Efficient Online Learning Algorithms for Joint Path and Beam Selection in Multihop Mmwave Networks

Tianchi Zhao* Chicheng Zhang† Ming Li* Jingcheng Li* Zhiwu Guo*
*Department of ECE, The University of Arizona, Tucson, AZ, 85721
†Department of Computer Science, University of Arizona,Tucson, AZ, 85721
Email: {tzhao7, chichengz, lim, jli2972, zhiwuguo}@email.arizona.edu

***Abstract*—To provide high coverage and combat high attenuation, mmWave networks typically require dense deployment of base stations, and adopt a self-backhauled network architecture where data are transmitted via multi-hop links. The unique characteristics of mmWave links (e.g., highly directional beams, sensitivity to blockage) bring challenges to designing an efficient online routing algorithm, where beam selection must be simultaneously considered. In this paper, we formulate the online joint path and beam selection (JPBS) problem for multihop mmWave networks. We exploit the Unimodal property of the mmWave channel to design a new and efficient combinatorial bandit algorithm for JPBS: Combinatorial Unimodal Lower Confidence Bound based Joint Path and Beam Selection (CULCB-JPBS). We prove a finite-time regret bound of CULCB-JPBS and show that it is independent of the number of beams in each link. Furthermore, our experimental and simulation results show that our proposed learning algorithm can significantly improve the convergence rate and yield much lower regret (thus lower end-to-end delay), compared with existing approaches.**

I. Introduction

To meet the explosive growth of wireless devices and mobile data traffic demand, the fifth-generation (5G) network aims to deliver massive connectivity and data rate, high reliability and low latency [1]. As the sub-6 GHz spectrum is becoming increasingly scarce, both academia and industry are exploring the underutilized mmWave frequency bands (30-300 GHz) [2], which promises high data rates. Due to the short wavelength, mmWave communications suffer from high attenuation and penetration loss, as well as are sensitive to blockage [3], which necessitates directional beamforming with a large number of antennas. Thus, in practice, to provide ubiquitous coverage, mmWave base station (BS) deployments typically require high density with short distances among them ($< 200m$). Since high-speed wired backhaul (e.g., optical fibre connections) may not always be available and incurs high cost, wireless self-backhauling was proposed as a promising alternative, in which the same spectrum is used for both coverage and backhaul connectivity to other BSes. This is referred to as ultra-dense self-backhauled small-cell deployments (a.k.a. integrated access and backhaul [4]–[6]), which has been adopted by the industry. In self-backhauled mmWave networks, only a fraction of BSes have fiber/wired connections and other BSes connect to them via multi-hop, short distance wireless links. This is also compliant with the cloud-RAN architecture in 5G, where the centralized unit (CU) and distributed unit (DU) are separated and the self-backhaul can be used to transmit high-speed baseband data from the CU to DU [4], [7].

Previous works have shown that self-backhauled mmWave networks can improve the coverage, throughput and transmission reliability [8]–[10]. However, using multi-hop transmissions may increase the end-to-end (E2E) delay, which plays a vital role in Ultra-reliable low-latency communication (URLLC) applications (one of the main scenarios specified by 3GPP 5G standards). Example URLLC applications include vehicle-to-everything (V2X), virtual-reality/augmented reality (VR/AR), remote collaborative surgery, networked unmanned vehicles, and etc. Thus, it is important to select the optimal path from a macro BS to the UE via multiple small-cell BSes to minimize the E2E delay, which must be updated in an online manner, due to wireless channel dynamics (e.g., natural channel fluctuations, blockage or UE mobility). On the other hand, due to the high directionality of mmWave beams, beam alignment and tracking has been a challenging problem for individual one-hop mmWave links (where most existing works focus on, e.g., [11]–[13]). Since there are typically a large number of beams in the beamforming codebook, beam training-based algorithms can incur large signaling overhead and delay. Applying them directly to a multi-hop network (by training each link's beams first and then choose the path using the best beams' costs as metric) will only exacerbate this issue by bringing even higher overhead and delay. Thus, path and beam selection must be considered jointly in an online manner.

In this work, we formulate the joint path and beam selection (JPBS) in self-backhauled mmWave networks as an online learning problem. Naively, if we apply a standard multi-arm bandit (MAB) algorithm, with each arm corresponding to (path, beam) combinations, the number of arms is exponential in the size of the network and the number of beams per link, which is not scalable in practice. To solve this problem, we propose a new combinatorial bandit algorithm by exploiting the unique property of mmWave channels. That is, the expected delay (inverse of packet delivery probability) of each link satisfies the Unimodality structure over the beam space, which we validate theoretically and experimentally. Based on this property, our proposed algorithm jointly chooses a path and beams for all the links on the path in an online manner, which enjoys fast convergence to the optimal path and beam choices, whose regret does not depend on the number of beams. Our main contributions are summarized as follows:

(1) We formulate the online joint path and beam selection problem in self-backhauled mmWave networks. We show that if the average SNR of a mmWave link has Unimodal property over the beam space, the successful packet delivery probability (and expected link delay) also has the Unimodal property. Experimental results confirm our theoretical analysis.

(2) We propose a new combinatorial Unimodal multi-armed bandit algorithm for joint path and beam selection: CULCB-JPBS. We derive an instance-dependent upper regret bound for the CULCB-JPBS, which does not depend on the number of beams for each link, in contrast to the linear dependence of regret on the number of beams for the state-of-the-art combinatorial bandit algorithm without exploiting Unimodal property. Our CULCB-JPBS is general enough so that it can be applied to other applications where Unimodality is satisfied.

(3) We carry out real-world experiments and collect data using a 28 GHz USRP-based mmWave communication platform. Using both experimental data and simulations, we validate our algorithm's efficiency and effectiveness in terms of cumulative regret, delay and number of successfully received packets. We show that CULCB-JPBS achieves much lower regret and delay, and converges faster than three baseline algorithms.

II. Related work

A. Beam Alignment for Single-hop MmWave Links

Existing beam alignment algorithms for mmWave can be divided into offline and online ones [14], [15]. An example of the former is the work of Hassanieh et al. [16], which proposed a fast mmWave beam alignment algorithm without scanning the whole search space. It can find the best beam alignment in $O(KlogN)$, where K is the number of channel paths and N is the number of beams. However, offline algorithms assume static channels and incur extra training overhead.

For the online setting, the closest work is Hashemi et al. [11], who proposed a Multi-Armed Bandit (MAB)-based algorithm for beam alignment in a mmWave link, exploiting the inherent correlation of the channel. The idea is to leverage the unimodal property of average RSS to significantly reduce the search space to the neighbors of optimal beam, by eliminating beams with worse performance. Wu et al. [12] also proposed an MAB-based algorithm that takes advantage of the correlation among beams, assuming multi-modal expected reward functions. Both works assume stochastic channels with fixed distributions. Aykin et al. [13] proposed an adaptive Thompson sampling algorithm to deal with user mobility, using a discount factor to emphasize current reward value and de-emphasize past reward value. However, the discount factor is difficult to set in practice because it depends on the velocity of the user. Note that, online algorithms can achieve data transmission and learning simultaneously.

B. Routing in Self-backhauled Mmwave Networks

Yuan et al. [17] formulated a joint path selection and scheduling problem to optimize QoS in self-backhauled mmwave networks. However, this algorithm is offline and they did not consider beam selection. Since this problem is NP-hard, they proposed an approximation algorithm with performance guarantee. Vu et al. [10] proposed a joint rate and path selection algorithm to select the best paths and allocate rates over these paths subject to latency constraints. Learning the best path is done by employing a reinforcement learning algorithm, and the rate allocation is solved by the successive convex approximation method. However, it adopts the ϵ-greedy algorithm to do exploration, which usually has a suboptimal regret guarantee. Sun et al. [18] proposed two MAB based handover mechanisms to reduce unnecessary handovers. However, their approaches cannot guarantee that the channel remains stationary within each block because of the mobility of users. To the best of our knowledge, we are the first to study the online JPBS problem in mmWave networks.

C. Unimodal Bandit

In a Unimodal bandit problem, the expected reward of arms forms a Unimodal function (a function is said to be Unimodal if it has only one peak (valley), e,g. parabola). If the Unimodal function has a peak, we will solve a maximization problem to find the largest point. Otherwise, we will solve a minimization problem to find the smallest point if the Unimodal function has a valley. Here, specialized algorithms have been designed to exploit the Unimodality structure, to achieve faster convergence rate (compared to standard bandit algorithms such as UCB and Thompson Sampling). Yu et al. [19] initiates the study of this problem, under continuous arm and discrete arm settings. Combes et al. [20] proposed Optimal Sampling for Unimodal Bandits (OSUB), and exploits the Unimodal structure under the discrete arm setting. They provided a regret upper bound for OSUB which does not depend on the number of arms. Zhang et al. [21] showed that the effective throughputs of mmWave codebooks possess the Unimodal property and proposed a Unimodal Thompson Sampling (UTS) algorithm to deal with mmWave codebook selection. Zhao et al. [22] study bandits with clustered arms, where the expected reward of each cluster has a Unimodal structure. It can be applied to multi-channel mmWave beam selection or the codebook selection problem. The main difference of our work with [22] is the system model. The above works are only applicable to the single link/one-hop setting. In contrast, our work deals with the more challenging semi-bandit routing settings, whereas bandits with clustered arm can be regarded as a special case, where each cluster can be viewed as a separate direct link from the source to destination.

D. Combinatorial Bandit

In a combinatorial bandit problem, each (combinatorial) arm is a combination of individual arms, and the reward function has a linear form. The semi-bandit feedback model assumes the availability of reward feedback from each individual arm belonging to the chosen arm. Gai et al. [23] proposed a Learning with Linear Rewards (LLR, a.k.a. combinatorial UCB, CUCB) algorithm to solve online combinatorial semi-bandit problems, with applications to maximum weighted matching, shortest path, minimum spanning tree, etc. Although LLR is

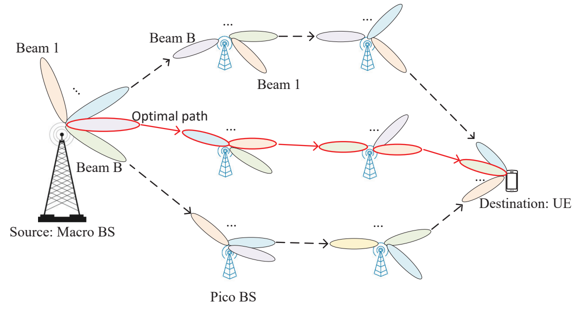


Fig. 1. The system model of 5G self-backhauled multihop mmWave networks.

applicable to our online JPBS problem, as we will see, a naive application of LLR has a regret bound that depends linearly on the number of beams per link, which can be large in practice. In contrast, our work assumes Unimodality of the cost of each link, and our algorithm exploits the Unimodality to achieve a significantly lower regret bound. He et al. [24] proposed another combinatorial bandit algorithm for online shortest path in multi-hop wireless networks using probing packets to gather feedback. The difference with our work is that it considers sub-6GHz bands. Talebi et al. [25] also formulated the shortest path routing problem as a combiatorial bandit optimization problem and proposed algorithms under different settings where routing decisions are made. Their work differs from Gai et al. [23] because their algorithms adopt the KL-UCB-style confidence bound construction instead of UCB. None of the above works exploit the Unimodal property.

III. System Model and Problem Formulation

A. System Model

We consider a 5G millimeter wave self-backhauled network which is shown in Fig. 1. We consider downlink (DL) unicast transmissions,where a macro base station (BS) with a wired backhaul is the source node, and a user equipment (UE) is the destination. Other pico BSes (aka. small cells) act as wireless backhaul, and are used as relay nodes between the source and destination to compensate the high path-loss in the mmWave band. Each node is equipped with an antenna array for analog beamforming. For the mmWave channel model, we consider a link between a transmitter (Tx) and receiver (Rx), which have N_t and N_r antennas respectively. $\boldsymbol{H}_t$ denotes the physical channel between Tx and Rx at time slot t, which is a $N_r \times N_t$ complex channel matrix. We denote the beamforming codebook for Tx as $\mathcal{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_{B_{Tx}}\}$ and the one for Rx as $\mathcal{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_{B_{Rx}}\}$, where B_{Tx} and B_{Rx} are the maximum number of narrow beams that can be generated by Tx and Rx, respectively, and $\boldsymbol{f}_m \in C^{N_t \times 1}, m = 1, 2, \cdots, B_{Tx}$, $\boldsymbol{q}_n \in C^{N_r \times 1}, n = 1, 2, \cdots, B_{Rx}$ are phase shift vectors for each beam. We denote $\mathcal{B} = \mathcal{F} \times \mathcal{Q}$, and the total number of beam vector pairs in each link as $B = |\mathcal{B}| = B_{Tx} \times B_{Rx}$. The received signal is

$$y(t) = \boldsymbol{q}_n^H \boldsymbol{H}_t \boldsymbol{f}_m s(t) + \boldsymbol{q}_n^H \boldsymbol{z}(t), \qquad (1)$$

where $s(t)$ is the transmitted signal, and $\boldsymbol{z}(t) \in C^{N_r \times 1}$ is a vector of complex white Gaussian noise. The channel $\boldsymbol{H}_t$ is time-varying, and we assume that at each time slot, it is independently drawn from a fixed but unknown distribution due to fading and possible environmental disturbances. This is called the stochastic setting which is also considered by previous works [11], [12]. Note that, to handle channel dynamics (blockage or UE mobility), we can assume that the channels have multiple states, which can be considered as stochastic in each state (e.g., different blockage states or UE locations), or piece-wise stationary [26] (will be discussed in Sec. VI).

We do not assume a specific channel model/distribution as it is unknown, but we assume that each link's average SNR or received signal strength (RSS) satisfies the Unimodal property over beam space. That is, if we fixed either the Tx's or Rx's beamforming vector, average SNR of the received signal is a Unimodal function w.r.t. different beam indices of the Rx or the Tx (respectively). In this paper, we assume that the Rx's beam vector is fixed and only the Tx can change its beam, which gives a one-dimensional (1-D) Unimodal function. We will consider the 2-D Unimodal property in our future work. Previous works provided theoretical analysis ([12], [27]) and experimental results ([18]) to show that such unimodality assumption holds for mmWave channels with a single path (or a dominant line-of-sight, LoS path). This is applicable to self-backhauled mmWave networks because they are deployed outdoors with LoS, and typically reflectors are faraway [6].

B. Problem Formulation

We represent the above mmWave self-backhauled network as a directed graph $G = \langle \mathcal{V}, \boldsymbol{E} \rangle$ with a source node s and a destination node d, where $\mathcal{V}$ is the set of all vertices and $\boldsymbol{E}$ is the set of all the directed edges. Denote $E = |\boldsymbol{E}|$. In our problem, the vertices include both BSes (source or relay nodes) and UEs (as destinations), each edge corresponds to a link $l = (i, j)$ between nodes i and j. Our goal is to choose a routing path and corresponding beams for each link on the path, to optimize certain Quality-of-Service (QoS) metric. We define (l, b) as a tuple of link $l \in \boldsymbol{E}$, and a beam (pair) $b \in \mathcal{B} = \{1, 2, \ldots, B\}$ for this link. Also, $a_{l,b}$ is an indicator variable: $a_{l,b} = 1$ means link l is selected and beam b is used for link l; otherwise, $a_{l,b} = 0$. We define $\theta_{l,b}$ as the expected cost using the b-th beam for link $l \in \boldsymbol{E}$.

The *online JPBS problem* is formulated as a combinatorial unimodal bandit. An individual arm is a tuple $(l, b), l \in \boldsymbol{E}, b \in \mathcal{B}$. A combinatorial arm $\boldsymbol{a}$ is defined as a set of chosen links and beams, represented by a binary vector $\boldsymbol{a}$ (its elements are $a_{l,b}$). The feasible set of arms is:

$$\boldsymbol{\Omega} = \{\boldsymbol{a} \in \{0,1\}^{BE} : \text{s.t. Eq.}(2-4)\};$$

$$\forall l \in \boldsymbol{E}, b \in \mathcal{B}, \quad a_{l,b} \in \{0,1\}, \tag{2}$$

$$\forall l \in \boldsymbol{E}, \sum_{b \in \mathcal{B}} a_{l,b} = 1, \tag{3}$$

$$\forall i, \sum_{l:pre(l)=i} \sum_{b \in \mathcal{B}} a_{l,b} - \sum_{l:suc(l)=i} \sum_{b \in \mathcal{B}} a_{l,b} = \begin{cases} 1 & i = s \\ -1 & i = d \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where Eq. (3) means that only one beam is chosen by every link, Eq. (4) ensures a valid loop-free path is selected; here we also define $pre(l)$ and $suc(l)$ as the transmitter and receiver vertices of link l, respectively.

At each discrete time step $t = 1,..,T$, the learner (the macro BS in our setting) chooses an arm $\boldsymbol{a}(t) \in \boldsymbol{\Omega}$, and observes instantaneous costs $X_{l,b}(t)$ for all $(l,b) \in A_{\boldsymbol{a}(t)}$, where $E[X_{l,b}(t)] = \theta_{l,b}$, and for any $\boldsymbol{a}$, $A_{\boldsymbol{a}}$ denotes the set of all (l,b) tuples such that $a_{l,b} = 1$. The instantaneous cost $X(t)$ of an arm is the summation of the instantaneous costs, $X_{l,b}(t)$, of each chosen individual arm (l,b): $X(t) = \sum_{l,b} a_{l,b}(t) X_{l,b}(t)$. Also, we define $\mu_{\boldsymbol{a}} = \sum_{l,b} a_{l,b}\theta_{l,b}$. Then, we have $E[\sum_{t=1}^{T} X(t)] = E[\sum_{t=1}^{T} \mu_{\mathbf{a}(t)}]$. Here, $X_{l,b}(t)$ are random variables.

Thus, we need to minimize the expected cumulative cost, defined as $X = E[\sum_{t=1}^{T} X(t)]$, and T is the total running time. We define the policy π as a strategy to select a sequence of arms $\boldsymbol{a}(t)$. Our objective is to sequentially choose arm $\boldsymbol{a}(t)$ at each time step to minimize the expected cumulative cost. Our online optimization problem can be formulated as

$$\textbf{OnlineOpt}: \min_{\pi} X = E^{\pi}[\sum_{t=1}^{T} \sum_{l,b} a_{l,b}(t) X_{l,b}(t)], \tag{5}$$

where π denotes a joint path-beam selection policy. The cumulative regret is defined as follows:

$$E[\text{Regret}(T)] = E[\sum_{t=1}^{T} (\mu_{\boldsymbol{a}(t)} - \mu_{\boldsymbol{a}^*})], \tag{6}$$

where $\mu_{\boldsymbol{a}^*}$ is the expected cost of the optimal arm in the offline problem. Note that, minimizing the expected cumulative costs in Eq. 5 is equivalent to minimizing the expected cumulative regret [23].

For online end-to-end (E2E) delay minimization, we set $X_{l,b}(t) = \tilde{d}_{l,b}(t)$ which is the instantaneous delay of a link-beam pair (the number of trials to transmit a packet until it is successfully delivered on a link l using beam b; $\tilde{d}_{l,b}(t)$ is a random variable). We focus on minimizing E2E delay as it plays a vital role in URLLC applications [1], [28]. In this case, $\theta_{l,b} = d_{l,b}$ where $d_{l,b}$ is the expected delay of successfully delivering a packet using the b-th beam on link $l \in \boldsymbol{E}$. Since we assume the channel is independent across time slots, the link delay (number of successive failures before success) follows a Geometric distribution with mean $d_{l,b} = 1/p_{l,b}$, where $p_{l,b}$ is the successful packet delivery probability using the b-th beam for link $l \in \boldsymbol{E}$. The formulation can also be extended to consider end-to-end reliability as the objective/constraint.

Additionally, we have the following assumption:

Assumption 1. *[Unimodality of expected link cost] The expected cost of every beam b under each link l, is a unimodal function w.r.t. the beam space $b \in \mathcal{B}$. Unimodal function means that there exists a unique valley $b^*(l)$ of $\theta_{l,1}, \ldots, \theta_{l,B}$;*

The expected link cost (delay) is the inverse of success probability. While it is a common assumption for mmWave that the link SNR/RSS has Unimodal properties w.r.t. beam space [11], [12], it is not straightforward that the success probability also has the unimodal property. We formalize this relation in the following Lemma.

Lemma 1. *For any fixed link l, if the average SNR has Unimodal property over the beam space and the SNR has same distribution shape with different means for different beams, the successful packet delivery probability and expected delay also have Unimodal properties over the beam space.*

Proof. To prove the unimodality of successful packet delivery probability, it suffices to prove the following: if $E[\gamma_{b_j}] \leq E[\gamma_{b_i}]$, then $p_{b_j}^{success} \leq p_{b_i}^{success}$, where $\gamma_{b_{i/j}}$ is the SNR for beam $b_{i/j}$, $p_{b_{i/j}}^{success}$ is the successful packet delivery probability under beam i/j, which can be derived by the outage probability. The outage probability of a beam b is defined as: $P_{out,b} = p(\gamma_b < \gamma_0) = \int_0^{\gamma_0} p_{\gamma_b}(\gamma) d\gamma$, where γ_0 is the minimum SNR required for successfully decoding a packet. Then, we have $p_b^{success} = 1 - P_{out,b}$. We denote $f_{i/j}(\gamma), F_{i/j}(\gamma)$ as the probability density function (PDF) and cumulative distribution function (CDF) for beam $b_{i/j}$, respectively. Also, we define $f_0(\gamma), F_0(\gamma)$ as the PDF/CDF which has same distribution shape with $\gamma_{b_{i/j}}$ except with a zero-mean. Then we have $P_{out,b_i} = F_i(\gamma_0) = \int_0^{\gamma_0} f_i(\tau) d\tau = \int_0^{\gamma_0} f_0(\tau - E[\gamma_i]) d\tau = \int_0^{\gamma_0 - E[\gamma_i]} f_0(\tau) d\tau = F_0(\gamma_0 - E[\gamma_i])$. Similarly, $P_{out,b_j} = F_j(\gamma_0) = F_0(\gamma_0 - E[\gamma_j])$. Note that $E[\gamma_j] \leq E[\gamma_i]$. Then $F_i(\gamma_0) \leq F_j(\gamma_0)$, and we have $p_{b_j}^{success} \leq p_{b_i}^{success}$.

Since the expected link delay is geometrically distributed with mean $1/p_b^{success}$ [25], which is a monotonically decreasing function w.r.t. $p_b^{success}$. Thus, the expected link delay also has Unimodal property over the beam space. □

Remark: Lemma 1 requires that the link SNR has the same distribution shape with different means for different beams. We now justify this condition using a generic mmWave signal propagation model adopted by previous works [18], [29]: $\gamma_j = 10\log(\frac{P^{Tx} G_j^{Rx} G_j^{Tx} PL(d)^{-1}}{P_n})$ dB, where P^{Tx} is the transmit power, P_n is noise power and G_j^{Rx} and G_j^{Tx} are the gains of the receive and transmit antenna arrays for beam j (in the directions of angle-of-arrival and angle-of-departure), respectively, d is the distance between transmitter and receiver, PL is the path loss. $PL(d) = \alpha + 10\beta \log_{10}(d) + \xi$, $\xi \sim N(0, \sigma^2)$, where α, β are the least square fittings of floating intercept and slope (respectively) over the measured distances, and ξ represents log-normal shadow fading with variance σ. Then, we have $E[\gamma_j] = 10\log(\frac{P^{Tx} G_j^{Rx} G_j^{Tx}}{P_n}) - E[PL(d)]$ dB.

TABLE I
FREQUENT NOTATIONS

$\boldsymbol{a}$	a combinatorial arm that contains individual arm (l, b)
$A_{\boldsymbol{a}}$	$\{(l, b) : l \in \boldsymbol{E}, b \in \mathcal{B}, a_{l,b} = 1\}$, the set of individual arms in $\boldsymbol{a}$
$\boldsymbol{a}^*$	the optimal arm
L	$\max_{\boldsymbol{a}\in\boldsymbol{\Omega}} \lvert\boldsymbol{a}\rvert$
$m_{l,b}(t)$	number of times that individual arm (l, b) has been selected up to round t.
$\hat{\theta}_{l,b}(t)$	empirical mean of individual arm (l, b) has been selected up to round t.
$\Delta_{\boldsymbol{a}}$	$\mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}^*}$
$T_{\boldsymbol{a}}(t)$	number of times arm $\boldsymbol{a}$ has been played in the first t time rounds.

We can see that $E[\gamma_j]$ is different under different beam j, but γ_j's PDF's shape does not change with j, as the path loss $PL(d)$ has the same distribution since the physical channel statistics does not depend on the beams.

IV. COMBINATORIAL BANDIT-BASED JOINT PATH AND BEAM SELECTION

In this section, we first introduce a naive combinatorial lower confidence bound (CLCB)-based JPBS algorithm, and then we introduce an improved CLCB algorithm, where we incorporate Unimodal bandit into CLCB, which has a much lower regret bound. We focus on the stochastic bandit setting.

A. Combinatorial LCB based Joint Path and Beam Selection

We first present a naive baseline, Combinatorial Lower Confidence Bound-based Joint Path and Beam Selection, abbreviated as CLCB-JPBS (Alg. 1). CLCB-JPBS is a direct adaptation of Gai et al's [23] LLR algorithm in our online joint path-beam selection problem; it uses the "optimism in the face of undertainty" principle: maintain lower bound estimates of the costs of individual arms (line 12, and solve an offline optimization problem in each step with the lower bound estimates using Dijkstra's algorithm (line 13). A small modification from LLR (Gai et al [23]) is that, we use a slightly tighter lower confidence bound for each individual arm's cost (we use $\sqrt{\frac{2\log(T)}{m_{l,b}(t)}}$, whereas LLR uses $\sqrt{\frac{(L+1)\log(T)}{m_{l,b}(t)}}$), which is still valid with high probability.

The CLCB-JPBS algorithm proceeds as follows: first, the algorithm initializes the maximum number of individual arms (i.e., L) in an arm (a combination of path and beams/set of link-beam pairs) (line 1). Then, it loops over all the links and beams in each link (line 3 and 4) to explore each l, b at least once (line 5) and obtain their initial empirical mean $\hat{\theta}_{l,b}$ and $m_{l,b}$. After the initialization stage, from time steps $BE + 1$, the CLCB-JPBS algorithm greedily pulls the optimum arm from its feasible set $\boldsymbol{\Omega}$ according to minimizing the optimistic lower bound on the expected cost, i.e., solving the optimization problem shown in line 13. Then, a packet is transmitted from s to d using selected arm (retransmitted in each link until the packet is received by the destination successfully). Lastly, in line 14, the empirical mean $\hat{\theta}_{l,b}$ and count $m_{l,b}$ are also updated according to the observed delay $X_{l,b}(t)$. The delay for each link and beam pair is obtained from ACK messages. We have the following regret guarantee of Alg. 1:

Theorem 1 The expected regret of Alg. 1 is:

$$E[\text{Regret}(T)] \leq O\left(\Delta_{\max}\frac{BEL^2\log(T)}{\Delta_{\min}^2} + BEL\right), \quad (7)$$

Algorithm 1 Combinatorial LCB based joint path and beam selection (CLCB-JPBS)

1: If $\max_{\boldsymbol{a}} |\boldsymbol{a}|$ is known, let $L = \max_{\boldsymbol{a}} |\boldsymbol{a}|$, else $L = E$.
2: For each combination of beam and link (l, b): $\hat{\theta}_{l,b}(0) = 0, m_{l,b}(0) = 0$
3: **for** link $l \in \boldsymbol{E}$ **do**
4: **for** beam $b \in \mathcal{B}$ **do**
5: Transmit a packet on a certain route with chosen combination of link and beam $a_{l,b} = 1$ in $\boldsymbol{a}$
6: Update $(\hat{\theta}_{l,b})_{1\times BE}$, $(m_{l,b})_{1\times BE}$
7: **end for**
8: **end for**
9: $t = BE$
10: **for** $t = BE \cdots T$ **do**
11: $t = t + 1$
12: For every l, b, define $\underline{\theta}_{l,b}(t) = \hat{\theta}_{l,b} - \sqrt{\frac{2\log(T)}{m_{l,b}(t)}}$
13: Select a path (set of links) and beams for each link on the path, according to $\boldsymbol{a}(t) = \arg\min_{\boldsymbol{a}\in\boldsymbol{\Theta}} \sum_{(l,b)\in A_{\boldsymbol{a}}} \underline{\theta}_{l,b}$. Then, transmit a packet from s to d using this path and beam combination (retransmit until the packet is successfully delivered to destination.
14: Obtain the feedback of delay $X_{l,b}(t)$ from ACK messages for each selected link and beam pair, and update $(\hat{\theta}_{l,b})_{1\times BE}$, $(m_{l,b})_{1\times BE}$ as follows:

$$(\hat{\theta}_{l,b}(t), m_{l,b}(t)) = \begin{cases} (\frac{\hat{\theta}_{l,b}(t-1)\cdot m_{l,b}(t-1) + X_{l,b}(t)}{m_{l,b}(t-1)+1}, \\ m_{l,b}(t-1)+1), \text{if link } l \text{ in } A_{\boldsymbol{a}(t)}, \\ (\hat{\theta}_{l,b}(t-1), m_{l,b}(t-1)), \\ \text{if link } l \text{ not in } A_{\boldsymbol{a}(t)}. \end{cases}$$

15: **end for**

where $\Delta_{\min} = \min_{\boldsymbol{a}\in\boldsymbol{\Omega},\boldsymbol{a}\neq\boldsymbol{a}^*}\{\mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}^*}\}$, $\Delta_{\max} = \max_{\boldsymbol{a}\in\boldsymbol{\Omega}}\{\mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}^*}\}$.

Proof outline for Theorem 1: The analysis of Alg. 1 is largely inspired by that of LLR [23]. Compared with the LLR algorithm, the difference in our proof is that we define a different event $\mathcal{E}$ to make sure if some arm has sufficient sampling, it will not be chosen anymore. The detailed proof is in Appendix VIII-A.

Remark: Compared with the result of LLR algorithm, our regret bound reduces from $O(\frac{L^3 BE\log(T)\Delta_{\max}}{\Delta_{\min}^2})$ [23] to $O(\frac{L^2 BE\log(T)\Delta_{\max}}{\Delta_{\min}^2})$ (from L^3 to L^2 in the $\log(T)$ term). This is due to the slightly tighter confidence bound construction of each individual arm's cost.

B. Combinatorial Unimodal LCB-based Joint Path and Beam Selection (CULCB-JPBS)

A main drawback of the baseline algorithm is that, its regret guarantee has a linear dependence with B, the number of beams for each link; this is impractical since B can be very large. In this section, we propose a new algorithm that utilizes Unimodal property to achieve a regret guarantee independent of B. The key challenge to combine CLCB-JPBS with a Unimodal bandit algorithm is the nonstationary nature of the rewards within a link (as each link's selected beam changes over time and may gradually converge to pulling the link's optimal beam). Directly using a Unimodal bandit as a subroutine for beam selection for a chosen link does not provide a theoretical regret guarantee. The design of the algorithm should follow the "optimism in the face of uncertainty" principle: it needs to adopt the regret guarantee of the Unimodal bandit algorithm during each link's beam selection to construct a valid lower confidence bound for each link's optimal cost. Our proposed algorithm is built upon the CLCB-JPBS and stochastic golden search for discrete arm (SGSD) [22], and we call it Combinatorial Unimodal LCB based joint path and beam selection (CULCB-JPBS), namely, Alg. 2. It has a provable regret guarantee. Before we present Alg. 2, we make the following assumption for each link l:

Assumption 2. *[Bounded expected cost difference] There exist positive constants D_L and D_H such that $\left|\theta_{l,b} - \theta_{l,b+1}\right| \leq D_H$, and $\left|\theta_{l,b} - \theta_{l,b+1}\right| \geq D_L$ for all $j \in \{1, \dots, B\}$.*

The reason for this assumption is to avoid sharp peak and flat plateau in the expected cost functions. The basic idea of CULCB-JPBS is to decouple the selection of the links and the beams: at every time step t, it first chooses a link vector $\boldsymbol{l}(t)$ from the feasible link vector set (line 10)

$$\tilde{\boldsymbol{\Theta}} = \{\boldsymbol{l} \in \{0,1\}^E : R_{\boldsymbol{l}} \text{ form a path from } s \text{ to } d\},$$

where for any $\boldsymbol{l} \in \{0,1\}^E$, we use $R_{\boldsymbol{l}} := \{l \in \boldsymbol{E} : \boldsymbol{l}_l = 1\}$ to denote the set of links represented by $\boldsymbol{l}$; then the algorithm selects beam for all selected links $\boldsymbol{l}(t)$ using a specialized procedure (line 12 to 23). Our algorithm design follows the "optimism in the face of uncertainty" principle: link sets are chosen according to their optimistic lower confidence bounds (LCBs) on their minimum expected costs $\nu_l = \min_{b \in \mathcal{B}} \theta_{l,b}$.

In more detail, the algorithm proceeds as follows: In the initialization phase (line 3 to 7), it begins by selecting each combination of beam and link at least once to ensure $M_l(t)$ and $\hat{\nu}_l(t)$ are updated. $M_l(t)$ is number of times that link l has been selected and $\hat{\nu}_l(t)$ is the empirical mean value for the link l. Once the initialization is completed, the algorithm selects the link set vector that minimize our designed LCB (line 10). The LCB for link l is:

$$\hat{\nu}_l(t) - \sqrt{\frac{2\log(t)}{M_l(t)}} - \frac{D_H}{D_L}\sqrt{\frac{\log(t)}{M_l(t)}},$$

where the first term is the empirical mean value of the $M_l(t)$ costs obtained by pulling the beams in link l. The second term accounts for the concentration between the sum of the noisy costs and the sum of their corresponding expected costs. The third term is one of our new algorithmic contributions – it accounts for the suboptimality of the arm selection in link l by SGSD so far, calculated by dividing SGSD's regret $O(\frac{D_H}{D_L}\sqrt{M_l(t)})$ by $M_l(t)$. The three terms jointly ensures that the LCB is indeed a high-probability lower bound of ν_i. In line 10, Alg. 2 selects a link set $R_{\boldsymbol{l}} \in \tilde{\boldsymbol{\Theta}}$, and subsequently select a beam for each link and obtain its cost $X_{l,b}$ (lines 12 to 23). Lastly, in line 25, the algorithm updates the chosen link l's statistics, empirical cost mean $\hat{\nu}_l(t)$ and count $M_l(t)$. Other links' statistics remain the same as time step $t-1$.

The optimization problem in line 10 of Alg. 2 is deterministic in each given round t, to obtain the current optimal link set vector $\boldsymbol{l}(t)$. This is also a shortest path problem and we apply Dijkstra's algorithm to solve it in polynomial time.

As mentioned above, for each link l, the algorithm runs a separate copy of the SGSD [22] (line 12 to 23). SGSD is a recent Unimodal bandit algorithm. Its high level idea is to reduce the discrete-arm Unimodal bandits problem to a continuous-arm Unimodal bandits problem, using the Stochastic Golden Search (SGS) algorithm in the continuous arm setting [19]. Specifically, the beam selection problem in link l is a discrete-arm Unimodal bandit problem with expected costs of B beams being $\theta_{l,1}, \dots, \theta_{l,B}$. Every beam j is associated to a point j/B in the $[0,1]$ interval and perform linear interpolation, inducing a function f over the continuous interval $[0,1]$, where for each $j \in \mathcal{B} = \{1, \dots, B\}$:

$$f(x) := \theta_{l,j-1} \cdot (j - Bx) + \theta_{l,j} \cdot (Bx - (j-1)), x \in [(j-1)/B, j/B). \tag{8}$$

SGS is then used to optimize f. Observe that f has minimum at $x^* = j^*/B$ (j^* is the optimal beam $\arg\min_{j \in \mathcal{B}} \theta_{l,j}$), and for $x \in [j/B, (j+1)/B)$, bandit feedback of $f(x)$ can be simulated by pulling a beam randomly from $\{j, j+1\}$ (Alg. 3). To this end, it narrows down the sampling interval, maintaining the invariant that $j^*/B \in [x_A^l, x_C^l]$ with high probability.

Theorem 2: Under the Assumptions 1, 2, when $T \geq 3$, the expected regret of Alg. 2 is:

$$E[\text{Regret}(T)] \leq O\left(\frac{\Delta_{\max} E(\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta_{\min}^2} + \frac{D_H L \log(T)}{(D_L)^2}\right),$$

where $\Delta_{\min} = \min_{\boldsymbol{a} \in \boldsymbol{\Omega}, \boldsymbol{a} \neq \boldsymbol{a}^*}\{\mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}^*}\}$, $\Delta_{\max} = \max_{\boldsymbol{a} \in \boldsymbol{\Omega}}\{\mu_{\boldsymbol{a}} - \mu_{\boldsymbol{a}^*}\}$, $L = \max_{\boldsymbol{a}} |\boldsymbol{a}|$.

Proof outline for Theorem 2: Compared with CLCB-JPBS analysis, the difference in the proof for CULCB-JPBS is that we divide sub-optimal arms into two types:

(1) At least one link in the sub-optimal arm does not share the same link with optimal arm.

(2) All links in the sub-optimal arm shares the same link with the optimal arm (but some beam(s) are different). We define the event $\mathcal{E}'$:

$$\mathcal{E}' \stackrel{\text{def}}{=} \left\{ |\hat{\nu}_l(t) - \nu_l| \leq \sqrt{\frac{2\log(T)}{M_l(t)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}, \forall l, t \right\}.$$

The high-level idea of the proof is as follows:

(1) We bound the probability that event $\mathcal{E}'$ does not happen using Hoeffding inequality.

(2) We bound the expected number of time steps when at least one link in the selected arm does not belong to the optimal arm's path when the event $\mathcal{E}'$ holds.

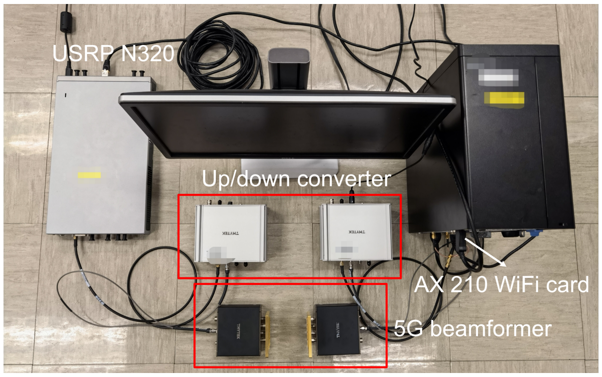


Fig. 2. Transmitter and receiver of a single link.

(3) Lastly, we bound the regret incurred when the algorithm chooses some sub-optimal arm which shares the exact same set of links with the optimal arm. The detailed proof is in the VIII-B.

Remark: Theorem 2 shows that the regret bound depends on the number of links (instead of the number of combination of arms and beams). Compared to the CLCB-JPBS algorithm with a total of $N = BE$ arms (B is the number of individual arms (beams) in each link), whose regret is $O(\frac{L^2 BE \log(T)\Delta_{max}}{\Delta_{min}^2})$, when $\frac{D_H}{D_L} \ll B$, CULCB-JPBS has a much better regret. In practice, to determine D_H, we can set a large enough number $C_{\max}$ for the largest delay, and set the smallest delay as 1. Then, $D_H = C_{\max} - 1$. For D_L, we can set a small enough number based on prior experience.

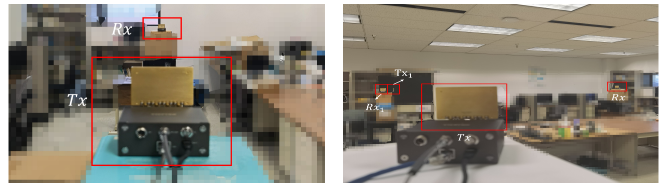
V. Evaluation

To evaluate our proposed algorithms, we use a combination of real-world experiments and simulation. We aim to achieve three goals: (1) Use data collected from real-world experiments to verify the unimodal property of successful packet delivery probability (PDP) on single mmWave links. (2) Use a real-world experimental setup with three links to emulate a multi-hop mmWave network, and evaluate the performance of our proposed algorithms with the collected data. (3) Use simulation with larger-scale networks to verify the scalability of our algorithms.

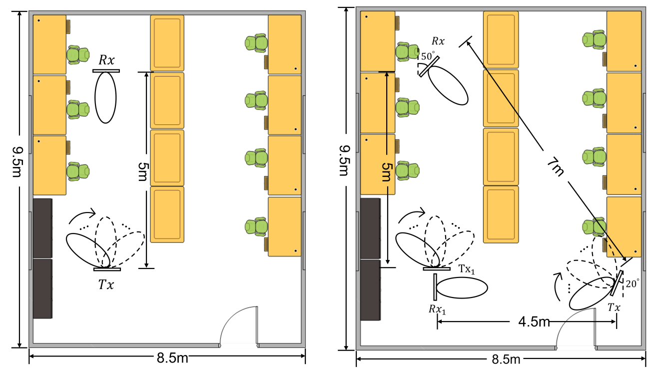
A. Experimental Evaluation

1) Testbed: We developed a mmWave communication testbed consisting of several NI USRPs and Tmytek's mmWave front-end [30] with a center frequency of 28 GHz (in the 5G NR bands).

The transmitter (Tx) is a USRP N320 programmable radio device which operates at 5.2 GHz center frequency. The USRP N320 is connected to a up/down converter which converts the generated signal to 28 GHz. Then the signal is transmitted by a 5G beamformer called BBox Lite, which provides 16 antenna elements for 2-D beamforming. The Half Power Beamwidth (HPBW) of the beamformer is 25°, and we set its gain to 10 dB. The transmit beam can be steered horizontally from -45° to 45° using the *TMXLAB Kit* [30]. To implement the receiver



(a) Single link setup (b) Multi-hop setup



(c) Single link layout (d) Multi-hop layout

Fig. 3. Experimental setups

(Rx), we use another 5G beamformer to receive the 28 GHz signal. Another up/down converter converts the signal from 28 GHz down to 5.2 GHz. An Intel AX 210 WiFi card is used to decode the packets in baseband, and measure the Received Signal Strength (RSS) and Channel State Information (CSI) of each packet using standard procedures [31]. All the hardware devices are controlled by a desktop. The transmitter and receiver of a single link are shown in Fig. 2.

To configure the USRP and extract packet data from the AX 210 WiFi card, we use a software tool called *PicoScenes* [32], [33]. PicoScenes is a CSI tool built upon many open-source software libraries. The integrated GUN radio controls the USRP, so it is compatible with all USRPs that support GUN radio. PicoScenes is a high-performance software implementation of 802.11 a/g/n/ac/ax standards, which allow us to fully control the baseband signal and access the complete physical layer information. We implement the IEEE 802.11ax standard based on an OFDM system with 234 subcarriers using PicoScenes. However, due to the limitation of this tool, we cannot implement a large bandwidth, e.g. 80 MHz and 160 MHz. So we choose 20 MHz bandwidth in all our experiments.

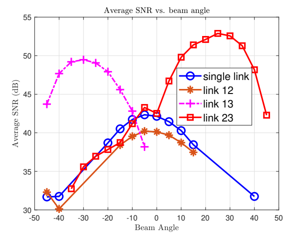
2) Experimental setup: We use two setups in our experiments. Both the first and second setup are employed to verify the unimodal property of successful packet delivery probability over the beam space on a single link, whereas the second one is used for the multi-hop experiment.

Setup 1. Our first setup and layout are shown in Fig. 3(a) and Fig. 3(c). For the transmit beamforming codebook, we use a 5° angle step which provides 19 different beams. We set the transmit power to 15 dBm and the antenna gain of Tx to 10 dB. The Rx fixes its beam direction at 0° (perpendicular to its antenna) so that it points toward the Tx, and the antenna gain is 8 dB, while the Tx steers its beam. Both Tx and Rx are placed 1.5m above the ground (to avoid blockage) and the distance between them is 5m. The Tx sends 10,000 100B-long packets under each beam with a 5ms interval between packets. For each Tx beam and each transmitted packet, we record whether the packet is received successfully or not (1/0) (which means it passes CRC check after error correction), and only when it is successful, CSI/RSS values are extracted.

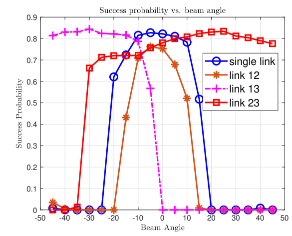
Setup 2. To emulate a multi-hop mmWave network, we first create a three link topology in the lab (a single-hop path and a two-hop path), and then collect over-the-air datasets from all three links. The second setup and layout are shown in Fig. 3(b) and Fig. 3(d). The setting for each node is similar to Setup 1. Since it is difficult to implement the online learning algorithms in real-time in our existing testbed (which requires real-time feedback and control), we collect over-the-air packet data offline (sequentially for each link) and simulate the online algorithms in Matlab. The data collection procedure for each link is the same as the single-link setup (we obtain 10,000 packet delivery success/failure results for each beam of each link). We calculate the average successful packet delivery rate from 10,000 packets as the ground truth for each link and beam. Then, we split all the link-beam's dataset into 3 groups, each of them contains 3300 data records. To simulate an online algorithm using each data group, in each round, the algorithm chooses a set of link (path) and beam combinations (individual arms), and "transmits" a packet on each chosen link and beam pair until it is successfully received (if the delivery fails, we regard the next packet in the dataset as a retransmission). The link delay (instantaneous reward) in this round is calculated as the number of time slots (each of them is 5ms+packet transmission time) taken to successfully deliver that packet. Then the algorithm updates the empirical means for all individual arms and picks another set of link-beam pairs in the next round.

We compare the performance of our proposed CULCB-JPBS algorithm with three baselines: (1) our proposed CLCB-JPBS; (2) Learning with linear rewards (LLR) [23], which is a well-known combinatorial UCB algorithm without considering unimodal property, and we define each link-beam pair as an individual arm; (3) Offline training based on Unimodal Beam Alignment (UBA) [11]. In phase 1, we run UBA algorithm for each link for T rounds (with probing packets). In phase 2, we compute an optimal path using the empirical average delays of optimal beams for each link (learned from phase 1), and we commit to using that path and beam combination.

3) Experimental Results: From the single-link datasets (collected from both setups), we show the average SNR and packet delivery probability (PDP) over 10,000 packets as a function of transmit beam angle in Fig. 4. Note that, in Fig. 4 SNR of a few angles are missing since no packets are received in those directions. Fig. 4 (b) shows that the PDP of each link also have similar trends. The above observations are consistent with our theoretical analysis in Lemma 1. In addition, when the maximum SNR/PDP is achieved, the beam angle is roughly the

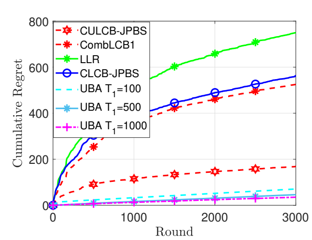


(a) Average SNR vs. beam angle for different links.

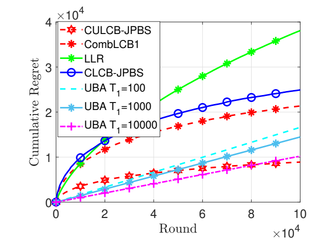


(b) Packet Delivery Probability vs. beam angle for different links

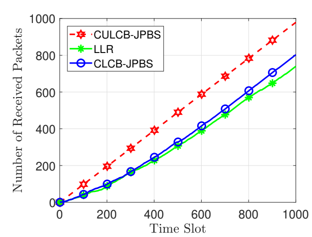
Fig. 4. Average SNR and PDP for single links vs. beam angle



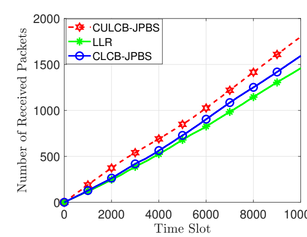
(a) Experiment: Average cumulative regret vs. round/step



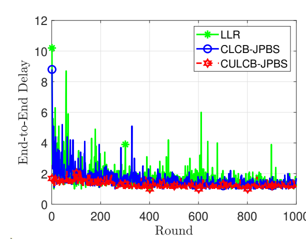
(b) Simulation: Average cumulative regret vs. Round



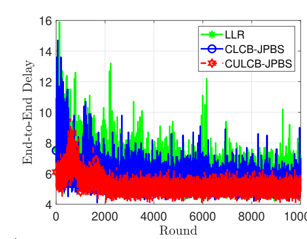
(c) Experiment: Average number of received packets vs. time slot



(d) Simulation: Average number of received packets vs. time slot



(e) Experiment: Average end-to-end delay (time slots) vs. Round



(f) Simulation: Average end-to-end delay (time slots) vs. Round

Fig. 5. Results from experimental data (setup 2), and pure simulation: average cumulative regret, number of received packets and average end-to-end delay

angle of the LoS direction. Note that, the Unimodal property of PDP over the beam space also implies the same for expected link delay.

For algorithm evaluation, Fig. 5 (a) shows the cumulative regret of the joint path and beam selection algorithms under Setup 2 (averaged over three runs, one for each data group). From Fig. 5 (a), we can see that CULCB-JPBS has lower regret than the CLCB-JPBS and LLR algorithm. This result matches our expectation since the Unimodal property of

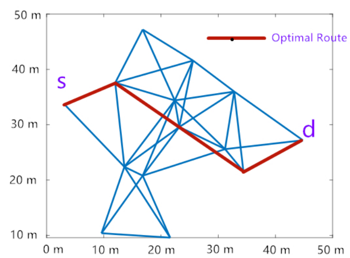


Fig. 6. Topology of the multihop mmWave network used in simulation.

expected delay over the beam space helps the algorithm to converge faster (consistent with our theoretical analysis). In addition, the cumulative regret of CLCB-JPBS is smaller than the LLR algorithm. This is because the exploration term for CLCB-JPBS uses an aggressive confidence bound for selecting arms, while LLR adopts a conservative confidence bound. Meanwhile, Figs. 5 (a) also shows the regret performance (in the online phase) using the UBA algorithm for training. The number of probing rounds in UBA for each link is chosen as $T = 100, 500, 1000$, and we average over 3 runs. We can see that the average cumulative regret grows linearly with the round. This is because running T rounds of UBA is similar to the explore-then-commit strategy [34], where a sub-optimal path-beam combination may be selected and committed to. Although its cumulative regret increases slower, it ultimately surpasses the online algorithms with time. Note that, offline training based on UBA incurs significant overhead because no data is transmitted during the training phase (for T rounds). Our CULCB-JPBS algorithm do not incur any training overhead as data packets are transmitted in each round.

To further examine the advantage of our proposed algorithm over baselines using concrete performance metrics, we evaluate the average number of received packets vs. time slots and average end-to-end (E2E) delay vs. round. From Figs. 5 (c) (e), we can see that CLCB-JPBS and LLR result in overall larger E2E delays and smaller number of received packets over time, while our CULCB-JPBS's E2E delay converges to the optimal value quickly after about 200 rounds (in setup 2, the direct link is the optimal path), and the number of received packets almost increases linearly from the beginning (which also implies high reliability).

B. Simulation-based Evaluation

To evaluate the performance of the algorithms over larger scale networks, we perform simulations using MATLAB. We generate a topology with 14 nodes randomly distributed within a square area of 50 m$\times$ 50 m. For the wireless channel model, we adopt the 3GPP Standard probabilistic LOS model [18], [29], which is often used in mmWave communication. Based on the 3GPP TS 38.101-1/2 standard [35], the system is assumed to operate at 28 GHz carrier frequency, has a bandwidth of 100 MHz, and we use 64-QAM modulation. For each link, there are a total of 16 beams and we only consider Tx beam selection (each beam's width is 5 degrees and the step between adjacent beams' angles is $10°$).

Simulation Results. Figs. 5 (b) (d) (f) show our simulation results. We can see that CULCB-JPBS still has better performance than other algorithms, which is consistent with the results from Setup 2. We can also see that LLR incurs much higher regret than other algorithms. This is partly because its regret depends on L^3, where L, the length of the longest path is 5 (Compared to $L = 2$ in Setup 2), and it does not exploit unimodality structure.

VI. Discussions

A. Extensions

Our algorithm can be easily extended to handle channel dynamics due to blockage or UE mobility. The basic idea is to detect changes of each link's channel state in an online manner, and reset the CULCB-JPBS algorithm once that happens. For example, a simple method is to detect a sudden decrease of link quality indicators (after an enough number of rounds), such as link SNR measurement (this is also adopted by [11]). We can also adopt quick change-point detection schemes (such as CUSUM [26], and others [19]), and integrate them with CULCB-JPBS. As CULCB-JPBS converges fast, it can find the optimal (or near-optimal) combinatorial arm in a few hundred rounds (i.e., less than one second). If the blockage or UE location changes slower than this time scale, it incurs minimal additional regret. As an ongoing work, we are currently studying efficient JPBS algorithms in the piece-wise stationary setting and analyzing their regret bounds.

B. Other Applications

Our algorithm is general enough to be applied to any combinatorial bandit problem with unimodal structure. For instance, in navigation applications in intelligent transportation, a vehicle driving from a source to destination has multiple route choices (paths), each consisting of several road segments (links). For each given link l, it can further choose a travel speed v_l. A combination of a link and speed corresponds to an individual arm. The expected cost for each link can be defined as follows: $r_l = e(v_l) + c \cdot p(v_l)$, where $e(v_l)$ is a function representing the efficiency of travel (e.g., delay), which is a decreasing function of speed, and $p(v_l)$ denotes a safety cost which is usually an increasing function of speed [**?**], and c is a coefficient. Thus, each individual arm's expected cost has the unimodal structure.

VII. Conclusion and Future Work

In this paper, we proposed the CULCB-JPBS algorithm for delay-optimal online joint path and beam selection for multi-hop mmWave networks. To address the scalability issue of combinatorial bandit algorithms under a large number of beams, our algorithm exploits the Unimodal property of expected reward over the beam space for each link. Compared with existing algorithms, we show that the cumulative regret bound of CULCB-JPBS has no dependence on the number of beams per link. We empirically evaluate our algorithm via both experiments and simulations, which significantly outperforms existing algorithms in terms of regret and E2E delay.

There are many interesting questions for future work: (1) We will extend to joint beam selection on both Tx and Rx sides, for which the bi-variate expected cost function has 2-D unimodal property. (2) Currently our algorithm only gathers feedback from the chosen links and beams. Allowing more feedback, such as periodically probing other links/beams, may speed up the learning, but also brings more overhead. We can study the tradeoff between feedback overhead and learning performance. Note that [24] also adopted a probing based online routing method. (3) We will consider different channel conditions, such as multi-path channels, where the reward functions are multimodal. (4) Our current algorithm is centralized. We will design distributed online learning algorithms, such as reinforcement learning using the Unimodal property.

VIII. Appendix

A. Proof of Theorem 1

Proof. Let $G = \{(l,b) : l \in \mathcal{E}, b \in \mathcal{B}\}$ be the collection of (link, beam). $|G| = BE$. Recall that we have defined $\boldsymbol{I}(\mathcal{E}) = 1$ if event $\mathcal{E}$ happens, $\boldsymbol{I}(\mathcal{E}) = 0$ otherwise. cases must happen: 1) an optimal arm a^* is played; 2) a non-optimal a arm is played. In the first case, $(\hat{T}_g(t))_{g\in G}$ will not be updated. When a non-optimal arm is selected, there must exist at least one $g \in a$ such that $g = \arg\min_{g\in a} m_g(t)$. If there is only one such arm, $\hat{T}_g(t)$ is increased by 1. If there are multiple such action, we arbitrarily pick one g, and increment $\hat{T}_g$ by 1. Meanwhile, We define $\hat{g}(t) = g$ to be that $\hat{T}_g(t)$ is increased by one (this implies that individual arm g is selected at time step t).

We define event $\mathcal{E}$ as $\cap_{g\in G}\{|\hat{\theta}_g(t-1) - \theta_g| \le \sqrt{\frac{2\log(T)}{m_g(t-1)}}\}$, where $\hat{T}_g(t) \le m_g(t)$. On event $\mathcal{E}$, we know that

$$UCB_t(\boldsymbol{a}) = \sum_{g\in a}\{\hat{\theta}_g(t-1) + \sqrt{\frac{2\log(T)}{m_g(t-1)}}\} \overset{on\mathcal{E}}{\le} \sum_{g\in \boldsymbol{a}}\{\theta_g + 2\sqrt{\frac{2\log(T)}{m_g(t-1)}}\} = V_t(\boldsymbol{a}). \tag{9}$$

On event $\mathcal{E}$, we define $\boldsymbol{I}(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g)$ ($l_t = \min_{g\in \boldsymbol{a}(t)} m_g(t)$).

$$\begin{aligned}
&Pr(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g) = Pr(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g, \mathcal{E})\\
&+Pr(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g, \mathcal{E}^c)\\
&\le Pr(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g, \mathcal{E}) + Pr(\mathcal{E}^c),
\end{aligned} \tag{10}$$

Next, we will prove the event $\mathcal{E}_1$:$(\hat{T}_g(t) \ge l_t, \hat{g}(t) = g, \mathcal{E})$ is impossible to happen. We need to derive $UCB_t(\boldsymbol{a}(t)) \le \mu_{\boldsymbol{a}^*}$ when event $\mathcal{E}$ happens. First, we know from Eq. (9) from Eq. (9) that $UCB_t(\boldsymbol{a}(t)) \le V_t(\boldsymbol{a}(t))$, therefore,

$$\begin{aligned}
&UCB_t(\boldsymbol{a}(t)) \le V_t(\boldsymbol{a}(t))\\
&= \sum_{g\in \boldsymbol{a}(t)} \{\theta_g + 2\sqrt{\frac{2\log(T)}{m_g(t)}}\}\\
&= \mu_{\boldsymbol{a}(t)} + 2\sum_{g\in \boldsymbol{a}(t)} \sqrt{\frac{2\log(T)}{m_j(t)}}\\
&\overset{(a)}{\le} \mu_{\boldsymbol{a}(t)} + 2\sum_{j\in \boldsymbol{a}(t)} \sqrt{\frac{2\log(T)}{l_t}},
\end{aligned} \tag{11}$$

where the last inequality is from the observation that, according to the definition of $\mathcal{E}_1$, $\forall g \in \boldsymbol{a}(t)$, $m_g(t) \ge l_t$.

Then, we have,

$$\begin{aligned}
&V_t(\boldsymbol{a}(t)) \le \mu_{\boldsymbol{a}(t)} + 2\sum_{j\in \boldsymbol{a}(t)} \sqrt{\frac{2\log(T)}{l_t}}\\
&\overset{(a)}{\le} \mu_{\boldsymbol{a}(t)} + 2L\sqrt{\frac{2\log(T)}{l_t}}
\end{aligned} \tag{12}$$

Where inequality (a) is based on the largest size of arm is L. We set $l_t \ge l = \frac{8L^2\log(T)}{\Delta^2_{\min}}$ where $\Delta_{\min} = \min_{\mathbf{a}\in\mathbf{\Omega},\mathbf{a}\ne\mathbf{a}^*}\{\mu_{\mathbf{a}^*} - \mu_{\mathbf{a}}\}$.

$$\begin{aligned}
&\mu_{a^*} - \mu_{a(t)} - 2L\sqrt{\frac{2\log(T)}{l_t}}\\
&\ge \mu_{\boldsymbol{a}^*} - \mu_{\boldsymbol{a}(t)} - 2L\sqrt{\frac{2\log(T)}{l}}\\
&= \mu_{\boldsymbol{a}^*} - \mu_{\boldsymbol{a}(t)} - 2L\sqrt{\frac{2\log(T)}{\frac{8M^2\log(T)}{\Delta^2_{\min}}}} \ge 0,
\end{aligned} \tag{13}$$

So, it is impossible to select arm $\boldsymbol{a}(t)$ because $V_t(\boldsymbol{a}(t)) \le \mu_{\boldsymbol{a}^*}$.

For the second term in Eq. (10), if event $\mathcal{E}$ does not hold, it means that at least one individual arm does not hold a certain threshold. We assume that all arm contains L individual arm. The statement becomes: if event $\mathcal{E}$ does not hold, it means that at least one individual arm g satisfies $|\hat{\theta}_g(t-1) - \theta_g| \ge \sqrt{\frac{2\log(T)}{m_g(t)}}$.

$$\begin{aligned}
&Pr(E^c) = Pr(\exists g \in G, |\hat{\theta}_g(t-1) - \theta_g| \ge \sqrt{\frac{2\log(T)}{m_g(t)}})\\
&= \sum_{g\in G} Pr(|\hat{\theta}_g(t-1) - \theta_g| \ge \sqrt{\frac{2\log(T)}{m_g(t)}})\\
&\overset{(a)}{\le} |G|T^{-4},
\end{aligned} \tag{14}$$

Inequality (a) is based on the Lemma 1 in useful facts (31). We define τ is the first time slot that $\hat{T}_g(t) \geq l$. Then, the expectation of $\hat{T}_g(t)$,

$$\begin{aligned}
E[\hat{T}_g(t)] &\leq l + \sum_{t=\tau+1}^{T} Pr(\hat{T}_g(t) \geq l, \hat{g}(t) = g) \\
&\overset{(a)}{\leq} l + \sum_{t=\tau+1}^{T} Pr(\hat{T}_g(t) \geq l, \hat{g}(t) = g, E) + \sum_{t=1}^{T} Pr(E^c) \\
&\overset{(b)}{\leq} l + \sum_{t=\tau+1}^{T} Pr(\hat{T}_g(t) \geq l, \hat{g}(t) = g, E) + O(|G|) \\
&\overset{(c)}{\leq} l + |G|, \qquad (15)
\end{aligned}$$

where inequality (a) is from Eq. (10); inequality (b) is from Eq. (14); inequality (c) is from Eq. (13) (it is impossible to select arm $\boldsymbol{a}(t)$ because $V_t(\boldsymbol{a}(t)) \leq \mu_{\boldsymbol{a}^*}$).

The regret of CUCB-JPBS is,

$$\begin{aligned}
E[R(T)] &= \sum_{\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a}^*}} \Delta_{\boldsymbol{a}} E[T_{\boldsymbol{a}}(T)] \\
&\leq \Delta_{\max} \sum_{\mu_{\boldsymbol{a}} < \mu_{\boldsymbol{a}^*}} E[T_{\boldsymbol{a}}(T)] = \Delta_{\max} \sum_{i=g}^{BE} E[\hat{T}_g(T)] \\
&\leq \Delta_{\max}\{BEl + O(BEM)\} \\
&= \Delta_{\max}\{BE\frac{8L^2 \log(T)}{\Delta_{\min}^2} + O(BEL)\}, \qquad (16)
\end{aligned}$$

Where l can ensure all super arm have sufficient sampling. □

B. Proof of Theorem 2

Proof. For each link l, let E be the collection of link. Recall that we have defined $\boldsymbol{I}(\mathcal{E}') = 1$ if event $\mathcal{E}'$ happens, $\boldsymbol{I}(\mathcal{E}') = 0$ otherwise. We define R_p as the link set $\tilde{\boldsymbol{\Theta}}$ is the set of R_p. At each time-slot after the initialization period, one of the two cases must happen: 1) $L_{\boldsymbol{a}(t)} = L_{\boldsymbol{a}^*}$ ($a(t)$ shares the same link with optimal arm a^* is played); 2) $L_{\boldsymbol{a}(t)} \neq L_{\boldsymbol{a}^*}$ (At least one link does not same with optimal arm a^*) is played. We define $M_l(t)$ as follows: In the first case, $(M_l(t))_{l\in\mathcal{E}}$ will not be updated. When R_p is selected, there must exist at least one $l \in R_p$ such that $l = \arg\min_{l\in R_p} m_l$. If there is only one such link, $M_l(n)$ is increased by 1. If there are multiple such action, we arbitrarily pick one l, and increment M_l by 1.

We define event $\mathcal{E}'$ as $\cap_{l\in\mathcal{E}}\{|\hat{\theta}_l(t-1) - \theta_l| \leq \sqrt{\frac{2\log(T)}{M_l(t-1)}} + \frac{D_H}{D_L}\sqrt{\frac{1}{M_l(t)}}\}$. On event $\mathcal{E}'$, we know that

$$\begin{aligned}
UCB_t(\boldsymbol{l}(t)) &= \sum_{l\in R_p}\{\hat{\nu}_l(t-1) + \sqrt{\frac{2\log(T)}{M_l(t-1)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}\} \\
&\overset{on\,\varepsilon'}{\leq} \sum_{l\in R_p}\{\nu_l + 2\sqrt{\frac{2\log(T)}{M_l(t-1)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}\} \\
&= V_t(\boldsymbol{l}(t)). \qquad (17)
\end{aligned}$$

On event $\mathcal{E}'$, we define $\hat{e}(t)$ to be the unique link $l \in E$ such that $\hat{T}_l(t)$ gets incremented at time step t, $\hat{l}(t) = \min_{l\in R_l} m_l(t)$.

$$\begin{aligned}
&Pr(\hat{T}_l(t) \geq \hat{l}(t), \hat{e}(t) = l) \\
&= Pr(\hat{T}_l(t) \geq \hat{l}(t), \hat{e}(t) = l, \mathcal{E}') + Pr(\hat{T}_l(t) \geq \hat{l}(t), \hat{e}(t) = l, \mathcal{E}'^c) \\
&\leq Pr(\hat{T}_l(t) \geq \hat{l}(t), \hat{e}(t) = l, \mathcal{E}') + Pr(\mathcal{E}'^c), \qquad (18)
\end{aligned}$$

Next, we will prove the probability of event $\mathcal{E}_1'$ happens: $Pr(\hat{T}_{g'}(t) \geq \hat{l}(t), \hat{e}(t) = l, \mathcal{E}')$ is impossible to happen. We need to derive $UCB_t(\boldsymbol{l}(t)) \leq \mu_{a^*}$ when event $\boldsymbol{\varepsilon}'$ happens. First, we know from Eq. (17) that $UCB_t(\boldsymbol{l}(t)) \leq V_t(\boldsymbol{l}(t))$, therefore,

$$\begin{aligned}
&UCB_t(\boldsymbol{l}(t)) \leq V_t(\boldsymbol{l}(t)) \\
&= \sum_{l\in R_{\boldsymbol{l}(t)}}\{\theta_l + 2\sqrt{\frac{2\log(T)}{M_l(t)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}\} \\
&= \mu_{R_{\boldsymbol{l}(t)}} + 2\sum_{l\in R_{\boldsymbol{l}(t)}}\{\sqrt{\frac{2\log(T)}{M_l(t)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}\} \\
&\overset{(a)}{\leq} \mu_{\boldsymbol{l}(t)} + 2\sum_{l\in R_{\boldsymbol{l}(t)}}\{\sqrt{\frac{2\log(T)}{l_t'}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{l_t'}}\} \qquad (19)
\end{aligned}$$

According to the definition of $\mathcal{E}_1'$, we know that $\forall l \in R_{\boldsymbol{l}(t)}$, $M_l(t) \geq \hat{l}(t)$. Then, when $T > 3$, we have,

$$\begin{aligned}
&V_t(\boldsymbol{l}(t)) \\
&\leq \mu_{\boldsymbol{l}(t)} + 2\sum_{l\in R_{\boldsymbol{l}(t)}}\{\sqrt{\frac{2\log(T)}{\hat{l}(t)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{\hat{l}(t)}}\} \\
&\leq \mu_{\boldsymbol{l}(t)} + 2(1 + \frac{D_H}{D_L})\sum_{l\in R_{\boldsymbol{l}(t)}}\sqrt{\frac{2\log(T)}{\hat{l}(t)}} \\
&\overset{(b)}{\leq} \mu_{\boldsymbol{l}(t)} + 2(1 + \frac{D_H}{D_L})L\sqrt{\frac{2\log(T)}{\hat{l}(t)}}, \qquad (20)
\end{aligned}$$

Where inequality (a) is based on the fact that $2\log(T) \geq 1$ when $T \geq 3$. Inequality (b) is based on the largest length of super arm is L. We set $\hat{l}(t) \geq l' = \frac{8(1+\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta_{\min}'^2}$ where $\Delta_{\min}' = \min_{p\in\tilde{\boldsymbol{\Theta}}} \mu_{p^*} - \mu_p$. $\mu_p = \sum_{l\in p} \nu_l$, and p^* is the route of the optimal (route, beam) combination.

$$\begin{aligned}
&\mu_{p^*} - \mu_{\boldsymbol{l}(t)} - 2(1 + \frac{D_H}{D_L})L\sqrt{\frac{2\log(T)}{l_t'}} \\
&\geq \mu_{p^*} - \mu_{\boldsymbol{l}(t)} - 2(1 + \frac{D_H}{D_L})L\sqrt{\frac{2\log(T)}{l'}} \\
&= \mu_{p^*} - \mu_{\boldsymbol{l}(t)} - 2(1 + \frac{C_H}{C_L})L\sqrt{\frac{2\log(T)}{\frac{8(1+\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta_{\min}'^2}}} \\
&\geq 0, \qquad (21)
\end{aligned}$$

So, it is impossible to select link set $R_{\boldsymbol{l}(t)}$ because $UCB_t(\boldsymbol{l}(t)) \leq V_t(\boldsymbol{l}(t))$, $V_t(\boldsymbol{l}(t)) \leq \mu_{a^*}$ and $\mu_{p^*} \leq UCB_t(p^*)$. Then, we have $UCB_t(\boldsymbol{l}(t)) \leq UCB_t(\boldsymbol{l}^*)$

For the second term, if event $\mathcal{E}'$ does not hold, it means that at least one link does not hold a certain threshold. We assume that the largest number of element in link set R_p is L links. The statement becomes: if event $\boldsymbol{\varepsilon}'$ does not hold, it means that at least one link is $|\hat{\theta}_l(t-1) - \theta_l| \geq \sqrt{\frac{2\log(T)}{M_l(t)}} + \frac{D_H}{D_L}\sqrt{\frac{\log(T)}{M_l(t)}}$.

$$\begin{aligned}
&Pr(\varepsilon^c) \\
&= \sum_{l\in E} Pr(\exists l \in E, |\hat{\theta}_l(t-1) - \theta_l| \geq \sqrt{\frac{2\log(T)}{m_l(t)}}) \\
&= \sum_{l\in E} Pr(|\hat{\theta}_l(t-1) - \theta_l| \geq \sqrt{\frac{2\log(T)}{M_l(t)}} + \frac{D_H}{D_L}\sqrt{\frac{1}{M_l(t)}}) \\
&\overset{(a)}{\leq} LT^{-1}, \qquad (22)
\end{aligned}$$

Inequality (a) is based on the fact the result of Eq. (40) in [36]. $L = \max_{\boldsymbol{a}} |\boldsymbol{a}|$. We define τ' is the first time slot that $\hat{T}_l(t) \geq l'$. Then, the expectation of $\hat{T}_l(t)$,

$$\begin{aligned}
&E[\hat{T}_l(t)] \leq l' + \sum_{t=\tau'}^{T} Pr(\hat{T}_l(t) \geq l', \hat{e}(t) = l) \\
&\leq l' + \sum_{t=\tau'}^{T} Pr(\hat{T}_l(t) \geq l', \hat{e}(t) = l, E) + \sum_{t=\tau'}^{T} Pr(\varepsilon^c) \\
&\leq l' + \sum_{t=\tau'}^{T} Pr(\hat{T}_l(t) \geq l', \hat{e}(t) = l, E) + O(L) \\
&\leq l' + O(E), \qquad (23)
\end{aligned}$$

We decompose the expected regret $E[R(T)]$ into the sum of two terms $E[R_1(T)]$ and $E[R_2(T)]$, where

$$E[R_1(T)] = E[\sum_{t=1}^{T} \boldsymbol{I}(\exists l \in R_{\boldsymbol{l}(t)} : l \notin R_{p^*})(\mu_{p^*} - \mu_{\boldsymbol{l}(t)})],$$

is the regret of the selected arm at least one link does not belong to the optimal arm, and

$$E[R_2(T)] = E[\sum_{t=1}^{T} \boldsymbol{I}(\forall l \in R_{\boldsymbol{l}(t)} : l \in R_{l^*})(\mu_{l^*} - \mu_{\boldsymbol{l}(t)})], \qquad (24)$$

is the regret in rounds in which the sub-optimal arm has all the same links as the optimal arm.

We bound the two terms respectively. For $E[R_1(T)]$,

$$\begin{aligned}
&E[R_1(T)] \leq \Delta_{max} \sum_{\mu_p < \mu_{a^*}} E[T_p(T)] \\
&= \Delta_{\max} \sum_{l=1}^{E} E[\hat{T}_l(T)] \leq \Delta_{\max}\{El' + O(E^2)\} \\
&\leq \Delta_{\max}\{E \frac{8(1+\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta'^2_{\min}} + O(E^2)\} \\
&\overset{(a)}{\leq} \Delta_{\max}\{E \frac{8(1+\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta^2_{\min}} + O(E^2)\}, \qquad (25)
\end{aligned}$$

where $\Delta_{\min} = \min_{\mathbf{a} \in \mathbf{\Omega}, \mathbf{a} \neq \mathbf{a}^*}\{\mu_{\mathbf{a}^*} - \mu_{\mathbf{a}}\}$, and Inequality (a) is based on the fact that $\Delta'_{\min} \geq \Delta_{\min}$. $\Delta_{\max} = \max_{\mathbf{a} \in \mathbf{\Omega}}\{\mu_{\mathbf{a}^*} - \mu_{\mathbf{a}}\}$. $\boldsymbol{A_1}$ is the arm set that shares the same link with optimal arm a^*

Next, we will bound $E[R_2(n)]$. In this scenario, the only difference is the beam. b_l is the selected beam in link l, Δ_{b_l} is the regret between optimal beam and selected beam b_l and $\boldsymbol{B_1}$ is the individual arm set for b_l. Here,we consider that link l is the one part of link of optimal arm $\boldsymbol{a}^*$.

$$\begin{aligned}
&E[R_2(T)] = E[\sum_{t=1}^{T} \boldsymbol{I}(\forall l \in R_{\boldsymbol{l}(t)} : l \in R_{l^*})(\mu_{l^*} - \mu_{\boldsymbol{l}(t)})] \\
&= \sum_{a \in \boldsymbol{A_1}} \Delta_a E[T_a(T)] = \sum_{a \in \boldsymbol{A_1}} \sum_{(l,b) \in A_a} \Delta_{b_l} \\
&E[T_a(T)] \overset{(a)}{=} \sum_{b_l \in \boldsymbol{B_1}} \sum_{a \cap b_l \neq \emptyset, a \in \boldsymbol{A_1}} \Delta_{b_l} E[T_a(T)] \\
&\overset{(b)}{=} \sum_{b_l \in \boldsymbol{B_1}} \Delta_{b_l} E[T_{b_l}(T)], \qquad (26)
\end{aligned}$$

The operation of Eq. (a) is to group all the sub-optimal arm a that contains b_l. Eq. (b) is based on the fact that the summation of number of time for sub-optimal arm which contains b_l equals to the total selected time for individual arm b_l. We can divide the regret analysis into each link. For each link l, the regret is (Theorem 2 in [36])

$$E[R_l(T)] \leq O\left(\frac{D_H}{(D_L)^2} \log(8T)\right). \qquad (27)$$

The regret of the selected route which shares the same link with optimal arm is

$$\begin{aligned}
E[R_2(T)] &\leq \sum_{l \in a^*} O\left(\frac{D_H}{(D_L)^2} \log(8T)\right) \\
&\overset{(a)}{\leq} O\left(L \frac{D_H}{(D_L)^2} \log(8T)\right). \qquad (28)
\end{aligned}$$

Inequality (a) is based on the fact that the route length cannot exceed L. Combining 25 and 28, we have

$$\begin{aligned}
&E[R(T)] = E[R_1(T)] + E[R_2(T)] \leq O\left(E \frac{D_H}{(D_L)^2} \log(8T)\right) \\
&+\Delta_{max}\{L \frac{8(1+\frac{D_H}{D_L})^2 L^2 \log(T)}{\Delta^2_{\min}} + O(E^2)\}. \qquad (29)
\end{aligned}$$

□

C. Useful Facts

Lemma 2. *For any fixed individual arm $g \in G$ and any time step t in the learning process,*

$$Pr(|\hat{\theta}_g(t-1) - \theta_g| \geq \sqrt{\frac{2\log(T)}{m_g(t)}}) \leq \frac{2}{T^4}, \qquad (30)$$

Proof. Applying Chernoff–Hoeffding Bound, we can get

$$\begin{aligned}
&Pr(|\hat{\theta}_g(t-1) - \theta_g| \geq \sqrt{\frac{2\log(T)}{m_g(t)}}) \\
&\leq Pr(m_g(t)\hat{\theta}_g(t-1) - m_g(t)\theta_g \geq \sqrt{2m_g(t)\log(T)}) \\
&+Pr(m_g(t)\hat{\theta}_g(t-1) - m_g(t)\theta_g \leq -\sqrt{2m_g(t)\log(T)}) \\
&\leq e^{-4\log(T)} + e^{-4\log(T)} = \frac{2}{T^4}, \qquad (31)
\end{aligned}$$

□

References

[1] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5g radio network design for ultra-reliable low-latency communication," *IEEE network*, vol. 32, no. 2, pp. 24–31, 2018.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5g wireless systems," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 136–143, 2014.

[4] A. Ghosh and M. Cudak, "Integrated access and backhaul: Why it is essential for mmwave deployments," https://www.nokia.com/blog/integrated-access-and-backhaul-why-it-is-essential-for-mmwave-deployments/,, 2020.

[5] G. Brown, "Exploring the potential of mmwave for 5g mobile access," *Qualcomm White Paper*, 2016.

[6] "Breaking the wireless barriers to mobilize 5g nr mmwave," *Qualcomm White Paper*, 2019.

[7] D. Pliatsios, P. Sarigiannidis, S. Goudos, and G. K. Karagiannidis, "Realizing 5g vision through cloud ran: technologies, challenges, and trends," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–15, 2018.

[8] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, 2015.

[9] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE communications magazine*, vol. 49, no. 6, pp. 101–107, 2011.
[10] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, and M. Latva-Aho, "Path selection and rate allocation in self-backhauled mmwave networks," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.
[11] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE INFOCOM 2018*. IEEE, 2018, pp. 2393–2401.
[12] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang, and X. Shen, "Fast mmwave beam alignment via correlated bandit learning," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5894–5908, 2019.
[13] I. Aykin, B. Akgun, M. Feng, and M. Krunz, "Mamba: A multi-armed bandit framework for beam tracking in millimeter-wave systems," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1469–1478.
[14] Y. Wang, Z. Wei, and Z. Feng, "Beam training and tracking in mmwave communication: A survey," *arXiv preprint arXiv:2205.10169*, 2022.
[15] W. Attaoui, K. Bouraqia, and E. Sabir, "Initial access & beam alignment for mmwave and terahertz communications," *IEEE Access*, vol. 10, pp. 35 363–35 397, 2022.
[16] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. ACM, 2018, pp. 432–445.
[17] D. Yuan, H.-Y. Lin, J. Widmer, and M. Hollick, "Optimal joint routing and scheduling in millimeter-wave cellular networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1205–1213.
[18] L. Sun, J. Hou, and T. Shu, "Spatial and temporal contextual multi-armed bandit handovers in ultra-dense mmwave cellular networks," *IEEE Transactions on Mobile Computing*, vol. 20, no. 12, pp. 3423–3438, 2020.
[19] J. Y. Yu and S. Mannor, "Unimodal bandits," 2011.
[20] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *International Conference on Machine Learning*. PMLR, 2014, pp. 521–529.
[21] Y. Zhang, S. Basu, S. Shakkottai, and R. W. Heath Jr, "Mmwave codebook selection in rapidly-varying channels via multinomial thompson sampling," in *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2021, pp. 151–160.
[22] T. Zhao, C. Zhang, and M. Li, "Hierarchical unimodal bandits," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 269–283.
[23] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.
[24] T. He, D. Goeckel, R. Raghavendra, and D. Towsley, "Endhost-based shortest path routing in dynamic networks: An online learning approach," in *2013 Proceedings IEEE INFOCOM*. IEEE, 2013, pp. 2202–2210.
[25] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, "Stochastic online shortest path routing: The value of feedback," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 915–930, 2017.
[26] F. Liu, J. Lee, and N. Shroff, "A change-detection based framework for piecewise-stationary multi-armed bandit problem," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[27] P. Yang, W. Wu, N. Zhang, and X. Shen, "Machine learning-based beam alignment in mmwave networks," in *Millimeter-Wave Networks*. Springer, 2021, pp. 37–71.
[28] D. Jiang and G. Liu, "An overview of 5g requirements," *5G Mobile Communications*, pp. 3–26, 2017.
[29] 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.
[30] [Online]. Available: https://tmytek.com
[31] T. Aoki, Y. Egashira, and D. Takeda, "Preamble structure for mimo-ofdm wlan systems based on ieee 802.11 a," in *2006 IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2006, pp. 1–6.
[32] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the barriers: Demystifying wi-fi baseband design and introducing the picoscenes wi-fi sensing platform," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4476–4496, 2021.
[33] [Online]. Available: https://ps.zpj.io/#
[34] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
[35] A. V. Lopez, A. Chervyakov, G. Chance, S. Verma, and Y. Tang, "Opportunities and challenges of mmwave nr," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 4–6, 2019.
[36] T. Zhao, C. Zhang, and M. Li, "Hierarchical unimodal bandits," https://realworldml.github.io/files/cr/paper31.pdf, ICML2022 Workshop on ReALML.

D. Algorithm

Algorithm 2 CULCB-JPBS

Input: D_H, D_L
For each link l: $\hat{\nu}_l(0) = 0, M_l(0) = 0, x_A^l = 0, x_B^l = \frac{1}{\phi^2}$, $x_c^l = 1$, precision schedule $(\epsilon_1, \ldots, \epsilon_S)$, $s_l = 1$ for all $l \in E$.
Same as lines 2 to 8 of Alg. 1 for initialization.
for link $l \in \boldsymbol{E}$ **do**
  $M_l(t) = \sum_{b\in\mathcal{B}} m_{l,b}(t)$
  $\hat{\nu}_l(t) = \sum_{b\in\mathcal{B}} \frac{m_{l,b}(t)}{M_l(t)} \hat{\theta}_{l,b}(t)$
end for
for $t = BE \cdots T$ **do**
  $t = t + 1$
  Select a link vector $\boldsymbol{l}(t) = \arg\min_{\boldsymbol{l}\in\tilde{\boldsymbol{\Theta}}} \sum_{l\in R_{\boldsymbol{l}(t)}} (\hat{\nu}_l(t) - \sqrt{\frac{2\log(t)}{M_l(t)}} - \frac{D_H}{D_L}\sqrt{\frac{\log(t)}{M_l(t)}})$.
  for each link l in the chosen path $R_{\boldsymbol{l}(t)}$: **do**
    if there are more than one discrete beam j such that j/B in $[x_A^l, x_C^l]$ **then**
      In a round-robin manner, select one point in $\{x_A^l, x_B^l, x_B'^l, x_C^l\}$ and choose beam $b_l(t)$ according to Alg. 3, receive cost $X_{l,b_l(t)}(t)$
      if all of $\{x_A^l, x_B^l, x_B'^l, x_C^l\}$ are selected $\frac{2}{\epsilon_{s_l}^2}\log(8T)$ times **then**
        $s_l = s_l + 1$; let $\hat{x}^l$ be the point with smallest empirical mean cost in this phase
        If $\hat{x}^l \in \{x_A^l, x_B^l\}$ then eliminate interval $(x_B'^l, x_C^l]$ and let $x_C^l = x_B'^l$,
        else eliminate interval $[x_A^l, x_B^l)$ and let $x_A^l = x_B^l$.
        Let $x_B'^l = x_B^l - \frac{1}{\phi^2}(x_B^l - x_A^l)$ if $x_B^l - x_A^l > x_C^l - x_B^l$, and $x_B'^l = x_B^l + \frac{1}{\phi^2}(x_C^l - x_B^l)$ otherwise
      end if
    else
      Break
    end if
    Select the only discrete beam j such that j/B in $[x_A^l, x_C^l]$, let $b_l(t) = j$ and receive cost $X_{l,b_l(t)}(t)$
  end for
  Update empirical mean costs and counts for all links:

$$(\hat{\nu}_l(t), M_l(t)) = \begin{cases} (\frac{\hat{\nu}_l(t-1)\cdot M_l(t-1) + X_{l,b_l(t)}(t)}{M_l(t-1)+1}, \\ M_l(t-1)+1), \quad \text{link } l \text{ in } R_{\boldsymbol{l}(t)}, \\ (\hat{\nu}_l(t-1), M_l(t-1)), \\ \text{link } l \text{ not in } R_{\boldsymbol{l}(t)}. \end{cases}$$

end for

Algorithm 3 Cost sampling for any continuous point x'

Input: x'
Output: a stochastic cost of conditional mean $f(x')$ (Eq. (8))
$j = \lfloor Bx' \rfloor$
set

$$b = \begin{cases} j & \text{with probability} \quad j + 1 - Bx' \\ j+1 & \text{otherwise,} \end{cases}$$

$r \leftarrow$ cost of pulling beam b
return r