad \widehat{R}(1, D).$$

To solve this problem, we first show that the objective function, $\widehat{R}(1, D)$, is strictly quasi-concave with D on the domain.

Theorem 2. *$\widehat{R}(1, D)$ is an strictly quasi-concave function of D on $\mathbb{R}^+$ and there exists a unique $D^* \in \mathbb{R}^+$ at which $\widehat{R}(1, D)$ is maximized.*

The proof is provided in Appendix C. Theorem 2, implies that problem $\mathbf{\Pi_3}$ is quasi-concave. Moreover, problem $\mathbf{\Pi_3}$ is a one-dimensional optimization problem, hence a simple one-dimensional bisection method can find D^* with logarithmic numerical complexity [15, Chapter 2].

C. Practical Considerations

Even with a large number of antennas, there is a minimum beamwidth that can be realized in practice. If we assume a minimum beamwidth θ_{min}, then the optimal data transmission duration (i.e. D^* in $\mathbf{\Pi_3}$) may not conform to such requirement. The following lemma, proved in Appendix D, provides the optimal frame structure given a minimum beamwidth constraint.

Lemma 3. *If there is minimum constraint on the beamwidth θ, i.e. if $\theta \geq \theta_{min}$, then the solution of problem $\mathbf{\Pi_1}$ is $K^* = 1$ and $D^* = \max\{D_1^*, D_2^*(\theta_{min})\}$ where D_1^* is the solution of problem $\mathbf{\Pi_3}$ and $D_2^*(\theta_{min}) = \frac{\theta_{min}}{2\Delta} - \frac{3}{2}$.*

Although we considered a single path for the channel between the BS and the user in previous sections, the channel may contain multiple paths in practice. In such conditions, the proposed scheme is guaranteed to find at least one of the paths for data transmission. When there are more than one path, receiving an ACK at the end of the probing time slot m implies that AoA of at least one of them belongs to the probed region Θ_m introduced in Section III-A. On the other hand, receiving NACK implies that all the remaining paths belong to the angular region Θ'_m defined in Section III-A. Therefore, the AoA of at least one of the paths belong to the updated AUR at the end of time slot m. This implies that at the beginning of data transmission, the AoA of one or more paths belong to the updated AUR.

We considered an ideal beam pattern with a sharp roll-off (i.e. sharp edges) to model analog beamforming in Section II. However, it is possible to generalize the proposed method such that it works properly with more practical beam patterns where the roll-off is not sharp [16]. For beams with sharp edges, we showed in Section III-A that Θ_m and Θ'_m should have an overlap of length Δ. If the roll-off is not sharp, the uncertainty regions given ACK and NACK, i.e. Θ_m and Θ'_m, should have a larger overlap to compensate the lower BF gain in the roll-off region and ensure that the algorithm does not miss the AoA. It should be noted that the cost of this reliability is a lower BF gain caused by widening the beam pattern.

V. Numerical Results and Simulations

In this section, we provide extensive simulation results to evaluate the performance of the proposed method. Table I lists the simulation parameters. As a benchmark, we consider an optimized beam sweeping approach proposed in [12] which supports user mobility. In [12], the authors also consider an adaptation of IEEE 802.11ad to their proposed model. We consider it as another benchmark and refer the reader to [12, Section IV] for more details due to the lack of space.

Fig. 5a illustrates the average throughput as a function of BS transmit power when $\omega_{max} = 300$ degree/sec. We observe that our proposed method (with optimized frame structure) outperforms beam sweeping proposed in [12]. The reason is that onward bisection used in our approach reduces the length of AUR more than beam sweeping, leading to a higher BF gain during DP. Furthermore, we observe a large gap between the performance of the proposed method and that of 802.11ad which is because fixed 7° beams are used in 802.11ad while the beamwidth is optimized in our proposed method.

Fig. 5b displays the percentage of beam tracking overhead in the proposed method using the optimal frame structure, which is defined as $\frac{K^*}{K^*+D^*} \times 100\%$ where $K^* = 1$ according to Theorem 1 and D^* is obtained by solving $\mathbf{\Pi_3}$. Considering $P = 20$ dBm, we observe that the tracking overhead is between 8% and 12%.

Fig. 5c illustrates the optimized average throughput $\widehat{R}(1, D^*)$ as a function of maximum angular velocity ω_{max} for different values of transmit power P. According to (2), AUR expands faster for higher values of ω_{max}. As a result, the BS is required to perform PP more frequently so as to avoid low BF gains. However, this increases beam tracking overhead leading to a lower average throughput.

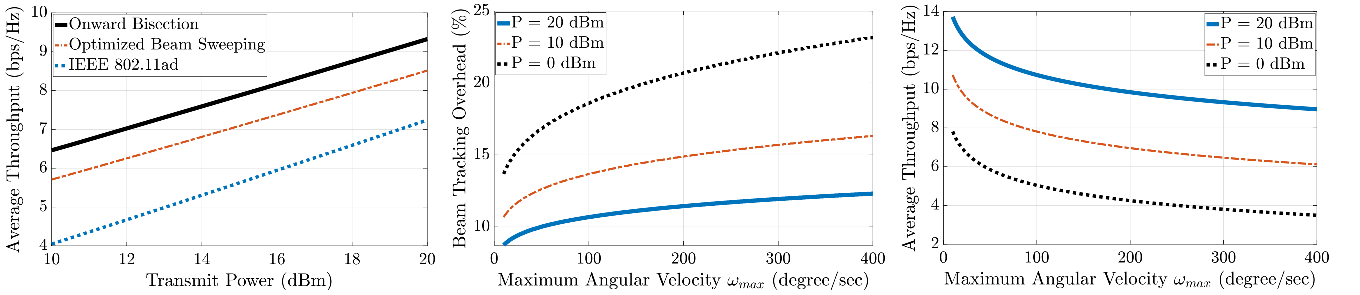
In Fig 6, we plot the average throughput as a function of BS transmit power when there is a minimum beamwidth constraint $\theta \geq \theta_{min} \in \{2°, 4°, 8°, 16°\}$ as discussed in Section IV-C. A larger θ_{min} restricts the BF gain more leading to a lower average throughput as expected. Furthermore, we note that adding 3 dB to the transmit power can approximately compensate for the throughput loss incurred by doubling the minimum feasible beamwidth.

VI. Conclusion

In this paper, we have proposed a joint beam tracking and data communication scheme for mmWave mobile networks. In this scheme, the BS increases the beamwidth during data transmission to compensate for location uncertainty caused by user mobility. To avoid low beamforming gains due to widening the beam, we have proposed a probing strategy to refine the estimation of AoA after each data communication phase. Furthermore, we have formulated an optimization problem to optimize the duration of probing and data communication

TABLE I: Simulation parameters

Parameter	**Value**
Distance between BS and user d	25 m
Probing time slot duration τ	100 μsec
Maximum angular velocity ω_{max}	$10 - 400$ degree/sec
Frequency	60 GHz
Bandwidth W	2 GHz
BS transmit power P	$0 - 20$ dBm
Noise spectral density N_0	-174 dBm/Hz
Pathloss $L(d)$ (in dB)	$68 + 20\log_{10}(d$ in m)



(a) (b) (c)

Fig. 5: (a) Average throughput (bps/Hz) as a function of BS transmit power (dBm) for different methods, (b) Beam tracking overhead (%) as a function of maximum angular velocity (degree/sec) for different transmit powers, (c) Average throughput (bps/Hz) as a function of maximum angular velocity (degree/sec) for different transmit powers.

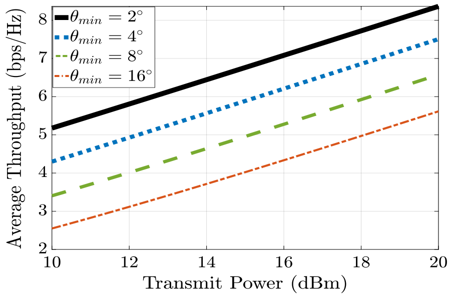


Fig. 6: Average throughput as a function of transmit power for different minimum beamwith constraints.

phases. A natural extension to this work is to consider directional beam patterns at user device, as well as the possibility of communication errors which can affect the performance.

Appendix

A. Proof of Lemma 1

We use dynamic programming to prove the statement. Let $\boldsymbol{\xi}^{[m]}, m \in \{1,2,\dots,K\}$ denote the vector including the first m elements of $\boldsymbol{\xi}$. Clearly, u_m is a function of $\boldsymbol{\xi}^{[m]}, m \in \{1,2,\dots,K\}$. First, we consider the last probing time slot, i.e. time slot K. According to (4) and (6), we have:

$$u_K(\boldsymbol{\xi}^{[K]}) = \begin{cases} \xi_K + 2\Delta, & \text{ACK}, \\ u_{K-1}(\boldsymbol{\xi}^{[K-1]}) - \xi_K + \Delta, & \text{NACK}. \end{cases}$$

Therefore, irrespective of $\boldsymbol{\xi}^{[K-1]}$, maximizing the minimum of $u_K(\boldsymbol{\xi}^{[K]})$ requires that $\xi_K^* + 2\Delta = u_{K-1}(\boldsymbol{\xi}^{[K-1]}) - \xi_K^* + \Delta$. This is due to the property of min max problem and leads to $\xi_K^* = (u_{K-1}(\boldsymbol{\xi}^{[K-1]}) - \Delta)/2$. Note that ξ_K^* is a feasible solution to problem $\mathbf{\Pi_1}$. This equality introduces the optimal value of ξ_K as a function of $\boldsymbol{\xi}^{[K-1]}$, leading to $u_K(\boldsymbol{\xi}^{[K]}) = (u_{K-1}(\boldsymbol{\xi}^{[K-1]}) + \Delta)/2$. Therefore, minimizing the maximum of $u_K(\boldsymbol{\xi}^{[K]})$, it is equivalent to use ξ_K^* introduced above as well as minimizing the maximum of $u_{K-1}(\boldsymbol{\xi}^{[K-1]})$. Repeating this top-down procedure leads to $\xi_m^* = (u_{m-1}(\boldsymbol{\xi}^{[m-1]}) - \Delta)/2, m = 1,2,\dots,K$.

B. Proof of Theorem 1

If $K = 1$ then the Lemma's statement is followed. Let $K \geq 2$ is optimal. We define $f(\widehat{P},K,D) \triangleq \widehat{R}(1,D/K) - \widehat{R}(K,D)$. The goal is to prove that $f(K,D,\widehat{P}) \geq 0, \forall K \in \mathbb{N}, \forall D \geq 0, \forall \widehat{P} \geq 0$. We note that $f(K,D,0) = 0$. It is sufficient to show that $\frac{\partial}{\partial \widehat{P}} f(K,D,\widehat{P}) \geq 0, \forall K \in \mathbb{N}, \forall D \geq 0, \forall \widehat{P} \geq 0$. Using Leibniz's integral rule, $\frac{\partial}{\partial \widehat{P}} f(K,D,\widehat{P})$ is equal to

$$\frac{K}{K+D} \log_e \left(\frac{\alpha + \widehat{P} + \frac{2D}{K}}{\alpha + \widehat{P} + \frac{D}{K}} \right) - \frac{1}{K+D} \log_e \left(\frac{\alpha + \widehat{P} + \frac{2^K D}{2^K - 1}}{\alpha + \widehat{P} + \frac{D}{2^K - 1}} \right).$$

It is straightforward to show that, $\frac{\partial}{\partial \widehat{P}} f(K,D,\widehat{P}) \geq 0$ is equivalent to

$$\frac{\left(Q + \frac{2D}{K}\right)^K (Q + \eta D)}{\left(Q + \frac{D}{K}\right)^K (Q + (\eta+1)D)} \geq 1, \tag{17}$$

where, $Q \triangleq \alpha + \widehat{P}$, and $\eta = \frac{1}{2^K - 1}$. Using binomial theorem it is straightforward to show that (17) is equivalent to

$$\frac{B_{-1} D^{K+1} + \sum_{i=0}^{K} B_i Q^{i+1} D^{K-i}}{C_{-1} D^{K+1} + \sum_{i=0}^{K} C_i Q^{i+1} D^{K-i}} \geq 1 \tag{18}$$

where,

$$B_{-1} = C_{-1} = \frac{2^K K^{-K}}{2^K - 1},$$
$$B_i = \binom{K}{i} \left(\frac{2}{K}\right)^{K-i} + \eta \binom{K}{i+1} \left(\frac{2}{K}\right)^{K-i-1},$$
$$C_i = \binom{K}{i} \left(\frac{1}{K}\right)^{K-i} + (\eta+1) \binom{K}{i+1} \left(\frac{1}{K}\right)^{K-i-1},$$

and $i \in \{0,1,\dots K\}$. To prove the correctness of (18), it is sufficient to prove that $B_i \geq C_i, i = 0,1,\dots,K$ for every $K \geq 2$ which is equivalent to proving the following inequality for $i = 0,1,\dots,K$:

$$E_i \triangleq (i+1)\left(2^{K-i+1} - 2\right) + (K-i)\left(\eta 2^{K-i-1} - \eta - 1\right) \geq 0.$$

We note that $E_0 = 2^{K+1} - 2 > 0$ since $K \geq 2$. Moreover we have $E_K = 0$. Furthermore, it can be shown that E_i is decreasing with i for $i \geq 1$. Consequently, we have $E_1 > E_2 > \dots > E_K = 0$ which concludes the proof.

C. Proof of Theorem 2

Let $f(D) \triangleq \widehat{R}(1,D)$ where $\widehat{R}(K,D)$ is given by (16). By taking derivative of $f(D)$ with respect to D and simplifying the expressions we have

$$f'(D) = \frac{1}{2(D+1)^2}\left[\log_2\left(1+\frac{\widehat{P}}{2D+\alpha}\right) + \log_2\left(1+\frac{\widehat{P}}{D+\alpha}\right) - 2\widehat{P}\log_2\left(1+\frac{D}{D+\alpha+\widehat{P}}\right)\right].$$

Let $g(D) \triangleq 2(D+1)^2 f'(D)$. Since $D \geq 0$, it is clear that $g(D)$ is of the same sign as $f'(D)$. Furthermore, by taking derivative of $g(D)$ with respect to D we have

$$g'(D) = -\widehat{P}\log_2(e)\left[\frac{2}{(2D+\alpha)(2D+\alpha+\widehat{P})} + \frac{1}{(D+\alpha)(D+\alpha+\widehat{P})} + \frac{2(\alpha+\widehat{P})}{(D+\alpha+\widehat{P})(2D+\alpha+\widehat{P})}\right], \tag{19}$$

where e is the Euler's number. We observe that $g'(D) < 0, \forall D \geq 0$. Therefore, $g(D)$ is monotonically decreasing over $[0,\infty)$. Besides, we have $g(0) = 2\log_2(1+\alpha^{-1}\widehat{P}) > 0$ and $\lim_{D\to\infty} g(D) = -2\widehat{P} < 0$. Therefore, according to the intermediate value theorem, there exist a finite D^* such that $g(D) = f'(D) = 0$. Furthermore, D^* is unique due to monotonicity of $g(D)$. Moreover, $g(D)$ (and consequently $f'(D)$) is positive, and negative over the intervals $[0,D^*)$ and (D^*,∞), respectively. We conclude that function $f(D)$ is monotonically increasing over $[0,D^*)$ and monotonically decreasing over (D^*,∞). Therefore, $f(D)$ is strictly quasi-concave and it takes its maximum value at D^*.

D. Proof of Lemma 3

It is straightforward to show that the minimum length of AUR occurs at the end of PP (equivalently at the beginning of DP) in steady state. Therefore, $\theta \geq \theta_{min}$ is equivalent to $u_K \geq \theta_{min}$ since the beamwidth θ is equal to the length of AUR during DP. Therefore, in order to model minimum beamwidth constraint, we merely need to add $u_K \geq \theta_{min}$ as a new constraint to the problem $\mathbf{\Pi_2}$ and solve it. Let $\mathbf{\Pi_4}$ denote the new optimization problem. Using a similar approach as the one in Section IV-B, we first show that $K^* = 1$ is still optimum. To this end, we also use the inequality $u_K(K^*,D^*) < u_K(1,D^*/K^*)$, which holds because $2^{K^*}-1 > K^*, \forall K^* > 1$), to show that if (K^*,D^*) is a feasible solution to problem $\mathbf{\Pi_4}$, then $(1,D^*/K^*)$ is also a feasible solution to that problem. Considering $K = K^* = 1$, problem $\mathbf{\Pi_3}$ is simplified to another problem, denoted by $\mathbf{\Pi_5}$, in which D^* is the only optimization variable. To find D^* we use the result of Theorem 2. Let D_1^* denote the solution of problem $\mathbf{\Pi_3}$. Noe that $\mathbf{\Pi_3}$ is a relaxed version of $\mathbf{\Pi_5}$ since it does not have a minimum beamwidth constraint. We consider the following two cases.

i) **Case 1:** If D_1^* is a feasible solution to $\mathbf{\Pi_5}$, i.e. if $u_K(1,D_1^*) \geq \theta_{min}$, then it is also the optimal solution of $\mathbf{\Pi_5}$, i.e. we have $D^* = D_1^*$.

ii) **Case 2:** Otherwise, if $u_K(1,D_1^*) < \theta_{min}$, D_1^* is not a feasible solution to $\mathbf{\Pi_5}$ and we have $D^* > D_1^*$ since u_K is increasing with D according to (14). Furthermore, we note that the objective function of $\mathbf{\Pi_5}$, i.e. $\widehat{R}(1,D)$, is strictly quasi-concave with respect to D and takes its maximum value at $D = D_1^*$ according to Theorem 2. Therefore $\widehat{R}(1,D)$ is decreasing with D for $D > D_1^*$. As u_K is increasing with D and $D^* > D_1^*$, we conclude that $u_K(1,D^*) = \theta_{min}$ is optimum, i.e. the minimum beamwidth constraint is active at the optimal solution. Using (14) and solving this equation for D^* leads to $D^* = D_2^*(\theta_{min}) \triangleq \frac{\theta_{min}}{2\Delta} - \frac{3}{2}$.

We conclude that $D^* = \max\{D_1^*, D_2^*(\theta_{min})\}$.

References

[1] 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, March 2014.

[2] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, 2015.

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

[4] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: mm-wave beam steering without in-band measurement," in *IEEE Conference on Computer Communications (INFOCOM)*, 2015.

[5] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmwave frequencies," *arXiv preprint arXiv:1804.01908*, 2018.

[6] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, "Comparative analysis of initial access techniques in 5G mmwave cellular networks," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 268–273.

[7] 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, 2016.

[8] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmwave communications," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 1926–1930.

[9] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *Information Theory and Applications Workshop (ITA), 2017*. IEEE, 2017, pp. 1–6.

[10] 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.

[11] "IEEE Std 802.11ad," *IEEE Standard*, pp. 1–634, March 2014.

[12] 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.

[13] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[14] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2015.

[15] M. Bartholomew-Biggs, "One-variable optimization," in *Nonlinear Optimization with Engineering Applications*. Springer, 2008, pp. 1–22.

[16] M. Hussain, D. J. Love, and N. Michelusi, "Neyman-pearson codebook design for beam alignment in millimeter-wave networks," in *Proceedings of the 1st ACM Workshop on Millimeter-Wave Networks and Sensing Systems 2017*. ACM, 2017, pp. 17–22.