Fast mmwave Beam Alignment via Correlated Bandit Learning

Wen Wu, *Student Member, IEEE,* Nan Cheng, *Member, IEEE,* Ning Zhang, *Senior Member, IEEE,*
Peng Yang, *Member, IEEE,* Weihua Zhuang, *Fellow, IEEE,* and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—Beam alignment (BA) is to ensure the transmitter and receiver beams are accurately aligned to establish a reliable communication link in millimeter-wave (mmwave) systems. Existing BA methods search the entire beam space to identify the optimal transmit-receive beam pair, which incurs significant BA latency on the order of seconds in the worst case. In this paper, we develop a learning algorithm to reduce BA latency, namely Hierarchical Beam Alignment (HBA) algorithm. We first formulate the BA problem as a stochastic multi-armed bandit problem with the objective to maximize the cumulative received signal strength within a certain period. The proposed algorithm takes advantage of the *correlation structure* among beams such that the information from nearby beams is extracted to identify the optimal beam, instead of searching the entire beam space. Furthermore, the *prior knowledge* on the channel fluctuation is incorporated in the proposed algorithm to further accelerate the BA process. Theoretical analysis indicates that the proposed algorithm is asymptotically optimal. Extensive simulation results demonstrate that the proposed algorithm can identify the optimal beam with a high probability and reduce the BA latency from hundreds of milliseconds to a few milliseconds in the *multipath* channel, as compared to the existing BA method in IEEE 802.11ad.**

Index Terms – **mmwave, beam alignment, correlation structure, prior knowledge, multi-armed bandit.**

I. Introduction

The ever-increasing data traffic driven by various emerging data-hungry applications, such as high-definition mobile video streaming, cordless virtual reality gaming and wireless fiber-to-home access, has placed a growing strain on the creaking traditional cellular networks. Millimeter-wave (mmwave) communication is envisioned as the most promising technology to accommodate the skyrocketing data traffic through harnessing multi-GHz bandwidths. Multiple standardization efforts, such as IEEE 802.11ad [1], [2] and ongoing IEEE 802.11ay [3], [4], and large-scale field-trials have paved the road for the commercialization of mmwave communications.

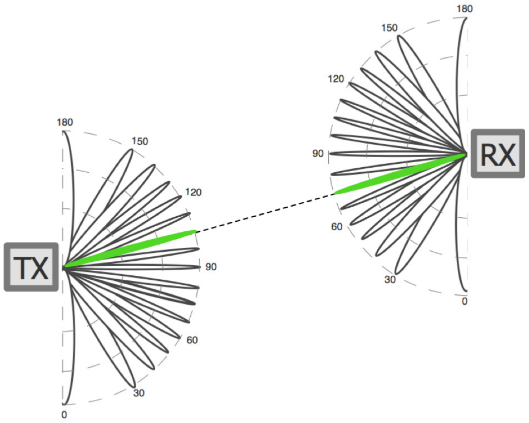


Fig. 1. A beam alignment example with 16 beams. The well-aligned transmitter and receiver beams are represented by solid green beams.

In mmwave communication systems, narrow directional beams are adopted at both the transmitter and receiver to compensate for the huge attenuation loss. Since beams are narrow, the communication is possible only when the transmitter and receiver beams are properly aligned [5], as shown in Fig. 1. Beam alignment (BA) is such a process to identify the optimal transmit-receive beam pair which attains the maximum received signal strength (RSS). Beam misalignment can dramatically reduce the link budget and drop the throughput from multiple Gbps to a few hundred Mbps [6]. As a key process in mmwave communications, BA is of significance to achieve multi-gigabit wireless transmission. To identify the best beam pair, a naive exhaustive search method scans all the combinations of the transmitter and receiver beams, which results in significant BA latency. Yet, a low-latency BA process is imperative for practical mmwave systems to accommodate real-time applications. Moreover, in mobile scenarios, user mobility changes the beam direction and thus frequently invokes BA, which further exacerbates the latency. To accelerate the beam search, IEEE 802.11ad protocol decouples the BA process into two steps. Firstly, the transmitter starts with a quasi-omnidirectional beam and the receiver scans the beam space for the best receiver beam. Secondly, the transmitter scans the beam space for the best transmitter beam while keeping the receiver quasi-omnidirectional. Still, the existing BA method in IEEE 802.11ad may take up to seconds with a large number of candidate beams [7]. To reduce BA latency, can we identify the optimal beam without searching the entire beam space?

In the literature, there are some initial research efforts to address this challenge. Utilizing the sparse characteristic of the mmwave channel, Marzi *et al.* developed a compressed sensing BA method [8]. Some out-of-band information, e.g.,

W. Wu, P. Yang, W. Zhuang, and X. Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada (e-mail:{w77wu, p38yang, wzhuang, sshen}@uwaterloo.ca). *Corresponding author: Peng Yang.*

N. Cheng is with the School of Telecommunication, Xidian University, Xian 710071, China (e-mail: dr.nan.cheng@ieee.org).

N. Zhang is with Department of Computing Sciences, Texas A&M University at Corpus Christi, TX, USA (e-mail: ning.zhang@tamucc.edu).

the Wi-Fi signal, is exploited to identify the optimal beam in [9]. These works perform BA with the assistance of excessive extra information besides RSS. Surprisingly, a crucial feature, the correlation structure among beams, is ignored in previous works. In fact, the RSS of nearby beams is similar which means nearby beams are highly correlated. In this way, if a beam does not perform well, its nearby beams are highly likely to perform worse either. The measurement of one beam not only reveals information about itself, but also its nearby beams. Hence, the information from nearby beams can be learned to identify the optimal beam without searching the entire beam space.

In this paper, we propose a fast BA algorithm, named Hierarchical Beam Alignment (HBA), by utilizing the *correlation structure* among beams and the *prior knowledge* on the channel fluctuation. In the BA problem, fast BA means identifying the optimal beam with the minimum latency. This problem boils down to sequentially selecting beams to maximize the cumulative RSS within a certain period, which can be formulated as a stochastic multi-armed bandit (MAB) problem. To solve this problem efficiently, two unique characteristics are incorporated in our proposed algorithm. Firstly, theoretical analysis indicates that the correlation structure among beams in the *multipath* channel can be characterized by a *multimodal* function. Utilizing this correlation structure, the proposed algorithm intelligently narrows the search space to identify the optimal beam. Secondly, incorporating the prior knowledge on the channel fluctuation to appropriately accommodate reward uncertainty, the proposed algorithm avoids excessive exploration and further accelerates the BA process. Theoretical analysis shows that the regret of HBA is *bounded* and thus the proposed algorithm is asymptotically optimal. Extensive simulation results demonstrate that HBA can identify the optimal beam with a high probability and reduce the number of beam measurements in the multipath channel, even with coarse prior knowledge. Particularly, the proposed algorithm reduces the BA latency by orders of magnitude as compared to the BA method in IEEE 802.11ad.

Our contributions in this paper are summarized as follows.

- We formulate the BA problem as a stochastic MAB problem, in which the objective is to sequentially select beams to maximize cumulative RSS within a certain period;
- We prove that the mean RSS function over the beam space follows a multimodality structure in the multipath channel, which characterizes the correlation structure among nearby beams;
- We propose a fast BA algorithm to accelerate beam search by exploiting the correlation structure and the prior knowledge on the channel fluctuation;
- We derive a sublinear analytical upper bound on the cumulative regret, i.e., $O(\sqrt{T \log T})$, indicating the proposed algorithm is asymptotically optimal.

The remainder of this paper is organized as follows. Section II reviews related works. The system model and problem formulation are presented in Section III. Section IV proposes a fast BA algorithm. Section V analyzes the regret performance of the proposed algorithm. Simulation results are given in Section VI. Finally, Section VII concludes this paper.

II. Related Work

The BA problem in mmwave systems garners much attention recently. Zhou *et al.* elaborated the challenges of the random access protocol in the BA process in dense networks [1]. In addition, the authors developed possible solutions from the MAC perspective. Utilizing the sparse characteristic that only a few paths exist in the mmwave channel, a compressed sensing solution can align beams with a low beam measurement complexity of $O(L \log N)$, where L is the number of channel paths and N is the number of beams [8]. The approach suits for mmwave systems where the accurate phase information is available. In another line of research, Wang *et al.* developed a fast-discovery multi-resolution beam search in [10], which probes the wide beam first and continues to narrow beams until identifying the best beam. While feasible, the method needs to adjust the beam resolution at every step. On the other hand, Xiao *et al.* proposed a hierarchical codebook search method to efficiently identify the optimal beam by jointly utilizing sub-array and deactivation techniques [11]. Moreover, they provide the closed-form expression of the hierarchical codebook. Sun *et al.* further developed an orthogonal pilot based low-overhead BA method for the multiuser mmwave systems [12]. Another solution exploits some out-of-band information, i.e., the Wi-Fi signal, to identify the optimal beam [9]. Similar works extract spatial information from sub-6 GHz signals to assist BA as well as boost throughput [13], [14]. Recent efforts leverage the multi-armed beams capability to improve BA performance. Hassanieh *et al.* proposed a fast BA protocol through scanning multiple directions simultaneously [7]. A similar method, which treats the problem of identifying the optimal beam as that of locating the error in linear block codes, is developed to reduce BA complexity [15]. The works in [1], [7]–[15] provide possible solutions for the BA problem in various scenarios. Different from prior works, our work considers the correlation structure among nearby beams to assist BA process.

MAB theory has been widely applied in wireless networks, such as power allocation in small base stations [16] [17], content placement in edge caching [18], [19], task assignment in mobile crowdsourcing [20] and mobility management in mobile edge computing [21]. Very recently, the BA problem is studied based on MAB theory, which makes online decision to strike the balance between *exploitation* and *exploration*. Gulati *et al.* applied the celebrated upper confidence bound (UCB) algorithm in beam selection in traditional MIMO systems [22]. Sim *et al.* developed an online beam selection algorithm in mmwave vehicular networks based on contextual bandit theory [23]. This work learns information from real-time environment to enhance the throughput of mmwave networks. A pioneering work in [6] exploits a unimodal structure among beams to accelerate the BA process in static environments. This solution focuses on aligning beams in the single-path channel. Another work developed a distributed BA search method based on adversarial bandit theory [24]. These works provide highly

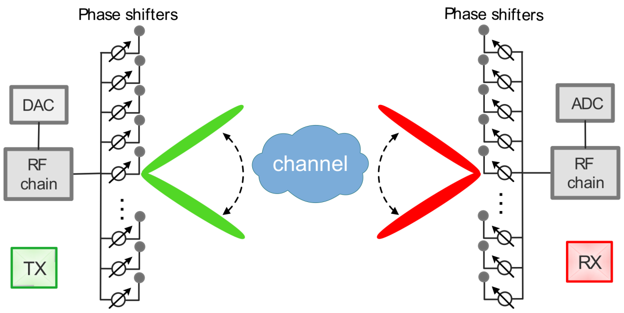


Fig. 2. The point-to-point mmwave system.

relevant insights on the BA problem in mmwave networks via bandit learning theory. However, they do not provide a method to quickly and accurately align beams, especially in complicated multipath channels. Different from existing works, we focus on leveraging the correlation structure and prior knowledge to accelerate the BA process in the multipath channel with only RSS.

III. System Model and Problem Formulation

A. Beam Alignment Model

As shown in Fig. 2, we consider a point-to-point mmwave system in a static environment, where the transmitter is equipped with N antennas. Uniform linear arrays are assumed in both the transmitter and receiver, and each antenna element is connected to a phase shifter to form narrow directional beams [25]. In the BA process, the receiver keeps quasi-omnidirectional while the transmitter scans the beam space to identify the best one. We consider the sparse clustered channel model, i.e., Saleh-Valenzuela model [26]. Suppose that the channel consists of L paths: one dominant line-of-sight (LOS) path and $L-1$ non-line-of-sight (NLOS) paths, due to strong reflections from the ground or side walls. The channel array response between the transmitter and receiver can be represented as a mixture of sinusoids,

$$h_n = g_0 e^{j\frac{2\pi d}{\lambda} n \vartheta_0} + \sum_{l=1}^{L-1} g_l e^{j\frac{2\pi d}{\lambda} n \vartheta_l} \tag{1}$$

where $0 \le n \le N-1$. Let d and λ denote the array element spacing and carrier wavelength, respectively. Typically, $d = \lambda/2$. Let g_0 and g_l represent the channel gains of the LOS path and the l-th NLOS path, respectively. Note that the channel gain of the NLOS path is around 10 dB weaker than that of the LOS path [27]. Let θ denote the physical angle of the channel. The corresponding spatial angle of the channel is denoted by $\vartheta = \cos\theta$. We vectorize the sinusoids $e^{j2\pi d n \vartheta/\lambda}, 0 \le n \le N-1$ into a vector $\mathbf{x}(\vartheta) \in \mathbb{C}^{N\times 1}$. Thus, the channel vector is given by

$$\mathbf{h} = g_0 \mathbf{x}(\vartheta_0) + \sum_{l=1}^{L-1} g_l \mathbf{x}(\vartheta_l) \in \mathbb{C}^{N\times 1}. \tag{2}$$

Since we consider a static environment, the channel vector keeps invariant during the BA process.

Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N] \in \mathbb{C}^{N\times N}$ denote the unitary discrete Fourier transform (DFT) matrix whose columns constitute the transmit beam space, given by

$$\mathbf{W} = \frac{1}{\sqrt{N}} [\mathbf{x}(\omega_1), \mathbf{x}(\omega_2), ..., \mathbf{x}(\omega_N)]. \tag{3}$$

In (3), $\omega_i = \frac{2i-N}{N}$ represents the spatial angle of the i-th beam [8]. According to the BA method in IEEE 802.11ad, the transmitter scans all the beams in $\mathbf{W}$, while the receiver beam keeps omni-directional. The received signal vector is given by

$$\mathbf{y} = \sqrt{P}\mathbf{h}^H \mathbf{W} + \mathbf{n} \tag{4}$$

where $\mathbf{n}$ denotes the additive Gaussian white noise vector. Let $N_o W$ denote the mean noise power, where W is the channel bandwidth and N_o is the noise power density.

The problem of identifying the optimal transmit beam boils down to identifying the element with the maximum magnitude within $\mathbf{y}$. Hence, to identify the optimal beam, the BA method in IEEE 802.11ad protocol needs to measure the RSS of all the transmit beams, leading to a high beam measurement complexity [7]. Searching the entire beam space incurs significant BA latency, especially when the beam space is large.

B. Problem Formulation

In this subsection, the BA problem is formulated as a stochastic MAB problem for *stationary* environments. Consider a time slotted system with T time slots of equal duration. In time slot $t \in \{1, 2, ..., T\}$, the transmitter selects a beam to transmit data. Let $\mathcal{B} = \{b_1, b_2, ..., b_N\}$ denote the set of candidate beams, which can be considered as *arms* in the bandit theory. At the beginning of time slot t, the transmitter selects a beam denoted by $b^t \in \mathcal{B}$. At the end of time slot t, the transmitter observes noisy RSS from the receiver, i.e., $r(b^t)$, which is considered as a *reward*. Rigorously, the reward is a random variable due to the channel fluctuation, such as shadow fading and the disturbance effect. For simplicity, we assume that the reward follows a Gaussian distribution with a variance σ^2. In other words, σ^2 also represents the variance of the channel fluctuation, which is utilized as *prior knowledge* in the following algorithm design. Note that the proposed algorithm can also be applied to non-Gaussian distribution settings, as validated in Section VI.

Let $b^{1:t} = \{b^1, b^2, ..., b^t\}$ denote the sequentially selected beams up to time slot t. The set of corresponding sequential rewards is represented by $r^{1:t} = \{r(b^1), r(b^2), ..., r(b^t)\}$. In the MAB setting, a sequential beam selection *policy* is how the transmitter selects the next beam based on previously selected beams $b^{1:t}$ and observed rewards $r^{1:t}$. Let Π be the set of all possible sequential beam selection policies. Our objective is to find a policy, $\pi \in \Pi$, that maximizes the expected cumulative reward (RSS) within a given time horizon of T slots, i.e., $\sum_{t=1}^{T} r(b^t)$. This objective conforms our target since a fast BA algorithm is to identify the optimal beam with the minimum latency.

In the MAB theory, *expected cumulative regret* is commonly adopted to evaluate the performance of a given policy, which denotes the expected cumulative difference between the reward of the selected beam and the maximum reward achieved by the optimal beam. The *expected cumulative regret* is defined as

$$\begin{aligned} R^{\pi}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\left(r(b^{\star}) - r(b^{t})\right)\right] \\ &= T \cdot \mathbb{E}\left[r\left(b^{\star}\right)\right] - \sum_{b_i \in \mathcal{B}} N_{b_i}^{\pi}(T)\mathbb{E}\left[r\left(b_i\right)\right] \end{aligned} \tag{5}$$

where $b^{\star}$ represents the optimal beam and $N_{b_i}^{\pi}(T)$ denotes the number of times that b_i has been selected up to time slot T. Hence, maximizing the cumulative reward is equivalent to minimizing the *expected cumulative regret* within T [6], which can be expressed as

$$\begin{aligned} \mathcal{P}1: \min_{\pi \in \Pi} \quad & R^{\pi}(T) \\ \text{s.t.} \quad & \sum_{b_i \in \mathcal{B}} N_{b_i}^{\pi}(T) \leq T \qquad \text{(6a)} \\ & N_{b_i}^{\pi}(T) \in \mathbb{Z}, \forall b_i \in \mathcal{B}. \qquad \text{(6b)} \end{aligned}$$

The preceding MAB problem $\mathcal{P}1$ can be solved by the celebrated UCB algorithm [28]. However, this problem has two characteristics that were not utilized in the UCB algorithm. Firstly, since the RSS of nearby beams is highly correlated, the correlation information from nearby beams can be utilized to select the next beam efficiently. Secondly, the prior knowledge on the channel fluctuation reflects the information of environment, which can be exploited to appropriately accommodate reward uncertainty such that the BA process can be further accelerated. In the following, we will leverage these two characteristics to accelerate the convergence speed, and hence reduce BA latency.

IV. Fast Beam Alignment

In this section, we first analyze and validate that the mean reward (RSS) over the beam space follows a multimodality structure, which characterizes the inherent correlation among beams. Next, by exploiting the correlation structure and the prior knowledge, a fast BA algorithm is proposed to identify the optimal beam.

A. Correlation Structure

Consider a *cyclic* undirected graph $G = (\mathcal{B}, E)$ whose vertices $\mathcal{B}$ stand for the beams. Let $(b_i, b_{i+1}) \in E$ denote the edge that connects neighboring beams b_i and b_{i+1}. In addition, $(b_N, b_1) \in E$ indicates that the last beam b_N and the first beam b_1 are neighbors since their beam orientations are close to each other. The unimodality structure is defined as follows.

***Definition** 1:* **(Unimodality)** Let $b_{i^{\star}}$ denote the optimal beam in G. The *unimodality* structure indicates that, $\forall b_i \in \mathcal{B}$, there exists a path, $(b_i, b_{i+1}, ..., b_{i^{\star}})$, along which the mean reward is strictly increasing.

In other words, the unimodality structure means that there is no local optimal beam over the beam space. Next, we aim to show that the correlation structure among beams follows above unimodality structure. Consider the single-path channel, where g and ϑ represent the channel gain and channel spatial angle of the path, respectively. With (4), the mean RSS is given by

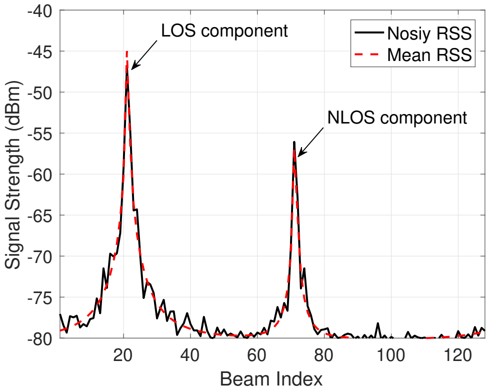


Fig. 3. The RSS function over the beam space in a two-path channel with 128 beams. The peak caused by the LOS link is around 10 dB higher than that by the NLOS link.

$$\begin{aligned} \mathbb{E}\left[r(b_i)\right] &= P\left|\mathbf{h}^{H}\mathbf{w}_i\right|^2 + N_o W \\ &= \frac{Pg^2}{N}\left|\mathbf{x}^{H}(\vartheta)\mathbf{x}(\omega_i)\right|^2 + N_o W \\ &= \frac{Pg^2}{N}\left|\sum_{n=0}^{N-1} e^{j\frac{2\pi d}{\lambda}n(\omega_i - \vartheta)}\right|^2 + N_o W \\ &= \frac{Pg^2}{N} D\left(\omega_i - \vartheta\right) + N_o W, \forall b_i \in \mathcal{B} \end{aligned} \tag{7}$$

where

$$D(x) = \frac{\sin^2(N\pi dx/\lambda)}{\sin^2(\pi dx/\lambda)} \tag{8}$$

denotes the antenna directivity function, which depends on the angular misalignment x. Hence, the mean RSS is a function of angular misalignment $\omega_i - \vartheta$.

***Theorem** 1:* In the single-path channel, the mean reward (RSS) over the beam space is a unimodal function.

***Proof** 1:* Proof is provided in Appendix A.

The linear combination of several unimodal functions is a *multimodal* function, which means that there exist several local optimums.

***Corollary** 1:* In the multipath channel, the mean reward (RSS) over the beam space is a multimodal function. The dominant peak of the multimodal function is caused by the LOS path, while other peaks are caused by NLOS paths.

***Proof** 2:* Proof is provided in Appendix B.

For example, Fig. 3 shows the RSS function over the beam space in a two-path channel. Even though the practical RSS is noisy due to the channel fluctuation, we observe that the mean RSS function follows the multimodality structure. For a two-path mmwave channel, there exist two peaks in the mean RSS function, where the dominant peak is due to the LOS path and another smaller peak is due to the NLOS path. Furthermore, the multimodality structure has been observed in many in-field measurements in mmwave systems, which further validates our theoretical results.

Remark 1: Theoretical analysis indicates that the RSS depends on the angular misalignment. As the angular misalignment of nearby beams is close, the RSS of nearby beams is similar such that nearby beams are highly correlated.

Since the RSS function over the beam space is a multimodal function, the BA problem boils down to identifying the optimal point of a multimodal function. In other words, our goal is to find optimal point $x^\star$ that maximizes multimodal reward function $f(x), x \in \mathcal{X}$. To solve this problem efficiently, the correlation structure of the reward function is leveraged. Specifically, the correlation structure is exploited based on a *dissimilarity function* that captures the smoothness of reward function [29].

***Definition** 2:* **Dissimilarity**. For space $\mathcal{X}$, a dissimilarity function for $x_1 \in \mathcal{X}$ and $x_2 \in \mathcal{X}$ is defined as $q(x_1, x_2) = w\|x_1 - x_2\|^\beta$, where $w > 0$, $\beta > 0$ and $\| \cdot \|$ denotes the Euclidean norm function. Note that $q(x, x) = 0$ for $x \in \mathcal{X}$.

The dissimilarity function is applied to characterize the discrepancy of two points in the reward function. Normally, two nearby points in the function have similar rewards, which means the dissimilarity between two nearby points is bounded. Such smoothness property of the reward function is exploited in the following algorithm design to accelerate the BA process.

B. Prior Knowledge

In addition to the aforementioned correlation structure, some prior knowledge can be leveraged to further speed up the BA process. As the reward is impacted by wireless environments, channel fluctuation statistics reflects the underlying information of the wireless environments. Leveraging the channel fluctuation statistics can appropriately accommodate the reward uncertainty such that less exploration is required. Specifically, the variance of the channel fluctuation σ^2 is assumed to be known *a priori* to accelerate the BA process. In practice, the prior knowledge can be obtained in the system initialization phase before the BA process is invoked. Practical mmwave systems also collect the variance of channel fluctuation periodically. Besides, since the channel statistical information changes slowly in static environments, there is no need to frequently collect the information. It is worth noting that the proposed algorithm works even with coarse prior knowledge at the expense of slower convergence or lower beam detection accuracy, which is presented in Section VI.

C. Hierarchical Beam Alignment (HBA) Algorithm

As discussed, the mean reward function exhibits the multimodality structure, and hence we adapt and extend the hierarchical optimistic optimization (HOO) algorithm [29] to the BA problem. Due to the lack of prior knowledge, HOO adopts a large confidence margin to accommodate the reward uncertainty, which results in slow convergence. Similar to the well-known Bayesian principles in [30], we leverage the prior knowledge to obtain an appropriate confidence margin, which avoids unnecessary exploration and further accelerates convergence. The proposed HBA algorithm is sketched in Algorithm 1. In the algorithm, $Ber(0.5)$ represents a Bernoulli distributed random variable with a parameter of 0.5, which means that the random variable is equally likely to take values 0 and 1. In addition, $leaf(\mathcal{T})$ represents the leaf node of tree $\mathcal{T}$.

The proposed algorithm is designed based on the correlation structure among beams. If a beam performs well, its nearby beams are highly likely to perform well too. Hence, the core idea is to explore intensively around good beams while loosely in others. For this purpose, a search tree is constructed, whose nodes are associated with search regions. A deeper node represents a smaller search region, as an illustrative example shown in Fig. 4(a). The algorithm operates in discrete time slots, and the binary tree is constructed in an incremental manner. At each time slot, a new node is selected by a node selection process and added to the search tree. Once selected, the beam located in the selected node is measured, and then the corresponding reward is observed. Then, the attributes of the search tree are updated based on the newly observed reward. In this way, the algorithm intelligently narrows the search region until the optimal beam is identified. It is worth noting that selecting a new node means exploring the region associated to the node, and the search tree explores the region based on previously selected beams and observed rewards.

Next, we elaborate the algorithm in detail. In the initialization phase, the beam space, $\mathcal{B}$, is mapped to a region $\mathcal{X} = [0, 1]$, which is uniformly partitioned by each beam. Similarly, the RSS function, $r(b_i), \forall b_i \in \mathcal{B}$, is mapped to a normalized reward function, $f(x), \forall x \in \mathcal{X}$, within $[0, 1]$. In the beginning, the search tree $\mathcal{T}$ only contains a root node $(0, 1)$. The node in the tree is represented by (h, j), where h denotes the depth from the root node and $j, 1 \leq j \leq 2^h$ denotes the index at depth h. In addition, each node in the tree is associated with a region. Let $C_{h,j}$ represent the region of (h, j). Specifically, the root node represents the entire region, i.e., $C_{0,1} = [0, 1]$. Let $(h + 1, 2j - 1)$ and $(h + 1, 2j)$ denote the left and the right child node of (h, j), respectively. Two child nodes partition the region of their parent node. Consider $C_{h,j} = [x_L, x_H]$, the left child node is associated with a region $C_{h+1,2j-1} = [x_L, x_a]$ and the right child node is associated with a region $C_{h+1,2j} = [x_a, x_H]$, where $x_a = x_L + (x_H - x_L)/2$ is the middle point of $C_{h,j}$. The HBA algorithm operates in a “zooming” manner, which intelligently narrows the search region via comparing the Q-values in the tree. The Q-value is designed based on the correlation structure of the reward function and the prior knowledge. At time slot t, HBA consists of the following three phases:

1. *New node selection*. In this phase, a new node will be selected. Let $\mathcal{T}_t$ denote the tree at time t. At each time slot, starting from the root node, the Q-values of two child nodes are compared until a new node $(H_t, J_t) \notin \mathcal{T}_t$ is selected. Specifically, traversing the tree, the child with a higher Q-value is chosen, otherwise breaking ties randomly (lines 5-6). The selected node is added to the tree, i.e., $\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{(H_t, J_t)\}$, and the path from the root node to the selected node is stored in $\mathcal{P}$.

2. *Attributes update*. In this phase, the attributes of all the nodes in the tree are updated. For the selected node in the previous phase, a beam located in the center of C_{H_t,J_t} is measured and then the corresponding reward r_t is obtained. Based on the newly observed reward, for node (h, j), $Q_{h,j}$ is

Algorithm 1: HBA algorithm

```
   Input: $\zeta$, $\rho_1$, $\gamma$ and $\sigma^2$
   Output: $b^\star$
1  Initialization: Set $\mathcal{T} = \{(0,1)\}$, $Q_{2,1} = Q_{2,2} = +\infty$,
     $x_L = 0$ and $x_H = 1$;
2  for t=1,2,3... do
3      $(h,j) \leftarrow (0,1)$, $\mathcal{P} \leftarrow \{(h,j)\}$;
4      ▷ New node selection
5      while $(h,i) \in \mathcal{T}_t$ do
               if $Q_{h+1,2j-1}(t) > Q_{h+1,2j}(t)$ then
                   $(h,j) \leftarrow (h+1, 2j-1)$, update $x_L = x_a$;
               else if $Q_{h+1,2j-1}(t) < Q_{h+1,2j}(t)$ then
                   $(h,j) \leftarrow (h+1, 2j)$, update $x_H = x_a$;
               else
                   $(h,j) \leftarrow (h+1, 2j - Ber(0.5))$, update the
                   search region;
               end if
           $\mathcal{P} \leftarrow \mathcal{P} \cup \{(h,j)\}$;
6      end
7      $(H_t, J_t) \leftarrow (h,j)$; $\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{(H_t, J_t)\}$;
8      ▷ Attributes update
9      Measure the beam located in the center $C_{H_t,J_t}$,
         and observe the reward $r^t$;
10     $\forall (h,j) \in \mathcal{P}$, update $N_{h,j}(t)$ and $R_{h,j}(t)$ with (9)
         and (10), respectively;
11     $\forall (h,j) \in \mathcal{T}_t$, update $E_{h,j}(t)$ with (11);
12     $Q_{H+1,2J-1}(t) = Q_{H+1,2J}(t) = +\infty$; $\hat{\mathcal{T}} = \mathcal{T}_t$;
13     for $(h,j) \in \hat{\mathcal{T}}$ do
14         $(h,j) \leftarrow leaf(\hat{\mathcal{T}})$, update $Q_{h,j}(t)$ with (12),
           $\hat{\mathcal{T}} \leftarrow \hat{\mathcal{T}} \setminus (h,j)$;
15     end
16     ▷ Terminating condition
         if $x_H - x_L < \zeta / N$ then
             Terminate beam search and select current beam $b^\star$;
         end if
17 end
```

updated by the following steps.

Firstly, as the new node is the descendant of all the nodes in path $\mathcal{P}$, $N_{h,j}(t)$, which represents the number of times that (h,j) has been selected until time slot t, is updated by

$$N_{h,j}(t) = N_{h,j}(t-1) + 1, \forall (h,j) \in \mathcal{P}. \tag{9}$$

Secondly, $R_{h,j}(t)$ represents the *mean measured reward* of (h,j) up to time slot t, which is updated by

$$R_{h,j}(t) = \frac{(N_{h,j}(t) - 1) R_{h,j}(t-1) + r^t}{N_{h,j}(t)}, \forall (h,j) \in \mathcal{P}. \tag{10}$$

Thirdly, for each node in the tree, the *initial estimated maximum mean reward* in region $C_{h,j}$, denoted by $E_{h,j}(t)$, is updated by,

$$E_{h,j}(t) = \begin{cases} R_{h,j}(t) + \sqrt{\frac{2\sigma^2 \log t}{N_{h,j}(t)}} + \rho_1 \gamma^h, & \text{if } N_{h,j}(t) > 0; \\ +\infty, & \text{otherwise,} \end{cases} \tag{11}$$

where $\sqrt{\frac{2\sigma^2 \log t}{N_{h,j}(t)}}$ is the confidence margin to accommodate for the uncertainty of rewards. As aforementioned, we adopt the Bayesian principle to design the confidence margin by leveraging the prior knowledge on the variance of channel fluctuation. In (11), $\rho_1 \gamma^h$ accounts for the maximum variation of the mean reward function in region $C_{h,j}$, where $\rho_1 > 0$ and $\gamma \in (0,1)$. This term is obtained via the correlation structure in the reward function. The maximum dissimilarity within region $C_{h,j}$ for the reward function is upper bounded by $\rho_1 \gamma^h$, i.e., $\max\limits_{x_1, x_2 \in C_{h,j}} q(x_1, x_2) \leq \rho_1 \gamma^h, \forall x_1, x_2 \in \mathcal{X}$, which holds due to the bounded diameter assumption in Section V. The values of ρ_1 and γ are selected based on extensive simulation trials. For a binary tree case, γ is typically set to 0.5 [29]. Note that E-values of all the unexplored nodes are set to infinity.

Finally, the *estimated maximum mean reward* in region $C_{h,j}$, $Q_{h,j}(t)$, should be recursively updated through the following bound

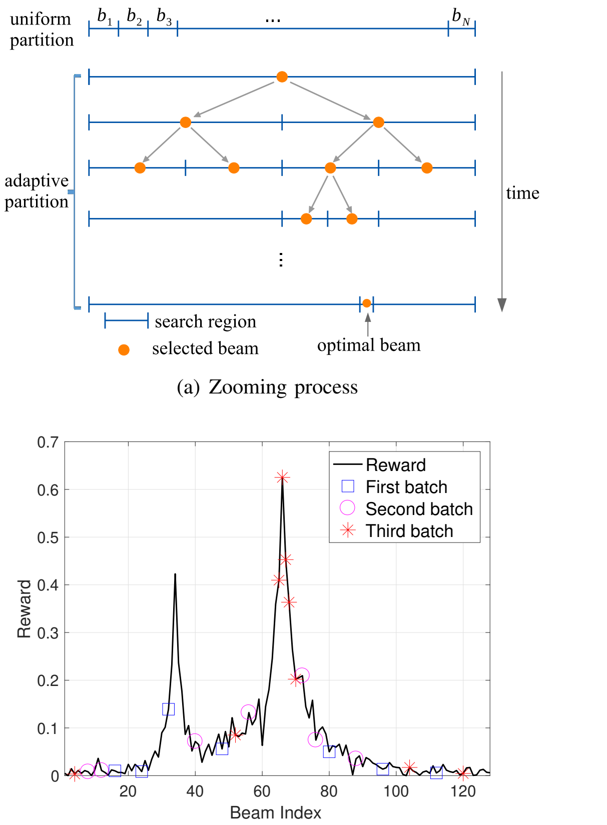
$$Q_{h,j}(t) = \begin{cases} \min\{E_{h,j}(t), \max\{Q_{h+1,2j-1}(t), Q_{h+1,2j}(t)\}\}, \\ \qquad\qquad \text{if } N_{h,j}(t) > 0; \\ +\infty, \qquad\qquad \text{otherwise.} \end{cases} \tag{12}$$

This bound depends on two terms. The first term, $E_{h,j}(t)$, is an upper bound for $Q_{h,j}(t)$ due to the definition of E-values. The second term, $\max\{Q_{h+1,2j-1}(t), Q_{h+1,2j}(t)\}$, is another valid upper bound of $Q_{h,j}(t)$. Since $C_{h,j} = C_{h+1,2j-1} \cup C_{h+1,2j-1}$, the maximum value between the Q-values in two subsets is the upper bound of Q-value in the union set. Combining both terms together, a tighter upper bound is obtained via taking the minimum value of these two bounds. Note that Q-values should be updated from the leaf node of the tree because Q-values of child nodes form the upper bound of their parent node (lines 12-15).

3. *Terminating condition.* As the tree is constructed over time, the search region gradually narrows as the depth of the tree increases. When the search region is sufficiently small, i.e., $x_H - x_L < \zeta / N$ where $0 < \zeta < 1$, the BA process is terminated and the beam located in the final region is selected as the optimal beam. The value of ζ should be carefully selected based on extensive simulation trials. Noteworthily, a larger ζ value results in faster convergence while lower beam detection accuracy.

Remark 2: A region attained a large Q-value represents that the potential maximum reward in the region is high, which means that the optimal beam (the maximum reward) locates in this region with a high probability. Hence, the HBA algorithm explores intensively in the regions with high estimated maximum rewards (Q-values) while loosely in others. In this way, the HBA algorithm is more efficient than the exhaustive search method, which accelerates the BA process.

Illustrative example: For better understanding of HBA, we provide two illustrative examples in Fig. 4. Firstly, as shown in Fig. 4(a), HBA operates similar to a "zooming" process. At the beginning, the search region is the entire region, which is uniformly partitioned by the beams. As time goes by, the search region is adaptively partitioned, and the algorithm gradually zooms to the region that contains the optimal beam.



(a) Zooming process

(b) Sequentially selected beams

Fig. 4. Illustrative examples of the HBA algorithm. (a) The proposed algorithm operates in a "zooming" manner. (b) The region that contains the dominant peak is explored intensively, while others are explored loosely.

Secondly, sequentially selected beams in the BA process are depicted in Fig. 4(b). The selected beams are divided into three batches according to the timeline. The first batch beams locate randomly in the whole region. The second batch beams get closer to the dominant peak. The last batch beams mainly focus around the optimal beam. We observe that the proposed algorithm explores intensively in the regions that contain good beams while loosely on the others.

D. Complexity Analysis

At time slot T, $\mathcal{T}_t$ contains T nodes as the tree increments by one node at each time slot. Hence, the storage complexity of the proposed algorithm is linear, i.e., $O(T)$. In addition, the attributes of all the nodes in the tree should be updated at each time slot, and hence the running time at each time slot is also linear. As the algorithm runs T time slots, the computational complexity of the HBA algorithm is a quadratic complexity $O(T^2)$. With the terminating condition, the tree is a finite tree and hence both storage complexity and computational complexity are bounded.

V. Regret Performance Analysis

In this section, we analyze the upper bound on the cumulative regret for the proposed algorithm. For the tractability of regret analysis, we have the following two assumptions.

***Assumption** 1:* **(Weak Lipschitz)** For any x around the optimal $x^\star$, there exist constants $c_H > 0$ and $\alpha > 0$ such that

$$f^\star - f(x) \leq c_H \|x^\star - x\|^\alpha \tag{13}$$

where $f^\star = f(x^\star)$ represents the optimum of function $f(\cdot)$. This assumption indicates that the reward function satisfies the week Lipschitz condition, which can avoid sharp valleys around the optimal point that induces high regret. Furthermore, the weak Lipschitz condition is mild, which only has the impact on the region in the vicinity of the optimal value. This assumption is well justified in many practical applications [17].

***Assumption** 2:*

1) **(Bounded diameter)** For a region, $C_{h,j}$, of depth h, the diameter of the region is defined as $D(C_{h,j}) = \max\limits_{x,y\in C_{h,j}} q(x,y)$. The diameter of the region is upper bounded by $\rho_1\gamma^h$ for constants $\rho_1 > 0$ and $0 < \gamma < 1$.
2) **(Well-shaped region)** For a region, $C_{h,j}$, of depth h, the region contains a ball with a radius of $\rho_2\gamma^h$ which locates in the center of $C_{h,j}$.

The bounded diameter condition is to upper bound the maximum variation of $f(x)$ within the region $C_{h,j}$. In contrast, the well-shaped region condition is to lower bound the minimum variation of $f(x)$ within the region $C_{h,j}$. Note that any region in the reward function satisfies the bounded diameter and well-shaped region conditions [29], which are utilized to bound the cumulative regret in the following analysis.

***Definition** 3:* **ϵ-optimal**. Let $f^\star_{h,j} = \max\limits_{x\in C_{h,j}} f(x)$ be the optimal reward in $C_{h,j}$. If $f^\star_{h,j} > f^\star - \epsilon_{h,j}$, $C_{h,j}$ is the $\epsilon_{h,j}$-optimal region.

For example, if $\epsilon_{h,j} = 0$, $C_{h,j}$ is the optimal region where optimal value $x^\star$ locates. Otherwise, if $\epsilon_{h,j} > 0$, $C_{h,j}$ is a sub-optimal region. Let $\epsilon_{h,j}$ represent the *suboptimality* of (h,j).

To obtain the regret bound, we first provide the following lemma.

***Lemma** 1:* For any node (h,j) whose suboptimality is larger than $\rho_1\gamma^h$, the expected number of times that (h,j) has been visited until time slot T, is upper bounded by

$$\mathbb{E}\left[N_{h,j}(T)\right] \leq \frac{8\sigma^2 \log T}{\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2} + c \tag{14}$$

where c is a constant.

***Proof** 3:* The detailed proof is given in Appendix C.

Remark 3: From Lemma 1, the number of times that a suboptimal node has been visited logarithmically increases with time, which implies the cumulative regret of the proposed algorithm is sublinear. In addition, the number of times that a suboptimal node has been visited, depends on the variance of the channel fluctuation. A larger variance of the channel fluctuation implies a more noisy wireless environment, which yields more exploration efforts to remove the reward uncertainty.

Based on above lemma, an upper bound is obtained in the following.

***Theorem** 2:* The upper bound on the cumulative regret of HBA is

$$R^\pi(T) = O\left(\sqrt{T\log T}\right). \tag{15}$$

Table I
SIMULATION PARAMETERS.

Parameter	Value
Noise spectrum density (N_o)	-174 dBm/Hz
System bandwidth (W)	2.16 GHz
Carrier frequency (f)	60 GHz
Path loss exponent (ξ)	1.74
Shadowing fading variance (σ)	2 dB
Signal range	$[-80, -20]$ dBm
SSW frame duration (T_{SSW})	15.8 *us*
Beacon interval duration (T_{BI})	100 *ms*
Number of beams (N)	{8-512}
EIRP (P_e)	50 dBm
Number of paths (L)	{1-5}
Algorithm parameters (ρ_1, γ)	$(3, 0.5)$
Terminating condition threshold (ζ)	0.1
Time horizon (T)	1000 time slots
Extra NLOS path loss	$U(7, 13)$ dB
Transmission distance (d)	20 m

Proof 4: The detailed proof is given in Appendix D.

Remark 4: Theorem 2 indicates the expected cumulative regret of HBA is sublinear in the time horizon T, i.e., $\lim_{T\to\infty} R^{\pi}(T)/T = 0$. Since the per-slot regret diminishes over time, the proposed algorithm is asymptotically optimal. Hence, the proposed algorithm converges to the optimal beam over time. Moreover, for finite time horizon T, the regret bound characterizes the convergence speed of the proposed algorithm.

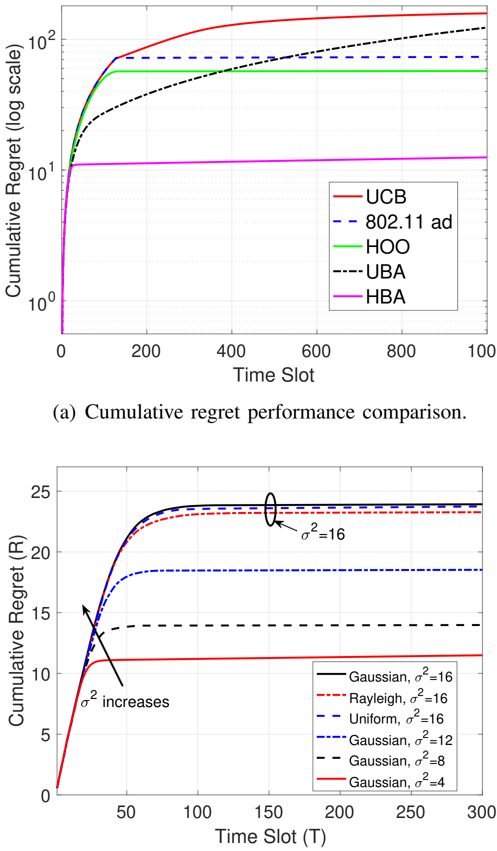
VI. SIMULATION RESULTS

A. Simulation Setup

We simulate an IEEE 802.11ad system, operating at 60 GHz with a bandwidth of 2.16 GHz [31]. Consider an outdoor scenario, such as university campus, where the transmission distance between the transmitter and the receiver is set to 20 m unless otherwise specified. The average effective isotropically radiated power (EIRP) P_e is fixed at 50 dBm[1], which is consistent with FCC regulations for 60 GHz unlicensed bands [32], [33]. Taking the directional antenna gain into consideration, the transmit power is $P = P_e - 10\log_{10} N$. For instance, the transmit powers are set to around 32 dBm and 23 dBm for 64 and 512 antenna arrays, respectively. It is worth noting that the mmwave channel is sparse, and hence we set the maximum number of channel paths to 5, which consists of one dominant LOS path and four NLOS paths. For the LOS path, the path loss is modeled as

$$PL(dB) = 32.5 + 20\log_{10}(f) + 10\xi\log_{10}(d) + \chi \qquad (16)$$

where f, ξ, d, and χ represent the carrier frequency, path loss exponent, transmission distance, and shadow fading, respectively. The shadow fading follows $N(0, \sigma^2)$ where σ is set to 2 dB [34]. Note that the channel fluctuation in the simulation is mainly caused by the shadow fading. In addition, according to practical in-field measurements, NLOS paths suffer around 10 dB more path loss than the LOS path [27]. We assume that the extra NLOS path loss follows a uniform distribution within $[7, 13]$ dB. Furthermore, for the HBA algorithm, the RSS within $[-80, -20]$ dBm is mapped to a reward within $[0, 1]$. The algorithm parameters, ρ_1, γ, and ζ are set to 3, 0.5, and 0.1, respectively, based on extensive simulation trials. Important simulation parameters are listed in Table I. We evaluate the performance via Monte-Carlo simulations. Simulation results are averaged based on 50000 samples with different channel fading and locations. The proposed HBA algorithm is compared to the following benchmarks:



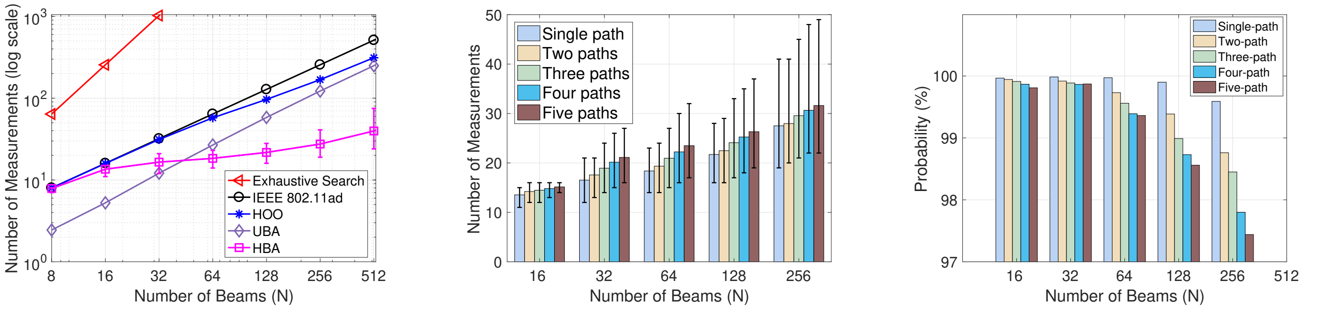
(a) Cumulative regret performance comparison.

(b) Impact of the channel fluctuation distribution and variance.

Fig. 5. Cumulative regret performance in the multipath channel.

- **IEEE 802.11ad** [2]: In this industrial method, one side (transmitter or receiver) scans the beam space, while the other side keeps omni-directional.
- **UCB** [28]: The celebrated algorithm selects the beam without exploiting both correlation structure and prior knowledge. The confidence margin is $\eta_u\sqrt{2\log t/N_{b_i}(t)}$, where the learning rate η_u is set to 0.2 based on extensive simulation trials.
- **Unimodal beam alignment (UBA)** [6]: The algorithm exploits the unimodal structure among beams to perform BA. Hence, it works in a "hill-climbing" manner, which selects the best beam among the neighboring beams at each time slot.
- **HOO** [29]: The algorithm selects the beam by exploiting beam correlation, without the prior knowledge. The confidence margin is $\eta_h\sqrt{2\log t/N_{h,j}(t)} + c_1\gamma^h$. Here, the learning rate η_h is set to 0.1, which is chosen based on

[1] For outdoor applications with the high antenna gain, the average EIRP limit is up to 82 dBm [32].



(a) Number of beam measurements in the single-path channel

(b) Number of beam measurements in the multi-path channel

(c) Beam detection accuracy in the multipath channel

Fig. 6. Performance comparison with respect to the number of paths. Error bars show the 90 percentile performance.

extensive simulation trials.

B. Regret Performance

Figure 5(a) shows the cumulative regret performance in two-path channels. Several important observations can be obtained from simulation results. First of all, HBA significantly outperforms other benchmarks. A "bounded regret" behavior is observed, which complies with the theoretical result in Theorem 2. In addition, HBA converges much faster than other benchmarks. Specifically, HBA only takes around 25 time slots to converge to the optimal beam. This is because HBA exploits both correlation structure and prior information to accelerate the BA process, while other benchmarks only exploit correlation structure or not. It is interesting to note that, as time goes by, the UBA algorithm performs even worse than the BA method in IEEE 802.11ad which does not exploit the correlation structure. The reason is that the UBA algorithm is designed based on the unimodal structure among beams, while the reward function evolves to a multimodal structure in the multipath channel. This model mismatch results in worse performance than not exploiting the correlation structure at all.

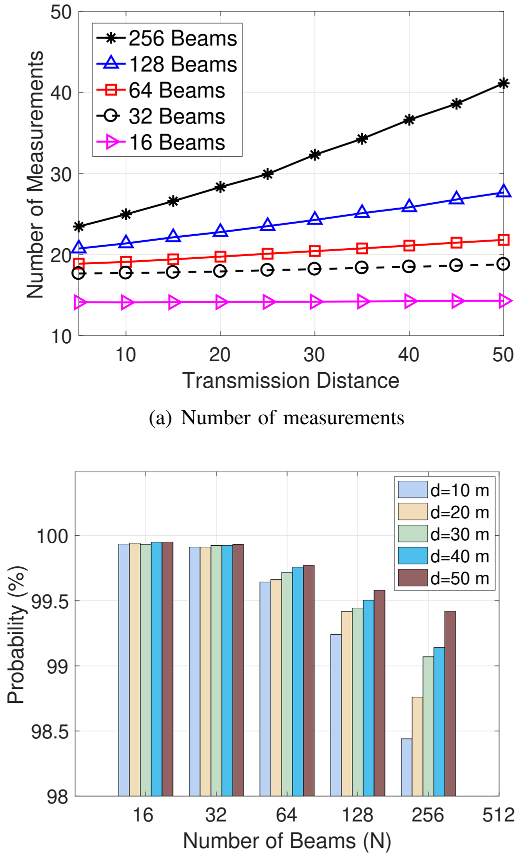
We further evaluate the impact of the channel fluctuation distribution on the regret performance in Fig. 5(b). To evaluate the dependency of the Gaussian distribution, the performance under Gaussian distribution is compared to that under two well-adopted non-Gaussian distributions, i.e., uniform distribution and Rayleigh distribution. The performance under non-Gaussian settings is very close to that under the Gaussian distribution, which means that the proposed algorithm can be applied in various settings. Furthermore, the impact of the channel fluctuation variance (σ^2) is studied in Fig. 5(b). As expected, the cumulative regret increases as the variance increases, because more exploration efforts are required in highly fluctuated channels.

C. Measurement Complexity and Beam Detection Accuracy

The regret performance only reflects the bounded fact of regret, not necessarily the actual performance. Next, we evaluate the performance of HBA using following two metrics: the number of measurements and beam detection accuracy.

We first evaluate the scalability of the proposed algorithm with the number of beams in single-path scenarios, as shown in Fig. 6(a). It is evident that the proposed algorithm significantly reduces the number of measurements as compared to the BA method in 802.11ad. For a small number ($N = 32$) of beams, the proposed algorithm reduces the number of measurements by 2 times as compared to the 802.11ad benchmark. Furthermore, the proposed algorithm achieves higher performance gains for larger numbers of beams. For instance, for a large number ($N = 512$) of beams, the proposed algorithm only needs around 40 measurements to identify the optimal beam, which reduces the number of measurements by 12 times as compared to the 802.11ad benchmark. The reason is that, different from the BA method in 802.11ad that explores all the beams, the proposed algorithm only needs to explore a few beams by leveraging the correlation structure and the prior knowledge. The results validate that the proposed algorithm is a scalable solution even with a large number of beams. In addition, we compare the HBA algorithm with the UBA algorithm. It can be seen that the UBA algorithm performs better than the HBA algorithm when the number of beams is small ($N \leq 32$). However, when the number of beams is large, HBA performs much better than UBA. Since UBA works in a "hill-climbing" manner to find the optimal beam, the number of measurements required by UBA increases with the number of beams due to a longer path to the optimal point. To avoid exceedingly high BA latency, the BA performance for a large number of beams is crucial. Thus, the proposed algorithm is more effective than the UBA algorithm when the number of beams is large. Besides, UBA does not work well in multipath scenarios, while the proposed algorithm does.

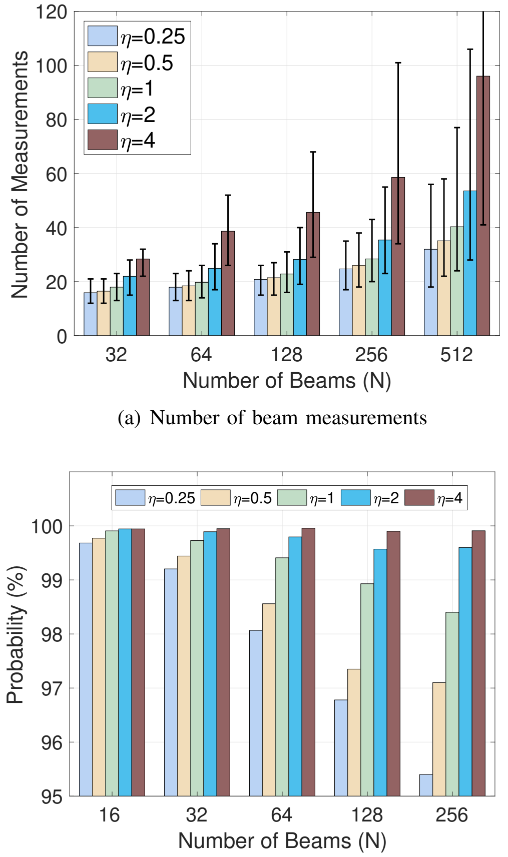
As shown in Fig. 6, we further study the performance in multipath channels. Due to the inherent sparse characteristics of the mmwave channel, the number of paths is selected from 1 to 5. Firstly, the numbers of measurements in terms of the number of paths are compared in Fig. 6(b). It can be seen that the number of measurements increases slightly as the number of paths increases. For example, for a 128-beam case, the number of measurements in the five-path channel increases by 15% as compared to that in the single-path channel. Secondly, beam detection accuracy performance is presented in Fig. 6(c). The HBA algorithm detects the optimal beam with a high probability, even in sophisticated multipath channels. Simulation results show that the beam detection accuracy is higher than 97%, even in the worst case. In addition, the



(a) Number of measurements

(b) Beam detection accuracy

Fig. 7. Performance comparison with respect to transmission distance in two-path channels.



(a) Number of beam measurements

(b) Beam detection accuracy

Fig. 8. Performance comparison with coarse prior knowledge in two-path channels.

beam detection accuracy slightly decreases as the number of paths increases. For a large number ($N = 256$) of beams, the beam detection accuracy decreases from 99.6% in the single-path channel to 97.4% in the five-path channel due to the sophisticated multipath channel.

Figure 7 shows the impact of the transmission distance on the performance. We first observe that the number of measurements increases in terms of the transmission distance, as shown in Fig. 7(a). Specifically, the number of measurements increases by 32% as the distance increases from 5 meters to 50 meters for $N = 128$. Because the RSS is weaker for a longer distance such that limited information can be extracted from nearby beams. Hence, the proposed algorithm needs to explore more beams to identify the optimal beam for remote users. Even for remote users, the proposed BA algorithm performs better than the 802.11ad benchmark. When the distance increases to 50 meters, our algorithm needs about 44 measurements for $N = 256$, which still reduces the number of measurements by 5.8 times as compared to the 802.11ad benchmark. Finally, the beam detection accuracy is presented in Fig. 7(b). Even in the low SNR case, the proposed algorithm can detect the optimal beam with a high probability.

For implementation consideration, Fig. 8 presents the performance of HBA under coarse prior knowledge conditions. The metric of the coarse prior knowledge is defined as a ratio between the estimated variance (σ_e^2) and the accurate one, i.e., $\eta = \sigma_e^2/\sigma^2$. Hence, the coarse prior knowledge can be divided into two categories: the underestimated prior knowledge when $\eta < 1$ and the overestimated prior knowledge when $\eta > 1$. We can see from Fig. 8(a) that the number of measurements increases as η increases from 0.25 to 4. Specifically, for a 256-beam case, the HBA algorithm with the overestimated prior knowledge for $\eta = 4$ requires more beam measurements as compared to that with accurate prior knowledge. Overestimating prior knowledge results in a larger confidence margin to accommodate reward uncertainty, such that more exploration efforts are needed and better beam detection accuracy can be achieved, as shown in Fig. 8(b). In contrast, when prior knowledge is underestimated, the number of measurements is slightly smaller than that with accurate prior knowledge, while the beam detection accuracy decreases due to insufficient exploration efforts. More importantly, even with the coarse prior knowledge, the proposed algorithm can substantially reduce the number of measurements as compared to benchmarks, and achieve high beam detection accuracy. For a 256-beam case, even in the worst case, the proposed algorithm reduces the number of measurements by 6 times in comparison with the BA method in 802.11ad.

D. BA Latency

Practical BA latency needs to take the 802.11ad protocol into consideration, which is different from a simple product

Table II
BA LATENCY WITH DIFFERENT NUMBERS OF BEAMS IN MULTIPATH CHANNELS.

	One user		**Four-user**	
N	**802.11ad**	**HBA**	**802.11ad**	**HBA**
16	0.51 ms	0.48 ms	1.26 ms	1.19 ms
32	1.01 ms	0.59 ms	2.53 ms	1.47 ms
64	2.02 ms	0.65 ms	103.03 ms	1.63 ms
128	4.04 ms	0.76 ms	304.04 ms	1.89 ms
256	106.07 ms	0.94 ms	706.07 ms	2.35 ms

of the number of measurements and the duration of each measurement. In the protocol, BA must be performed in the associated beamforming training (A-BFT) stage, which contains 8 A-BFT slots, and each A-BFT slot contains 16 sector sweep (SSW) frames. Each SSW frame can only provide one measurement for one beam and has a duration about 15.8 *us* [2], [35]. If the BA process cannot be finished in the A-BFT stage of the current beacon interval (BI), this BA process has to wait for the A-BFT stage in the next BI, which increases the BA latency for a whole BI duration. In the simulation, the duration of BI is set to 100 *ms* [2]. In addition, since the HBA algorithm requires the feedback of RSS of the selected beam at each round, the feedback latency should also be incorporated into the calculation of BA latency. The duration of a feedback frame at each round is about 1 *us* in 802.11ad [36]. Taking the above protocol and the feedback latency into consideration, BA latency is calculated based on the average number of measurements. Table II presents the BA latency with different numbers of beams in the two-path channel. As expected, the BA latency increases as the number of beams increases. For the case with one user, the proposed algorithm reduces the BA latency significantly as compared to the BA method in 802.11ad. In particular, for a large number ($N = 256$) of beams, the BA latency drops from 106.07 *ms* to only 0.94 *ms*. This is because the BA process with the proposed algorithm can be finished in one BI as a small number of measurements is required to identify the optimal beam. Furthermore, a larger performance gain can be observed in the four-user case. In contrast to the BA method in 802.11ad which incurs more than 700 *ms* latency for a 256-beam phase arrays, the proposed algorithm takes about 2.35 *ms*, which corresponds to two orders of magnitude gain.

VII. Conclusion

In this paper, we have investigated the BA problem in mmwave systems to find the optimal beam pair. We have developed HBA, a learning algorithm which leverages the inherent correlation structure among beams and the prior knowledge on the channel fluctuation to accelerate the BA process. The proposed HBA algorithm can identify the optimal beam with a high probability using a small number of beam measurements, even when the number of beams is large. HBA can be applied to meet the demand of delay-sensitive Gbps applications, such as cordless virtual reality gaming. Beyond the BA problem, the design principle of leveraging correlation structure is useful in other optimization problems in wireless networks, such as power allocation and interference mitigation. For our future works, it would be interesting to extend the proposed algorithm to mobile scenarios, where the environment is highly dynamic and delay requirement is more stringent. In such scenario, the main challenge lies in extracting information from the real-time environment to speed up BA.

Appendix

A. Proof of Theorem 1

According to (7), the maximum RSS can be achieved with the minimum angular misalignment denoted by, $\delta = \omega_{i^\star} - \vartheta$, where $\omega_{i^\star}$ is the spatial angle for the optimal transmit beam. Hence, $D\left(\omega_i - \vartheta\right)$ can be rewritten as

$$\begin{aligned} D\left(\omega_i - \vartheta\right) &= D\left(\delta + \frac{2(i-i^\star)}{N}\right) \\ &= \frac{\sin^2(N\pi d\delta/\lambda)}{\sin^2\left(\pi d\left(\delta + \frac{2(i-i^\star)}{N}\right)/\lambda\right)}, \forall b_i \in \mathcal{B}. \end{aligned} \tag{17}$$

From simple analysis in (17), $D\left(\omega_i - \vartheta\right)$ monotonically increases in $[i^\circ, i^\star]$ and decreases in $[i^\star, i^\star + \frac{N}{2}]$, where $i^\circ = i^\star - \frac{N}{2}$. Hence, the mean RSS function over the beam space increases along path $(b_{i^\circ}, b_{i^\circ+1}, ..., b_{i^\star})$ and decreases along path $(b_{i^\star}, b_{i^\star+1}, ..., b_{i^\circ-1})$, i.e., $r(b_{i^\circ}) < r(b_{i^\circ+1}) < ... < r(b_{i^\star}) > ... > r(b_{i^\circ-2}) > r(b_{i^\circ-1})$. With the definition of the unimodality structure, the mean RSS function is unimodal over the beam space in the single-path channel, and the theorem statement follows.

B. Proof of Corollary 1

Similar to (7), the mean RSS in the multipath channel is represented by

$$\mathbb{E}\left[r(b_i)\right] = \underbrace{\frac{Pg_0^2}{N} D\left(\omega_i - \vartheta_0\right)}_{\text{LOS component}} + \underbrace{\sum_{l=1}^{L-1} \frac{Pg_l^2}{N} D\left(\omega_i - \vartheta_l\right)}_{\text{NLOS component}} + N_o W. \tag{18}$$

Above equation indicates that the aggregated RSS consists of a LOS component and several NLOS components. For each individual path of the mmwave channel, the corresponding RSS function is unimodal function based on Theorem 1. Hence, the RSS function in the multipath channel is the aggregation of several unimodal functions, which can be considered as a multimodal function. Specifically, L paths exist in the mmwave channel, which correspond to L peaks in the multimodal function. As the channel gain of the LOS path is significantly larger than that of NLOS paths, i.e., $g_0^2 > g_l^2$. Hence, the dominant peak corresponds to the LOS path while other peaks correspond to NLOS paths. Hence, the Corollary 1 is proved.

C. Proof of Lemma 1

For any integer $m > 0$, according to the definition, the average times that node (h, j) has been visited up to time slot

T, is given by

$$\begin{aligned}
\mathbb{E}\left[N_{h,j}(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{(H_t,J_t)\in C_{h,j}}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{(H_t,J_t)\in C_{h,j}, N_{h,j}(t)\le m\}}\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{(H_t,J_t)\in C_{h,j}, N_{h,j}(t)> m\}}\right] \\
&\le m + \mathbb{E}\left[\sum_{t=m+1}^{T} \mathbb{1}_{\{(H_t,J_t)\in C_{h,j}, N_{h,j}(t)> m\}}\right] \\
&= m + \sum_{t=m+1}^{T} \mathbb{P}\left((H_t,J_t)\in C_{h,j}, N_{h,j}(t)>m\right).
\end{aligned} \tag{19}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $(H_t,J_t)\in C_{h,j}$ denotes the selected node (H_t,J_t) locates within $C_{h,j}$. The first equality is because $N_{h,j}(t)>m$ only occurs when t is larger than m.

We apply a case study to obtain an upper bound of $\mathbb{E}\left[N_{h,j}(T)\right]$. Assume node (h,j) is selected at time slot t. The path from root node $(0,1)$ to (h,j) is given by, $\mathcal{P}=\{(0,1),(1,j_1^\star),...,(k,j_k^\star),(k+1,j_{k+1}^o),...,(h,j)\}$, where k denotes the largest depth of the optimal node in the path. Before node $(k,j_k^\star)$, the optimal nodes are selected. For notation simplicity, we omit the time slot t in $Q_{k,j}(t)$. After traversing node $(k,j_k^\star)$, a sub-optimal node $(k+1,j_{k+1}^o)$ is selected instead of the optimal node $(k+1,j_{k+1}^\star)$ because the suboptimal node has a larger Q-value than the optimal node, i.e., $Q_{k+1,j^o}\ge Q_{k+1,j^\star}$. As Q-values increase along path $\mathcal{P}$, we have $Q_{k+1,j^\star}\le Q_{k+1,j_{k+1}^o}\le,...,\le Q_{h,j}$. Note that Q-values are upper bounded by E-values according to the definition, such that $Q_{k+1,j^\star}\le E_{h,j}$. Further, event $Q_{k+1,j^\star}\le E_{h,j}$ can be interpreted as the union of two events, $\{Q_{k+1,j^\star}\le f^\star\}\cup\{E_{h,j}\ge f^\star\}$. Hence, the probability that (H_t,J_t) locates within $C_{h,j}$ is upper bounded by

$$\mathbb{P}\left((H_t,J_t)\in C_{h,j}\right)\le \mathbb{P}\left(Q_{k+1,j^\star}\le f^\star\right)+\mathbb{P}\left(E_{h,j}\ge f^\star\right). \tag{20}$$

With the definition of Q-value, the Q-value of a node is the minimum value among the E-value of the node and Q-values of its child nodes. Hence, event $\{Q_{k+1,j^\star}\le f^\star\}$ can be interpreted as the union of two new events, $\{E_{k+1,j^\star}\le f^\star\}\cup\{Q_{k+2,j_{k+2}^\star}\le f^\star\}$. Since event $\{Q_{k+2,j_{k+2}^\star}\le f^\star\}$ can be further recursively expanded as $\bigcup\limits_{s=k+2}^{t-1}\{E_{s,j_s^\star}\le f^\star\}$, we have

$$\mathbb{P}\left(Q_{k+1,j^\star}\le f^\star\right)\le \sum_{s=k+1}^{t-1}\mathbb{P}\left(E_{s,j_s^\star}\le f^\star\right). \tag{21}$$

Substituting (21) and (20) into (19), (19) can be rewritten as

$$\begin{aligned}
\mathbb{E}\left[N_{h,j}(T)\right]\le m+\sum_{t=m+1}^{T}\left(\sum_{s=k+1}^{t-1}\mathbb{P}\left(E_{s,j^\star}(t)\le f^\star\right)\right. \\
\left.+\mathbb{P}\left(E_{h,j}(t)\ge f^\star, N_{h,j}(t)>m\right)\right).
\end{aligned} \tag{22}$$

The following analysis is to bound the three terms in (22) separately.

Firstly, since m is an arbitrary integer, taking m as the smallest integer that satisfies the condition $m\ge\frac{8\sigma^2\log T}{(\epsilon_{h,j}-c_1\gamma^h)^2}$. Hence m is bounded by

$$m\le\frac{8\sigma^2\log T}{\left(\epsilon_{h,j}-\rho_1\gamma^h\right)^2}+1. \tag{23}$$

Secondly, we aim to bound the first term $\mathbb{P}\left(E_{s,j^\star}\le f^\star\right)$. For the optimal nodes $(h,j^\star)$, according to the definition of E-values, $E_{h,j^\star}=\infty$ when $N_{h,j^\star}=0$. Hence, event $E_{h,j^\star}\le f^\star$ only occurs when $N_{h,j}\ge 1$. As a result, $\mathbb{P}\left(E_{h,j^\star}\le f^\star\right)$ can be rewritten as

$$\begin{aligned}
&\mathbb{P}\left(E_{h,j^\star}\le f^\star, N_{h,j}\ge 1\right) \\
&=\mathbb{P}\left(R_{h,j^\star}+\sqrt{\frac{2\sigma^2\log t}{N_{h,j^\star}}}+\rho_1\gamma^h\le f^\star, N_{h,j^\star}\ge 1\right) \\
&=\mathbb{P}\left(\left(f^\star-R_{h,j^\star}-\rho_1\gamma^h\right)N_{h,j^\star}\ge\sqrt{2\sigma^2N_{h,j^\star}\log t},\right. \\
&N_{h,j^\star}\ge 1) \\
&\overset{(a)}{=}\mathbb{P}\left(\sum_{s=1}^{t}\left(f^\star-f(X_s)+\rho_1\gamma^h\right)\mathbb{1}_{(H_t,J_t)\in C_{h,j^\star}}\right. \\
&+\sum_{s=1}^{t}\left(f(X_s)-Y_s\right)\mathbb{1}_{(H_t,J_t)\in C_{h,j^\star}}\ge\sqrt{2\sigma^2N_{h,j^\star}\log t}, \\
&N_{h,j^\star}\ge 1) \\
&\overset{(b)}{\le}\mathbb{P}\left(\sum_{s=1}^{t}\left(f(X_s)-Y_s\right)\mathbb{1}_{(H_t,J_t)\in C_{h,j^\star}}\ge\sqrt{2\sigma^2N_{h,j^\star}\log t},\right. \\
&N_{h,j^\star}\ge 1) \\
&\overset{(c)}{=}\mathbb{P}\left(\sum_{p=1}^{N_{h,j^\star}}\left(\tilde{Y}_p-f(\tilde{X}_p)\right)\ge\sqrt{2\sigma^2N_{h,j^\star}\log t}, N_{h,j^\star}\ge 1\right).
\end{aligned} \tag{24}$$

In (24), the first step follows from the definition of E-value in (11); (a) is obtained from the definition of $N_{h,j^\star}$, where $X_s,\forall s=1,2,...,t-1$ denotes the sequentially selected beams up to time $t-1$ and the corresponding reward sequence is represented by Y_s; (b) follows from the fact that $f^\star-f(X_t)-\rho_1\gamma^h<0$ holds for all the beams in the optimal region $C_{h,j^\star}$; (c) is because the definition of a new beam selection sequence $\tilde{X}_p,\forall p=1,2,3,...$ whose corresponding reward sequence is $\tilde{Y}_p$.

Let $T_p=\min\{t:N_{h,j}(t)=p\}$ represent the time sequence for the selected node in $C_{h,j}$. The sequentially selected beams can be represented by a new sequence $\tilde{X}_p=X_{T_p},\forall p=$

1, 2, 3, ..., and (24) can be further bounded by

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{p=1}^{N_{h,j_h^\star}}\left(\tilde{Y}_p - f(\tilde{X}_p)\right) \geq \sqrt{2\sigma^2 N_{h,j^\star}\log t}, N_{h,j_h^\star} \geq 1\right)\\
&\overset{(a)}{\leq} \sum_{s=1}^{t}\mathbb{P}\left(\sum_{p=1}^{s}\left(\tilde{Y}_p - f(\tilde{X}_p)\right) \geq \sqrt{2\sigma^2 s\log t}\right)\\
&\overset{(b)}{\leq} \sum_{s=1}^{t}\exp\left(-\frac{4\sigma^2 s\log t}{s\sigma^2}\right) = t^{-3}.
\end{aligned}
\tag{25}
$$

In (25), (a) can be acquired via the union bound that takes all possible values of $N_{h,j_h^\star}$; as $\tilde{D}_p = \tilde{Y}_p - f(\tilde{X}_p)$ can be considered as martingale differences, (b) is obtained via the Hoeffding-Azuma inequality [29]

$$
\mathbb{P}\left(\sum_{p=1}^{k}\tilde{D}_p \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{p=1}^{k}\sigma^2}\right). \tag{26}
$$

Thirdly, for suboptimal nodes (h, j), the upper bound of $\mathbb{P}\left(E_{h,j} \geq f^\star, N_{h,j} > m\right)$ can be obtained via a similar method of bounding $\mathbb{P}\left(E_{h,j^\star} \leq f^\star, N_{h,j} \geq 1\right)$, such that

$$
\begin{aligned}
&\mathbb{P}\left(E_{h,j} \geq f^\star, N_{h,j} > m\right)\\
&= \mathbb{P}\left(R_{h,j} + \sqrt{\frac{2\sigma^2\log t}{N_{h,j}}} + \rho_1\gamma^h \geq f_{h,j}^\star + \epsilon_{h,j}, N_{h,j} > m\right)\\
&\overset{(a)}{\leq} \mathbb{P}\left(R_{h,j} \geq f_{h,j}^\star + \frac{\epsilon_{h,j} - \rho_1\gamma^h}{2}, N_{h,j} > m\right)\\
&= \mathbb{P}\left(\left(R_{h,j} - f_{h,j}^\star\right)N_{h,j} \geq \frac{\epsilon_{h,j} - \rho_1\gamma^h}{2}N_{h,j}, N_{h,j} > m\right)\\
&= \mathbb{P}\left(\sum_{s=1}^{t}\left(Y_s - f_{h,j}^\star\right)\mathbb{1}_{(H_s,J_s)\in C_{h,j}} \geq N_{h,j}\frac{\epsilon_{h,j} - \rho_1\gamma^h}{2},\right.\\
&\left.N_{h,j} > m\right)\\
&\leq \mathbb{P}\left(\sum_{s=1}^{t}\left(Y_s - f(X_s)\right)\mathbb{1}_{(H_s,J_s)\in C_{h,j}} \geq N_{h,j}\frac{\epsilon_{h,j} - \rho_1\gamma^h}{2},\right.\\
&\left.N_{h,j} > m\right)\\
&\overset{(b)}{=} \mathbb{P}\left(\sum_{p=1}^{N_{h,j}}\left(\hat{Y}_p - f(\hat{X}_p)\right) \geq N_{h,j}\frac{\epsilon_{h,j} - \rho_1\gamma^h}{2}, N_{h,j} > m\right)
\end{aligned}
\tag{27}
$$

In (27), (a) is due to the substitution of $N_{h,j}(t) \geq \frac{8\sigma^2\log t}{(\epsilon_{h,j} - \rho_1\gamma^h)^2}$ where $m \geq \frac{8\sigma^2\log t}{(\epsilon_{h,j} - \rho_1\gamma^h)^2}$; (b) is obtained via a similar method as (24)(c), where a new beam sequence $\{\hat{X}_1, \hat{X}_2, ..., \hat{X}_p\}$ is formed to represent the sequentially selected beams in $C_{h,j}$. Next, (27) can be further bounded by

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{p=1}^{N_{h,j}}\left(\hat{Y}_p - f(\hat{X}_p)\right) \geq N_{h,j}\frac{\epsilon_{h,j} - \rho_1\gamma^h}{2}, N_{h,j} > m\right)\\
&\overset{(a)}{\leq} \sum_{k=m+1}^{t}\mathbb{P}\left(\sum_{p=1}^{k}\left(\hat{Y}_p - f(\hat{X}_p)\right) \geq \frac{k(\epsilon_{h,j} - \rho_1\gamma^h)}{2}\right)\\
&\overset{(b)}{\leq} \sum_{k=m+1}^{t}\exp\left(-\frac{k\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2}{2\sigma^2}\right)\\
&\leq t\exp\left(-\frac{m\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2}{2\sigma^2}\right)\\
&\overset{(c)}{\leq} t\exp\left(-4\log T\right) = tT^{-4}
\end{aligned}
\tag{28}
$$

In (28), (a) is due to a similar union bound in (25)(a); (b) is obtained via the Hoeffding-Azuma inequality; (c) is obtained via the substitution of $m \geq \frac{8\sigma^2\log T}{(\epsilon_{h,j} - \rho_1\gamma^h)^2}$.

Finally, substituting (23), (25) and (28) into (22), the upper bound is given by

$$
\begin{aligned}
\mathbb{E}\left[N_{h,j}(T)\right] &\leq \frac{8\sigma^2\log T}{\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2} + 1 + \sum_{t=m+1}^{T}\left(\sum_{k+1}^{t-1}t^{-3} + tT^{-4}\right)\\
&\leq \frac{8\sigma^2\log T}{\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2} + 1 + \sum_{t=1}^{T}\left(t^{-2} + T^{-3}\right)\\
&\leq \frac{8\sigma^2\log T}{\left(\epsilon_{h,j} - \rho_1\gamma^h\right)^2} + c
\end{aligned}
\tag{29}
$$

where c is a constant. The last step is because $\sum_{t=1}^{T}t^{-2}$ is bounded. Hence, Lemma 1 is proved.

D. Proof of Theorem 2

All nodes with depth h can be divided into two subsets: Φ_h that denotes the set of all the $2\rho_1\gamma^h$-optimal nodes, and Ω_h that denotes the set of nodes whose parents belong to Φ_{h-1} while itself does not belong to Φ_h. Let $H \geq 1$ be an integer whose value is determined later. With above definition, $\mathcal{T}$ can be divided into three subtrees: $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$. Let $\mathcal{T}_1$ contain Φ_H and its decedents. Let $\mathcal{T}_2$ include all the $2\rho_1\gamma^h$-optimal nodes at all the depths smaller than H, i.e., $\mathcal{T}_2 = \bigcup_{h=1}^{H-1}\Phi_h$. Let $\mathcal{T}_3$ include all the nodes in Ω_h at all the depths smaller than H, i.e., $\mathcal{T}_3 = \bigcup_{h=1}^{H}\Omega_h$. Hence the cumulative regret can be partitioned as

$$
R^\pi(T) = \mathbb{E}\left[R^\pi(\mathcal{T}_1)\right] + \mathbb{E}\left[R^\pi(\mathcal{T}_2)\right] + \mathbb{E}\left[R^\pi(\mathcal{T}_3)\right] \tag{30}
$$

where

$$
\mathbb{E}\left[R^\pi(\mathcal{T}_i)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(f^\star - f(X_t)\right)\mathbb{1}_{\{(H_t,J_t)\in\mathcal{T}_i\}}\right].
$$

Next, the regret analysis follows the idea of bounding the regret on each subtree separately.

Step 1: Bounding the regret on $\mathcal{T}_1$. As each node in Φ_H is $2\rho_1\gamma^H$-optimal, all the beams located in Φ_H are $4\rho_1\gamma^H$-optimal, i.e., $f^\star - f\left(X_t\right) \leq 4\rho_1\gamma^H$, $X_t \in \Phi_H$. In addition, it is obvious that the number of nodes in subtree $\mathcal{T}_1$ is smaller than the time horizon, i.e., $|\mathcal{T}_1| \leq T$ where $|\cdot|$ represents the cardinality operator. Therefore, the regret on $\mathcal{T}_1$ is upper bounded by

$$\mathbb{E}\left[R^\pi\left(\mathcal{T}_1\right)\right] \leq 4\rho_1\gamma^H T. \tag{31}$$

Step 2: Bounding the regret on $\mathcal{T}_2$. As $\mathcal{T}_2 = \bigcup\limits_{h=1}^{H-1} \Phi_h$ and each beam in Φ_h is $4\rho_1\gamma^h$-optimal, the regret on $\mathcal{T}_2$ can be written as $\mathbb{E}\left[R^\pi\left(\mathcal{T}_2\right)\right] \leq \sum_{h=1}^{H-1} 4\rho_1\gamma^h|\Phi_h|$. Based on the results in [29], we have $|\Phi_h| \leq c_1\left(\rho_2\gamma^h\right)^{-\kappa}$ where $\kappa = \frac{1}{\beta} - \frac{1}{\alpha}$. Specifically, α and β are give in the weak Lipschitz assumption and the dissimilarity function, respectively. The regret on $\mathcal{T}_2$ can be further bounded by

$$\begin{aligned}\mathbb{E}\left[R^\pi\left(\mathcal{T}_2\right)\right] &\leq \sum_{h=1}^{H-1} 4\rho_1\gamma^h c_1\left(\rho_2\gamma^h\right)^{-\kappa} \\ &= 4\rho_1 c_1\rho_2^{-\kappa}\sum_{h=0}^{H-1}\gamma^{h(1-\kappa)} \leq \frac{4\rho_1 c_1\rho_2^{-\kappa}}{1-\gamma^{1-\kappa}}.\end{aligned} \tag{32}$$

From (32), we can see that $\mathbb{E}\left[R^\pi\left(\mathcal{T}_2\right)\right]$ is upper bounded by a constant as $\mathcal{T}_2$ is a finite tree.

Step 3: Bounding the regret on $\mathcal{T}_3$. For each node in Ω_h, its parents should be included by Φ_{h-1}. Thus, all the beams in Ω_h are $4\rho_1\gamma^{h-1}$-optimal, and the cardinality of Ω_h is smaller than $2|\Phi_{h-1}|$. Besides, with the results in Lemma 1, $\mathbb{E}\left[N_{h,j}(t)\right] = \frac{8\sigma^2\log t}{\left(\rho_1\gamma^h\right)^2} + c$, for any $2\rho_1\gamma^{h-1}$-optimal nodes. Thus, the regret on $\mathcal{T}_3$ is given by

$$\begin{aligned}\mathbb{E}\left[R^\pi\left(\mathcal{T}_3\right)\right] &\leq \sum_{h=1}^{H} 4\rho_1\gamma^{h-1}2|\Phi_{h-1}|\mathbb{E}\left[N_{h,j}(T)\right] \\ &\leq 8\rho_1 c_1\rho_2^{-\kappa}\sum_{h=1}^{H}\gamma^{(h-1)(1-\kappa)}\left(\frac{8\sigma^2\log T}{\left(\rho_1\gamma^h\right)^2} + c\right).\end{aligned} \tag{33}$$

Finally, substituting (31), (32) and (33) into (30), we have

$$\begin{aligned}R^\pi\left(T\right) &\leq 4\rho_1\gamma^H T + \frac{4\rho_1 c_1\rho_2^{-\kappa}}{1-\gamma^{1-\kappa}} \\ &\quad + 8\rho_1 c_1\rho_2^{-\kappa}\sum_{h=1}^{H}\gamma^{(h-1)(1-\kappa)}\left(\frac{8\sigma^2\log T}{\left(\rho_1\gamma^h\right)^2} + c\right) \\ &= O\left(\gamma^H T + \log T\gamma^{-H(1+\kappa)}\right) \\ &= O\left(T^{\frac{\kappa+1}{\kappa+2}}\left(\log T\right)^{\frac{1}{\kappa+2}}\right).\end{aligned} \tag{34}$$

The last step is obtained from setting γ^H as the order of $\left(T/\log T\right)^{-1/(\kappa+2)}$ [29]. If the smoothness of the function is known, we can set $\alpha = \beta$ such that $\kappa = 0$ [29]. Hence, (34) can be rewritten as $O\left(\sqrt{T\log T}\right)$, and then the theorem is proved.

References

[1] P. Zhou, X. Fang, Y. Fang, Y. Long, R. He, and X. Han, "Enhanced random access and beam training for millimeter wave wireless local networks with high user density," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7760–7773, Dec. 2017.

[2] IEEE Standards, "IEEE standards 802.11ad-2012: Enhancement for very high throughput in the 60 GHz band," 2012.

[3] Y. Ghasempour, C. R. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 186–192, Dec. 2017.

[4] W. Wu, Q. Shen, K. Aldubaikhy, N. Cheng, N. Zhang, and X. Shen, "Enhance the edge with beamforming: Performance analysis of beamforming-enabled WLAN," in *Proc. IEEE WiOpt*, 2018.

[5] J. Qiao, Y. He, and X. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[6] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE INFOCOM*, 2018, pp. 2393–2401.

[7] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proc. ACM SIGCOMM*, 2018, pp. 432–445.

[8] Z. Marzi, D. Ramasamy, and U. Madhow, "Compressive channel estimation and tracking for large arrays in mm-Wave picocells," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 514–527, Apr. 2016.

[9] S. Sur, I. Pefkianakis, X. Zhang, and K. H. Kim, "WiFi-assisted 60 GHz wireless networks," in *Proc. ACM MOBICOM*, 2017, pp. 28–41.

[10] J. Wang, Z. Lan, C. Pyo, T. Baykas, C. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[11] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[12] X. Sun, C. Qi, and G. Y. Li, "Beam training and allocation for multiuser millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1041–1053, Feb. 2019.

[13] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[14] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 875–888, Feb. 2018.

[15] Y. Shabara, C. E. Koksal, and E. Ekici, "Linear block coding for efficient beam discovery in millimeter wave communication networks," in *Proc. IEEE INFOCOM*, 2018, pp. 2285–2293.

[16] Z. Wang and C. Shen, "Small cell transmit power assignment based on correlated bandit learning," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1030–1045, May 2017.

[17] C. Shen, R. Zhou, C. Tekin, and M. van der Schaar, "Generalized global bandit and its application in cellular coverage optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 218–232, Feb. 2018.

[18] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.

[19] S. Müller, O. Atan, M. van der Schaar, and A. Klein, "Context-aware proactive content caching with service differentiation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1024–1036, Feb. 2017.

[20] P. Yang, N. Zhang, S. Zhang, K. Yang, L. Yu, and X. Shen, "Identifying the most valuable workers in fog-assisted spatial crowdsourcing," *IEEE Internet of Things J.*, vol. 4, no. 5, pp. 1193–1203, Oct. 2017.

[21] Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017.

[22] N. Gulati and K. R. Dandekar, "Learning state selection for reconfigurable antennas: A multi-armed bandit approach," *IEEE Trans. Antennas Propag.*, vol. 62, no. 3, pp. 1027–1038, Mar. 2014.

[23] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2487–2500, Dec. 2018.

[24] I. Chafaa, E. V. Belmega, and M. Debbah, "Adversarial multi-armed bandit for mmwave beam alignment with one-bit feedback," in *Proc. ACM ValueTools*, 2019.

[25] W. Wu, Q. Shen, M. Wang, and X. Shen, "Performance analysis of IEEE 802.11.ad downlink hybrid beamforming," in *Proc. IEEE ICC*, 2017.

[26] M. R. Akdeniz, Y. Liu, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[27] A. Maltsev, R. Maslennikov, A. Sevastyanov, A. Khoryaev, and A. Lomayev, "Experimental investigations of 60 GHz WLAN systems in office environment," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1488–1499, Oct. 2009.

[28] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[29] S. Bubeck, G. Stoltz, C. Szepesvári, and R. Munos, "Online optimization in X-armed bandits," in *Proc. NIPS*, 2009.

[30] P. B. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proc. IEEE*, vol. 102, no. 4, pp. 544–571, Apr. 2014.

[31] W. Wu, N. Zhang, N. Cheng, Y. Tang, K. Aldubaikhy, and X. Shen, "Beef up mmWave dense cellular networks with D2D-assisted cooperative edge caching," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3890–3904, Apr. 2019.

[32] FCC, "Report and order and further notice of proposed rulemaking, federal communications commission," 2016.

[33] J. Du and R. A. Valenzuela, "How much spectrum is too much in millimeter wave wireless access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1444–1458, July 2017.

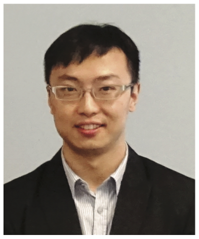
[34] 3GPP, "Technical specification group radio access network: Study on channel model for frequencies from 0.5 to 100 GHz," 2017.

[35] K. Jo, S. Park, H. Cho, J. Kim, S. Bang, and S. G. Kim, "Short SSW frame for A-BFT," *IEEE 802.11 Documents, doc.:IEEE 802.11-17/0117-00-00ay*, Jan. 2017.

[36] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "Towards scalable and ubiquitous millimeter-wave wireless networks," in *Proc. ACM MOBICOM*, 2018, pp. 257–271.

Wen Wu (S'13) earned the Ph.D. degree in Electrical and Computer Engineering from University of Waterloo, Waterloo, ON, Canada, in 2019. He received the B.E. degree in Information Engineering from South China University of Technology, Guangzhou, China, and the M.E. degree in Electrical Engineering from University of Science and Technology of China, Hefei, China, in 2012 and 2015, respectively. Starting from 2019, he works as a Postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include millimeter-wave networks and AI-empowered wireless networks.

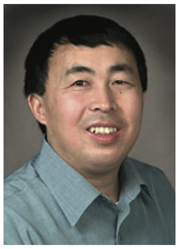
Nan Cheng (S'12-M'16) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, and B.E. degree and the M.S. degree from the Department of Electronics and Information Engineering, Tongji University. He is currently working as a joint Postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Toronto and the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include performance analysis, MAC, opportunistic communication for vehicular networks, unmanned aerial vehicles, and cellular traffic offloading.



Ning Zhang (M'15-SM'18) received the Ph.D degree from University of Waterloo, Canada, in 2015. After that, he was a postdoc research fellow at University of Waterloo and University of Toronto, Canada, respectively. Since 2017, he has been an Assistant Professor at Texas A&M University-Corpus Christi, USA. He serves as an Associate Editor of IEEE Internet of Things Journal, IEEE Transactions on Cognitive Communications and Networking, IEEE Access and IET Communications, and an Area Editor of Encyclopedia of Wireless Networks (Springer) and Cambridge Scholars. His current research interests include wireless communication and networking, mobile edge computing, machine learning and physical layer security.

Peng Yang (S'16-M'18) received his Ph.D. and B.E. degrees from School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2018 and 2013, respectively. He is a postdoctoral fellow with the BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada, where he was a visiting Ph.D. student from Sept. 2015 to Sept. 2017. His current research focuses on software defined networking and mobile edge computing.

Weihua Zhuang (M'93–SM'01–F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks. She is the recipient of 2017 Technical Recognition Award from IEEE Communications Society Ad Hoc & Sensor Networks Technical Committee, and a co-recipient of several best paper awards from IEEE conferences. Dr. Zhuang was the Editor-in-Chief of IEEE Transactions on Vehicular Technology (2007-2013), Technical Program Chair/Co-Chair of IEEE VTC Fall 2017 and Fall 2016, and the Technical Program Symposia Chair of the IEEE Globecom 2011. She is a Fellow of the IEEE, the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. Dr. Zhuang is an elected member in the Board of Governors and VP Publications of the IEEE Vehicular Technology Society



Xuemin (Sherman) Shen (M'97-SM'02-F'09) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Dr. Shen received the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, the Joseph LoCicero Award in 2015 and the Education Award in 2017 from the IEEE Communications Society. He has also received the Excellent Graduate Supervision Award in 2006 and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He served as the Technical Program Committee Chair/Co-Chair for the IEEE Globecom'16, the IEEE Infocom'14, the IEEE VTC'10 Fall, the IEEE Globecom'07, the Symposia Chair for the IEEE ICC'10, the Tutorial Chair for the IEEE VTC'11 Spring, the Chair for the IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He is the Editor-in-Chief of the IEEE INTERNET OF THINGS JOURNAL and the Vice President on Publications of the IEEE Communications Society.