Context-Aware Beam Management via Online Probing in Combinatorial Multi-Armed Bandits

Zhizhen Li*, Xuanhao Luo*, Mingzhe Chen†, Chenhan Xu*, Yuchen Liu*
*North Carolina State University, USA, †University of Miami, USA

***Abstract*—Millimeter-wave (mmWave) communication, a cornerstone in the evolution of next-generation wireless networks, offers substantial bandwidth and plays a crucial role in advancing wireless connectivity capabilities. Nevertheless, the inherent directionality and susceptibility to blockages pose significant challenges for a cost-effective beam management in densely deployed networks. This paper presents a Contextual Combinatorial Beam Management (CCBM) framework, leveraging both location-aware link qualities and beam correlation to tackle the joint access point (AP) and beam selection problem in mmWave networks, with a specific focus on mitigating coordination overhead and balancing the load across APs. Built upon a formulated multi-armed bandit problem, CCBM significantly reduces the uncertainty during online probing process by employing early stopping and attention-based selection mechanisms. Theoretical analysis establishes the asymptotically optimality of the proposed approach, complemented by extensive evaluation results showcasing the superiority of our framework over other state-of-the-art schemes in multiple dimensions.**

I. Introduction

Millimeter-wave (mmWave) communication is regarded as a highly promising technology due to its ability to provide high-bandwidth and low-latency wireless connectivity. The substantial data rates offered by mmWave can effectively cater to the escalating demands of densely deployed devices and bandwidth-intensive applications in wireless local-area networks (WLANs). In a mmWave WLAN scenario, the dense deployment of access points (APs) becomes crucial to ensuring ubiquitous coverage and network robustness, given the directional nature of communications. However, this brings about significant challenges in the system design. First, managing directional beams between a large number of clients and APs introduces formidable overhead [1], which escalates linearly with the number of communication entities. Second, the paired beams are highly sensitive to both static and dynamic blockages [2], attributed to the limited propagation distance and poor penetration capabilities of mmWave links. Third, the uncertain and time-varying nature of the mmWave channel poses challenges in performing an adaptive beamforming, especially when considering load balancing for a consistent level of service in such multi-AP multi-user environments.

Traditionally, several solutions have been developed for beam management and resource allocation. However, a key assumption in these works is that the channel condition is well-known from the start, a challenge particularly in densely deployed mmWave scenarios given their time-varying nature. Recently, various learning-based approaches have emerged to address the uncertainty in beam alignment and selection. For instance, in [3], a deep learning framework was proposed to predict link quality between beams. While effective in reducing overhead and achieving high performance in a site-specific vehicular network, its implementation on network devices demands significant computational resources. Alternatively, a multi-armed bandit (MAB) based online learning framework appears more suitable, as it negates the need for offline data collection and strikes a balance between exploration and exploitation in uncertain environments. In particular, [4] adopted a contextual MAB approach to address the beam selection problem, using user location as side information to aid decision-making. In [5], the correlation between beams is considered as arm context, and a unimodal beamforming algorithm was proposed. However, none of these works considered a practical mmWave network scenario with numerous obstacles, and load-balanced resource allocation is not jointly considered in their schematic designs. Notably, [6] partially addressed the resource allocation problem with a coarse-level AP probing algorithm, but it did not manage the beam pairing for each deployed AP. To our knowledge, there has been no comprehensive study on a joint AP and beam selection scheme with load balancing considered in an obstacle-rich mmWave network, which is the subject of this work herein.

In our prior works [7], [8], we developed a regression-based machine learning framework to predict link quality under both static and dynamic blockages. This framework achieves an accuracy rate of up to 94% and requires only a few environmental information as input. Particularly, it seamlessly adapts to different indoor scenarios by merely modifying input data, eliminating the need for additional training. By utilizing such link quality predictions as prior contextual knowledge, the overhead produced in AP probing and beamforming processes can be greatly reduced, as intuitively, APs offering high signal strength at specific locations can be selected for optimal beam pairing. This prior work lays a foundation for the beam management study, facilitating a context-aware online probing approach towards coordination-minimal wireless environment.

In this paper, we present a novel contextual combinatorial beam management (CCBM) framework designed to tackle the joint AP and beam selection problem in mmWave wireless networks, ensuring a balanced load distribution among dense APs for consistent user services with minimum coordination overhead. In CCBM, each beam is treated as an arm, with the received power serving as the reward for selecting specific beams. The objective is to sequentially choose these

978-1-7281-9054-9/24/$31.00 ©2024 IEEE

arms to maximize cumulative rewards within a given time horizon, particularly allowing users to explore *multiple* arms and evaluate their rewards before finalizing the AP-beam selection. This approach significantly minimizes uncertainty by revealing arm rewards before the decision-making process, while minimizing coordination overhead between users and APs. Additionally, by leveraging link quality predictions of unknown beam directions from our prior works [7], [8], the CCBM framework prioritizes APs based on their predicted link quality at the user location. Only beams associated with higher predicted values from these APs are considered during the online probing process. This strategy expedites the assessment of network conditions by avoiding unnecessary searches among irrelevant candidate beams, as both the environmental context and arm context implicitly contribute to the rapid identification of optimal beams. The main contributions of this work are summarized as follows.

- We innovatively frame the joint AP selection and beam management as a contextual combinatorial MAB problem, naturally leveraging the correlation between nearby beams and location-aware link qualities as context information to expedite the beam selection process.
- In our proposed CCBM framework, we incorporate a novel attention-based selection scheme along with an early stopping criterion to prevent excessive exploration during the online probing process. Theoretical analysis establishes an upper bound on cumulative regret, i.e. the gap to the results obtained from an oracle search, which demonstrates the asymptotic optimality of our approach.
- We develop a reward function within the MAB algorithm that explicitly considers load balancing among candidate APs, guiding users in the online probing process to select a globally optimal AP and beam for connection, thereby optimizing overall network performance.
- Comprehensive performance evaluations demonstrate the superiority of our CCBM framework over baseline approaches in various dimensions, including lower regret, increased user throughput, and improved load balancing across densely deployed mmWave APs.

II. Problem Formulation

In this section, we elaborate on the process of transforming the joint AP and beam selection problem into a contextual combinatorial MAB problem, and then derive an online probing algorithm for effective beam management.

Let N denote the number of APs in a wireless network environment and C represent the number of orthogonal beam patterns associated with each mmWave AP. Additionally, assume that M clients are moving randomly within the space. Let X represent the set of environmental contexts corresponding to the user locations. At each time step $t = 1, ..., T$, where T denotes a predetermined time horizon, the location $x_m^t \in X$ of user m at t can be observed. Subsequently, the link quality predictions obtained from the spatial-temporal model are utilized to rank APs based on the maximum signal strength they can offer at each user location. We establish an AP candidate set with the size of A for each user location x_m^t by selecting the top-A APs. Considering beam pairing, all the beams from each AP candidate set collectively form a beam set $\mathcal{A}_m^t = \{a_i^j | i \leq C, j \leq A\}$, where a_i^j represents the i-th beam of the j-th candidate AP.

Based on the above setup, each beam from $\mathcal{A}_m^t$ can be treated as an arm in a MAB problem. At each time step t, instead of playing just one arm, a subset of arms $\mathcal{S}_m^t \subset \mathcal{A}_m^t$ will be selected to play. There exists a budget B that limits the maximum number of arms that can be probed, i.e., $|\mathcal{S}_m^t| \leq B$. To optimally choose the subset $\mathcal{S}_m^t$, we incorporate *arm context* information $\mathcal{X} = \{O_a | a \in \mathcal{S}_m^t\}$. Specifically, in our considered scenario, arm context refers to the direction of each beam, where the details about the arm selection will be introduced in Sec. III. Here we define the reward of selecting a beam a at the user location x as corresponding to the signal strength of the beam alignment process. We denote this reward by $r_{a|x}$ and its expected value by $\mu_{a|x} = \mathbb{E}[r_{a|x}]$. To model the probing overhead and adhere to the load constraint, the reward of playing a single arm can be further formulated as $\frac{K-k_a}{K} r_{a|x}$, where K is the maximum number of users that can be connected to a single beam, k_a is the current number of users that have connected to beam a. Overall, the design of this reward function effectively guides the users to select beams of some APs with lower traffic load while maintaining a relatively high link quality.

As mentioned earlier, in our context, we can probe a *subset* of arms $\mathcal{S}_m^t \subset \mathcal{A}_m^t$ to assess the qualities of these arms. We then select the arm that yields the highest reward in $\mathcal{S}_m^t$. Let $\mathbf{r} = \{r_{a|x}\}_{a \in \mathcal{S}_m^t}$ denote the collection of rewards of arms in the probing set. The reward of probing $\mathcal{S}_m^t$ can then be formulated as $R(\mathcal{S}_m^t, \mathbf{r})$, signifying that the reward is jointly determined by the selection of subset and the individual reward of each arm in the subset:

$$R(\mathcal{S}_m^t, \mathbf{r}) = \max_{a \in \mathcal{S}_m^t} \frac{K - k_a}{K} r_{a|x_m^t}. \tag{1}$$

It is worth noting that the $\max$ function in Eq. (1) is a submodular function [6]. It satisfies the diminishing returns property – given the arm sets $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A} \subseteq \mathcal{B}$, for all arms $a \notin \mathcal{B}$, we have:

$$R(\mathcal{A} \cup m, \mathbf{r}) - R(\mathcal{A}, \mathbf{r}) \geq R(\mathcal{B} \cup m, \mathbf{r}) - R(\mathcal{B}, \mathbf{r}). \tag{2}$$

This property of the reward function aligns well with our formulated contextual combinatorial MAB problem, which is specifically tailored for submodular functions.

Given the above property, an online probing method with known expected rewards is outlined in Algorithm 1. The primary objective is to maximize the *cumulative* reward expectation over T rounds, i.e., $\sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}[R(\mathcal{S}_m^t), \mathbf{r}]$. Assuming an optimal algorithm could consistently select the best arm set $\mathcal{S}_m^{*,t}$ at every round t for each user m, the performance of our algorithm can be measured by the expected cumulative regret in Eq. (3), which quantifies the expected cumulative

difference between the maximum reward achieved by the optimal algorithm and the reward obtained by our algorithm.

$$Reg(T) = \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}[R(\mathcal{S}_m^{t,*}, \mathbf{r}) - R(\mathcal{S}_m^t, \mathbf{r})] \quad (3)$$

Hence, the objective of maximizing the expected cumulative reward is equivalent to minimizing the expected regret.

It has been proven that maximizing a submodular set function with known reward expectation is NP-hard [9]. However, a greedy probing algorithm has been proposed in [10] that guarantees achieving no less than $(1 - 1/e)$ of the optimal solution. Therefore, as described in Algorithm 1, beams can be sequentially selected based on their marginal reward to ensure an asymptotic optimality. Given that no polynomial time algorithm can achieve a better approximation for the submodular function maximization problem, our objective is to find an algorithm that achieves sublinear $(1-\frac{1}{e})$-approximation regret, as formulated in Eq. (4):

$$Reg(T) = (1-\frac{1}{e}) * \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}[R(\mathcal{S}_m^{t,*}), \mathbf{r}] - \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}[R(\mathcal{S}_m^t, \mathbf{r})]. \quad (4)$$

Algorithm 1 Online Probing Algorithm

Input: arm set $\mathcal{A}$, reward function R, budget B
Output: super arm $\mathcal{S}$

1: $\mathcal{S} \leftarrow \varnothing$
2: $i = 0$
3: **while** $i \leq B$ **do**
4: $\quad m = \underset{m \in \mathcal{A} \setminus \mathcal{S}}{\operatorname{argmax}} R(\mathcal{S} \cup \{m\}, \mathbf{r}) - R(\mathcal{S}, \mathbf{r})$
5: $\quad \mathcal{S} = \mathcal{S} \cup \{m\}$
6: $\quad i = i + 1$

III. Contextual Combinatorial Beam Management

Building upon the problem formulation as discussed in Sec. II, this section introduces a novel contextual combinatorial MAB approach for beam management in mmWave wireless networks. In practical scenarios, it is infeasible to have prior knowledge of the expected rewards for arms. Consequently, direct application of Algorithm 1 is not viable. Instead, we aim to learn the expected values of arms using a contextual combinatorial MAB framework as illustrated in Sec. II. Such a framework was originally designed for general bandit problems with submodular reward function [11]. However, our approach differs by incorporating both arm context (beam correlation) and environment context (location-aware link qualities). Additionally, we subtly integrate a beam selection scheme to enhance rewards during the exploration period.

Algorithm 2 summarizes our CCBM framework using online probing to achieve the expected rewards for beam arms. Initially, the environment context space X is divided into a uniform grid set $X' = \{x_1, x_2, ..., x_n\}$, with n representing the total number of grids into which the space is partitioned. At each time step t, when a user location x_m^t is observed, it is mapped to the corresponding grid x in X' to which it belongs (Lines 4-5). In addition to managing the environment context space, we partition the arm context space $\mathcal{X} = [0, 2\pi]$, where it is first normalized to $[0, 1]$, and then divided into h_T *hypercubes* with a size of $\frac{1}{h_T}$.

In specific, at each time step t, for each user m, the algorithm observes the environment context x_m^t (e.g., the user's location) and then maps it to the corresponding grid $x \in X'$. For each arm a in $\mathcal{A}_m^t$ with arm context O_a, the algorithm determines a hypercube $p_a \in \mathcal{X}$ such that $O_a \in p_a$ holds. The collection of the hypercubes at time slot t is denoted as $\mathbf{p}^t = p_{a a \in \mathcal{A}_m^t}$ (Lines 10-11). Subsequently, the algorithm identifies the hypercubes $p_a \in \mathbf{p}^t$ that are explored less frequently based on the following criteria:

$$P^{ue,t} = \{p_a \in \mathbf{p}^t | x \in X', a \in \mathcal{A}^t, C^t(p_a|x) < K(n_x)\}, \quad (5)$$

where $C^t(p_a|x)$ is a counter that keeps track of the number of times the arms within the hypercube p_a are selected when the user location is mapped to x during time periods $1, 2, .., t-1$. The $K(n_x)$ is a deterministic, monotonically increasing control function, and n_x represents the number of times the grid x has been visited in the previous time periods.

Next, the algorithm determines whether to explore or exploit based on the number of arms located in under-explored hypercubes. If the set of under-explored arms is non-empty, the algorithm enters an exploration phase. Let q be the size of under-explored arm set. If the under-explored arm set contains at least B arms, i.e., $q \geq B$, we employ an arm selection scheme called *attention-based selection* (Lines 23-28) instead of randomly selecting arms as in prior MAB works.

Attention-based Selection: Let $\mathcal{Z}$ be the set of arms in the under-explored arm set with $C^t(p_a|x) = 0$, indicating that at grid x, the hypercube to which arm a belongs has never been chosen until time period $t-1$. If $|\mathcal{Z}| \geq B$, then randomly select B arms from $\mathcal{Z}$. If $0 < |\mathcal{Z}| < B$, select all arms in the set $\mathcal{Z}$ and randomly select other under-explored arms. The rationale behind this step is intuitive: If we only randomly select arms without attentions, there is a possibility that certain hypercubes providing good rewards may not be identified in the initial rounds, leading to sub-optimal exploration.

To be specific, when $|\mathcal{Z}| = 0$, indicating that the under-explored hypercubes have been chosen at least once, our algorithm first identifies the arm a_m^{t-1} chosen for the user m in the last time step $t-1$. Since we are considering a continuous movement (action space), for a single user, the location at time step t should be close to the location at time step $t-1$. Therefore, we can still assume a_m^{t-1} is a good candidate arm that can provide satisfied rewards at round t, and it will be included in the probing set $\mathcal{S}_m^t$, while the other arms are chosen randomly. In this way, we strategically incorporate attention-based exploitation into the exploration phase.

In some cases, if the under-explored arm set contains fewer than B elements, i.e., $q \leq B$, then the algorithm selects

all q arms (Lines 12-16). The remaining arms are selected sequentially by exploiting the estimated rewards as follow:

$$a = \underset{a \in \mathcal{A}_m^t \setminus \mathcal{S}_m^t}{\operatorname{argmax}} R(\mathcal{S}_m^t \cup \{m\}, \hat{\mathbf{r}}) - R(\mathcal{S}_m^t, \hat{\mathbf{r}}), \tag{6}$$

where $\hat{r}$ is used to denote the sampled reward of each arm a. If there are no under-explored arms, all B arms will be selected based on Eq. (6).

Algorithm 2 Contextual Combinatorial Beam Management

Input: user number M, arm set $\mathcal{A}_m^t$, reward function R, budget B, time horizon T, control function $K(n_x)$, arm context space $\mathcal{X}$, grid set X

Initialization: $\forall x \in X, n_x = 0$

$\forall p_a \in \mathcal{X}, C^t(p_a|x) = 0, \hat{r}(p_a|x) = 0$

1: **for** $t = 1, 2, \ldots, T$ **do**
2: **for** $m = 1, 2, \ldots, M$ **do**
3: $\mathcal{S}_m^t = \emptyset$
4: Receive the client position information x_m^t
5: Map the position context to grid $x \leftarrow x_m^t$
6: $n_x = n_x + 1$
7: **if** $t > t_\tau$ **then**
8: $\mathcal{S}_m^t \leftarrow$ select $\frac{B}{2}$ arms based on Eq. (6)
9: **else**
10: Find $\mathbf{p}^t$, such that $\forall a \in \mathcal{A}_m^t$, $O_a \in p_a, p_a \in \mathcal{X}$
11: Compute the under-explored hypercubes $P^{ue,t}$ using Eq. (5)
12: **if** $P^{ue,t} = \emptyset$ **then**
13: $\mathcal{S}_m^t \leftarrow$ select B arms based on Eq. (6)
14: **else**
15: **if** number of unexplored arms $q < B$ **then**
16: $\mathcal{S}_m^t \leftarrow$ select all q arms and the other $B - q$ arms based on Equation.6
17: **else**
18: run Attention-based Selection(void)
19: **for** each arm $a \in \mathcal{S}_m^t$ **do**
20: observe the quality $r_{a|x}$ of a
21: update $\hat{r}(p_a|x) = \frac{\hat{r}(p_a|x)C^t(p_a|x) + r_{a|x}}{C^t(p_a|x)+1}$
22: update counter $C^t(p_a|x) = C^t(p_a|x) + 1$

Function: Attention-based Selection(void):

23: **if** $|\mathcal{Z}| \geq B$ **then**
24: $\mathcal{S}_m^t \leftarrow$ randomly select arms from $\mathcal{Z}$
25: **else if** $0 < |\mathcal{Z}| < B$ **then**
26: $\mathcal{S}_m^t \leftarrow$ select all arms in $\mathcal{Z}$ and other arms randomly
27: **else**
28: $\mathcal{S}_m^t \leftarrow \{a_m^{t-1}\} \cup \{B - 1$ arms randomly selected$\}$

Since we consider a mmWave wireless network scenario with fixed APs placed in the space, it is intuitive that after a certain number of rounds of exploration, we can have a relatively comprehensive knowledge of the network condition. Thus, it will be more rewarding to perform exploitation after a certain time step. To this end, we incorporate a *early stopping criterion* to guide the algorithm into an exploitation phase (Lines 7-8).

Early Stopping Criterion: We assume that after a time threshold t_τ the algorithm enters a pure exploitation period. Since arms yield different rewards in terms of different grids x, if all the grids have been visited by users several times, the network condition can be well revealed. Thus, we set t_τ equal to the number of grids n across the space. It is also worth noting that we reduce the size of the probing set to $\frac{B}{2}$ during the pure exploitation period. Numerical results in Sec. V will show that this added criterion greatly reduces the beam search overhead while maintaining a competitive reward.

IV. Theoretical Analysis

In this section, we provide a theoretical analysis of the regret bond using our CCBM approach. The upper bound is derived under the principle that arms belonging to similar context space should have similar expected reward values.

Assumption 1. *(Lipschitz-continuous) There exists $C > 0$ such that for any arm a, a' with arm context $O_a, O_{a'} \in \mathcal{X}$, we have $|r_a - r_{a'}| \leq C||O_a - O_{a'}||_1$.*

Assumption 2. *(Bounded Reward) The reward of each arm is bounded by $0 < r < r^{max}$.*

We set the $h_T = \lceil T^{\frac{1}{4}} \rceil$ for the arm context partition and $K(n_x) = n_x^{\frac{1}{2}} log(n_x)$ as the control function to identify the under-explored arm hypercubes. Then, the regret can be bounded as follow:

Theorem 1. *Let $h_T = \lceil T^{\frac{1}{4}} \rceil$ and $K(n_x) = n_x^{\frac{1}{2}} log(n_x)$, if Assumptions 1 and 2 hold true, the regret $R(T)$ is bounded by:*

$$R(T) \leq (1 - \tfrac{1}{e}) B r^{max} 2M (M^{\frac{1}{2}} log(MT) T^{\frac{3}{4}} + T^{\frac{1}{4}}) + (1 - \tfrac{1}{e}) B r^{max} M \binom{|\mathcal{A}_m^t|}{B} \tfrac{\pi^2}{3} + (3BL + \tfrac{8}{3} B (r^{max} + L)) M T^{\frac{3}{4}}.$$

Proof. The regret $R(T)$ can be divided into the summands:

$$\mathbb{E}[R(T)] = \mathbb{E}[R_{explore}(T)] + \mathbb{E}[R_{exploit}(T)],$$

where the term $\mathbb{E}[R_{explore}(T)]$ is the regret due to the exploration process, and the term $\mathbb{E}[R_{exploit}(T)]$ corresponds to the regret in the exploitation phase. We first derive a bound on $\mathbb{E}[R_{explore}(T)]$. According to Algorithm 2, the set of under-explored hypercubes $P_T^{\text{ue},t}$ is non-empty during the exploration phase, which implies that there exists at least one hypercube p with $C^t(p|x) \leq K(n_x) = n_x^{\frac{1}{2}} \log(n_x)$. Because we only explore in the first t_τ rounds, $n_x < M t_\tau < MT$ holds. Certainly, there can be a maximum of $\lceil (MT)^{\frac{1}{2}} \log(MT) \rceil$ exploration phases in which p is under-explored. Given h_T hypercubes in the partition and a total of M users, the upper limit for exploration phases is $h_T M \lceil (MT)^{\frac{1}{2}} \log(MT) \rceil$. Owing to the submodularity of reward function and its bounded nature, the maximum regret for an incorrect selection in one exploration phase is constrained by $(1 - 1/e) B r^{\max}$. Therefore, we have

$$\begin{aligned} \mathbb{E}[R_{explore}(T)] &\leq (1 - \frac{1}{e}) B r^{\max} h_T M \lceil (MT)^{\frac{1}{2}} \log(TM) \rceil \\ &= (1 - \frac{1}{e}) B r^{\max} M \lceil T^{\frac{1}{4}} \rceil \lceil (MT)^{\frac{1}{2}} \log(TM) \rceil. \end{aligned}$$

Given $\lceil T^{\frac{1}{4}} \rceil \leq 2T^{\frac{1}{4}}$, we can further bound the maximum regret as:

$$\mathbb{E}[R_{explore}(T)] \leq (1 - \frac{1}{e}) B r^{\max} 2M(M^{\frac{1}{2}} T^{\frac{3}{4}} \log(MT) + T^{\frac{1}{4}}).$$

Applying similar reasoning, the regret bound during the exploitation phase can also be deduced. We omit it here due to the space limitation, but all proof details can be found in our supplementary technical report [12].

□

In summary, the leading order of the cumulative regret is $O(T^{\frac{3}{4}} log(T))$, indicating a sublinear growth over the time horizon T. This implies that our CCBM scheme exhibits asymptotic optimality and converges toward optimal strategy.

V. Performance Evaluations

In this section, we evaluate the performance of our CCBM approach through comprehensive numerical simulations. We begin by outlining the simulation setup, followed by a comparison of our algorithm with other baseline schemes in terms of multiple performance metrics.

A. Network Settings

We consider a 3-D indoor network scenario with a size of 40m×40m×3m, consisting of wooden tables, wooden chairs, metal cabinets, and 15 humans randomly moving at a speed of 0.8m/s to emulate the dynamic obstacles. In this setup, we place four 60GHz mmWave APs randomly in the space at a height of 2.9m. Specifically, each AP as the wireless transmitter is equipped with 8 orthogonal beam patterns with equal beam widths, covering a 360° azimuth. The environment context X is uniformly divided into 1600 (40×40) small grids, each measuring 1m^2. In particular, the neighbouring beams are regarded as arms with the similar context, hence the beams from the same AP are categorized into 4 hypercubes, resulting in a total of 16 hypercubes. We employ the commercial ray tracer *Wireless Insite*® [13] to generate the realistic network environment and mmWave signal profiles. Additionally, we introduce noise following a normal distribution $\mathcal{N}(0, 5)$ into the obtained received signal strength (RSS) values to account for potential measurement errors in the context information, mirroring the conditions encountered in practical scenarios.

Baseline Algorithms: We conduct a thorough performance analysis by comparing our algorithm with the following baseline schemes:

- **Optimal scheme**. This algorithm relies on an oracle search, indicating a priori knowledge of the expected reward $\mu_{a|x}$ for each arm a within $\mathcal{A}_m^t$ at grid x. It always selects an optimal subset $\mathcal{S}^*$ to probe the best beam at each time step, offering an upper-bound performance for comparison with other feasible schemes.
- **UCB-based scheme**. This state-of-the-art scheme, proposed in [6], employs an upper confidence bound approach. In each time step, it strategically selects B arms with the highest estimated upper confidence bounds on their expected rewards.
- **CC-MAB scheme**: We add the basic contextual MAB algorithm from [11] as the comparison point. The key distinction with our CCBM approach lies in the fact that CC-MAB incorporates a completely randomized arm selection process during the exploration phase.

B. Cumulative Regret for Beam Selection

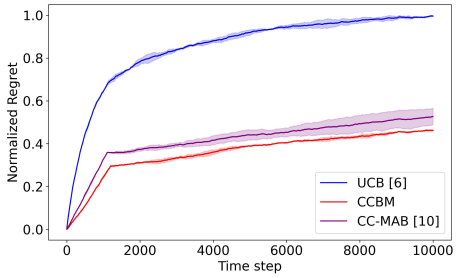

Fig. 1. Comparison of regret among different algorithms.

To evaluate the disparity between the total reward achieved by a practical probing algorithm and the optimal reward attainable by consistently selecting the best beam, Fig. 1 shows the cumulative regret over time for three distinct algorithms. Obviously, our proposed CCBM exhibits superior performance compared to the other two baselines. Specifically, the UCB-based scheme exhibits the highest regret, consistently maintaining a curve above the others throughout the time horizon. This can be attributed to the fact that it does not account for the properties of a submodular reward function. Besides, our CCBM scheme achieves a lower regret than CC-MAB, which demonstrates the effectiveness of incorporating exploitation into the exploration phase via our attention-based selection. Furthermore, a noticeable turning point occurs at time step around 1600, corresponding to the implementation of our early stopping criterion that prevents extensively useless searches.

C. Beam Management Overhead and Network Throughput

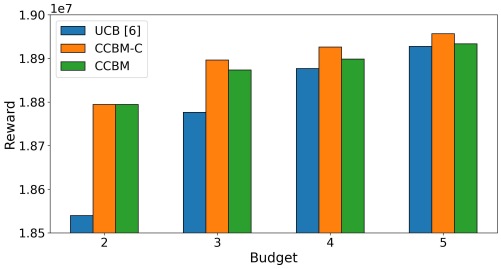

Fig. 2. Reward under different beam probing budgets.

In this section, we evaluate the performance in terms of beam management overhead and average user throughput in mmWave networks. First, we consider both our CCBM and its variant, CCBM-C, based on the rewards obtained under different probing budgets B. The higher B implies a potential manegement overhead. The only difference between the two schemes is that CCBM-C constantly probe beams with a budget of B while CCBM searches a subset of beams with a size of $\frac{B}{2}$ in early stopping phase. As depicted in Fig. 2, an increase in the budget leads to an augmentation in rewards for

all three algorithms. This trend is attributed to the fact that a larger budget enhances the probability of encompassing the optimal beam, thereby increasing the likelihood of identifying the most advantageous beam. Under different budgets, the CCBM-C algorithm consistently secures the highest rewards. Concurrently, the CCBM achieves rewards marginally lower than those of CCBM-C while utilizing only *half* of the budget. This demonstrates how efficiently the CCBM algorithm can leverage limited resources to optimize rewards. In this regard, we conclude that both CCBM and CCBM-C can obtain a higher reward at lower overhead, indicating that CCBM is effective in achieving high performance with a constrained beam search budget.

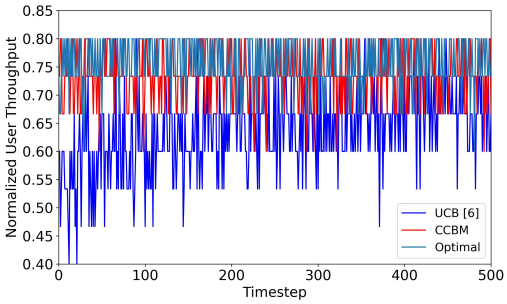

Fig. 3. Comparison of average user throughput among different schemes.

Fig. 3 compares the average user throughput of our proposed CCBM against the UCB-based scheme from [6] and the Optimal method with an oracle search. We observe that the throughput of the Optimal scheme is always the highest, benefiting from its priori knowledge about the expected reward of each arm. Consistently, the throughput of CCBM is maintained at a relatively high level, close to the optimal results and significantly surpassing the results of the UCB-based scheme. Additionally, the throughput of CCBM exhibits a lower variance than that of UCB-based scheme, signifying its greater stability. Such consistently higher throughput performance underscores the robustness and efficiency of our CCBM scheme in dynamic network environments.

Lastly, load balancing is another critical aspect addressed by our CCBM approach. To evaluate the load balancing performance, we utilize the maximum load utilization L_{max} as the metric to qualitatively reflect network congestion, where a higher L_{max} indicates more server congestion and unbalanced resource usage. The value of L_{max} corresponds to the maximum load among all beams in the network. Fig. 4 shows the average L_{max} across randomly located users. As expected, the Optimal scheme achieves the lowest L_{max} value, while our CCBM approach performs closely to the optimal results. It is observed that the gap is especially smaller under higher density of users in the network, which validates the load balancing capability owing to the strategical reward function design.

VI. Conclusion

This paper introduces a contextual combinatorial beam management scheme for the joint AP and beam selection in mmWave wireless networks. It incorporates an early stopping criterion and an attention-based mechanism to mitigate excessive search during online probing. Theoretical analysis establishes its asymptotic optimality by setting an upper bound on cumulative regret. Additionally, a carefully designed reward function takes into account load balancing among APs, aiding in selecting an global-view optimal AP and beam and thereby enhancing overall network performance. Through a series of simulations and theoretical analyses, CCBM has demonstrated its superiority over other baseline schemes in optimizing AP-beam selection in dense mmWave wireless networks.

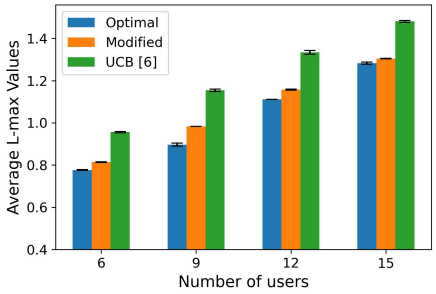

Fig. 4. Comparison of L-max among different schemes.

Acknowledgment

This research was supported by the National Science Foundation through Award CNS–2312138 and CNS–2312139.

References

[1] D. Zhang, P. S. Santhalingam, P. Pathak, and Z. Zheng, "CoBF: Coordinated beamforming in dense mmwave networks," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2023.

[2] Y. Liu, Q. Hu, and D. M. Blough, "Joint link-level and network-level reconfiguration for mmwave backhaul survivability in urban environments," in *Proceedings of International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2019.

[3] M. Polese, F. Restuccia, and T. Melodia, "DeepBeam: Deep waveform learning for coordination-free beam management in mmWave networks," in *ACM MobiHoc*, 2021.

[4] A. Asadi, S. Müller, and G. H. Sim et al., "FML: Fast machine learning for 5G mmWave vehicular communications," in *IEEE INFOCOM*, 2018.

[5] 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, 2019.

[6] T. Xu, D. Zhang, and Z. Zheng, "Online learning for adaptive probing and scheduling in dense wlans," in *IEEE INFOCOM*, 2023, pp. 1–10.

[7] Y. Liu and D. M. Blough, "Environment-aware link quality prediction for millimeter-wave wireless LANs," in *ACM International Symposium on Mobility Management and Wireless Access*, 2022, pp. 1–10.

[8] Z. Li, M. Chen, G. Li, and Y. Liu, "Spatial-temporal attention-based mmWave link quality prediction under dynamic blockages," in *Proc. of IEEE Global Communications Conference, to appear*, 2023.

[9] A. Goel, S. Guha, and K. Munagala, "Asking the right questions: Model-driven optimization using probes," in *Proceedings of ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2006.

[10] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions–I," *Mathematical programming*, vol. 14, pp. 265–294, 1978.

[11] L. Chen, J. Xu, and Z. Lu, "Contextual combinatorial multi-armed bandits with volatile arms and submodular reward," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] Z. Li, "Technical report for CCBM." [Online]. Available: https://github.com/lzzz1998/CCBM

[13] Remcom Inc. [Online]. Available: https://www.remcom.com/wireless-insite-em-propagation-software