Debamita Ghosh\textsuperscript{1}, Manjesh K. Hanawal\textsuperscript{2}, and Nikola Zlatanov\textsuperscript{3}

\textsuperscript{1}IITB-Monash Research Academy, IIT Bombay
\textsuperscript{2}MLioNS Lab, IEOR, IIT Bombay
\textsuperscript{3}Innopolis University

March 07, 2024
UB3: Fixed Budget Best Beam Identification in mmWave Massive MISO via Pure Exploration Unimodal Bandits

Debamita Ghosh, IITB-Monash Research Academy, IIT Bombay, India
Manjesh K. Hanawal, MLioNS Lab, IEOR, IIT Bombay, India
Nikola Zlatanov, Innopolis University, Russia

Abstract— One of the core problems in millimeter wave (mmWave) massive multiple-input-single-output (MISO) communication systems, which significantly affects the data rate, is the misalignment of the beam direction of the transmitter towards the receiver. In this paper, we investigate strategies that identify the best beam within a fixed duration of time. To this end, we develop an algorithm, named Unimodal Bandit for Best Beam (UB3), that exploits the unimodal structure of the mean received signal strength as a function of the available beams and identifies the best beam within a fixed time duration using pure exploration strategies. We derive an upper bound on the probability of misidentifying the best beam, and we prove that the upper bound is of the order \( O \left( \log_2 K \exp \left\{ -\frac{\alpha n A}{K \log K} \right\} \right) \), where \( K \) is the number of beams, \( A \) is a problem-dependent constant, and \( \alpha n \) is the number of pilots used in the channel estimation phase. In contrast, when the unimodal structure is not exploited, the error probability is of order \( \mathcal{O} \left( \log_2 K \exp \left\{ -\frac{\alpha n A}{K \log K} \right\} \right) \). Thus, by exploiting the unimodal structure, we achieve a much better error probability, which depends only logarithmically on \( K \). We demonstrate that UB3 outperforms the state-of-the-art algorithms through extensive simulations.

Index Terms—mmWave, transmit beamforming, multi-armed bandits, pure exploration

I. INTRODUCTION

The millimeter wave (mmWave) band, with a spectrum ranging from 30 GHz to 300 GHz, offers an abundant bandwidth and can support data rates of gigabits per second. Significant efforts are underway in standardising mmWave systems, such as IEEE 802.11ad [2], IEEE 802.15.3c [3], and the ongoing IEEE 802.11ay [4]. Moreover, the road to commercialization of 5G networks with mmWave technology has been prepared by extensive field trials [5].

Although mmWave massive multiple-input-multiple-output (MIMO) systems offer higher data rates, they come with a set of challenges. Small wavelengths of mmWave allow antennas with small form factors to be densely packed, which can focus the energy of the transmit signal in specific directions, forming sharp beams. This gives rise to one of the main challenges in mmWave massive MIMO, which is that the transmitter’s beam needs to be focused towards the position of the receiver in order for the received signal to have a high signal-to-noise ratio (SNR). A few degrees of misalignment of the beam direction of the transmitter towards the receiver can reduce the data rate from gigabits per second to a few megabits per second, thereby diminishing the gain obtained from the wide spectrum in mmWave massive MIMO systems [6]–[9].

In this work, we have considered the problem of aligning the transmit beam to the receiver’s position of a mmWave massive multiple-input-single-output (MISO) system to achieve the best SNR/data rate, a problem known as transmit beamforming (TBF). The TBF problem is critical to building better 5G communication networks with mmWave massive MISO systems. Our goal in this paper is to contribute towards improving the TBF by proposing an algorithm that learns the best beam of the transmitter within a given time duration with high probability and provide theoretical guarantee of our algorithm.

As mmWave massive MISO can only point the beam in different directions, finding the optimum direction effectively necessitates sweeping the beam through all feasible directions, where the sweep has to be done in steps of the beamwidth. One naive approach to performing transmit beamforming is an exhaustive search of all available directional beams. Since the beamwidth for mmWaves can be as low as 1°, and the number of beams can be quite large for mmWave massive MISO systems, the exhaustive search approach can significantly increase the beam training overhead [10], [11].

We address the TBF problem in mmWave massive MISO using the framework of the fixed-budget pure exploration Multi-Armed Bandit (MAB) framework, where pure exploration is performed within a given time duration. The TBF problem has already been addressed using the MAB framework, which uses cumulative regret minimization algorithms that balance exploration and exploitation to find the best beam, see [7], [12], [13]. However, due to the continuous exploration, sub-optimal beams can be used for data transfer, resulting in outages. Moreover, in the TBF problem, the goal is to identify the best beam during an initial channel estimation phase of fixed duration and only then start the data transfer between the transmitter and the receiver. Therefore, in this paper, we propose an efficient learning strategy that identifies the best beam within a fixed time duration set by the channel estimation time duration phase (i.e., budget).

In this work, we exploit the structural properties of the received signal strengths (RSS) as a function of the properly discretized beams to accelerate the learning process.
Several studies validate that the RSS as a function of the discrete beams in mmWave massive MISO systems follows a multi-modal structure, with one peak corresponding to the line-of-sight (LOS) path and other peaks corresponding to the non-line-of-sight (NLOS) paths [7], [12]. In mmWave massive MISO systems, the power of the LOS component is around 10 dB higher than the total power of the NLOS components, according to channel measurement campaigns in [14], [15]. Therefore, often there is one dominant peak for the LOS path, and the peaks due to the NLOS components are negligible compared to the dominant peak. As a result, the multimodal structure of the RSS behaves as an unimodal structure. Bandits with an unimodal structure are well studied in the literature in the cumulative regret setting with optimal algorithms provided in [16]–[18]. However, the fixed-budget pure exploration bandit with unimodal structure is not well studied and optimal algorithms are not known. This work fills in this gap in the literature. In this work, we develop a new fixed-budget pure exploration algorithm that exploits the unimodal structure of the RSS and, in addition, we also provide theoretical guarantees for our proposed algorithm. We provide a proposed algorithm named Unimodal Bandit for Best Beam (UB3) and is based on the idea of sequential elimination of the sub-optimal beams. For a fixed $K$ number of beams, UB3 achieves an error probability of the order of $O(\log K \exp(-\alpha n \Delta^2))$, where $\alpha n$ is the number of pilots used in the channel estimation phase and $\Delta$ is the minimum gap between two successive mean RSS of the beams. When no unimodal structure is assumed (unstructured), the best known achievable error probability is $O(\log K \exp(-\frac{\alpha n \Delta^2}{\kappa \log K}))$ [19]. Thus, by exploiting the unimodal structure, we achieve a much better error probability, which depends only logarithmically on $K$. We demonstrate this behaviour by establishing a lower bound for any unimodal bandits.

In summary, our contributions are as follows:

- We study the problem of TBF in mmWave massive MISO systems as a fixed-budget pure exploration multi-armed bandit problem.
- We exploit the unimodal structure of the RSS as a function of the beams and propose an algorithm named Unimodal Bandit for Best Beam (UB3) that identifies the best beam with a high probability in a fixed time duration.
- We provide an upper bound on the error probability of UB3 in identifying the best beam and show that the upper bound depends only logarithmically on the number of beams. We demonstrate this by establishing a lower bound for unimodal bandits.
- We perform extensive simulations to validate the superior performance of UB3 as compared to other state-of-the-art algorithms.

### A. Related Work

There is a growing interest in mmWave massive MIMO systems in academia and industry, so various aspects of these systems are being studied. For the recent advances in the mmWave massive MIMO/MISO systems, we refer to surveys [8], [20], [21].

Several approaches are proposed to solve the TBF problem. The compressive sensing-based methods [22]–[25] utilize the sparse characterization of the mmWave massive MIMO channel to learn the best beam of the transmitter that aligns towards the receiver. These methods work well when accurate channel state information (CSI) is available. The authors of [26] propose an Optimized Two-Stage Search (OTSS) algorithm where a suitable candidate set of transmit beams is identified in the first stage based on the received signal profile, and in the second stage, the best beam from the candidate set is selected using additional measurements. Codebook-based hierarchical methods are proposed in [27], [28], but require substantial channel knowledge for beamforming and are computationally complex. The authors of [29] utilize the location information to perform fast beamforming, which is feasible only when the location information of the receiver is available at the transmitter. The authors of [30] use Kalman filters to detect the angles of arrivals and departures to track the receiver. Recently, machine learning [31] and deep learning [32] methods have been used for transmit beamforming, which requires substantial offline channel observations and is computationally complex.

The Multi-Armed Bandit (MAB) approach has been widely used to tackle the TBF problem [7], [12], [13], [33]–[39]. The authors in [7], [12], [13], [33], [37]–[39] employ the MAB framework on regret minimization setting and the authors in [34]–[36] employ the MAB framework on pure exploration setting. The authors in [7] develop an algorithm named Unimodal Beam Alignment (UBA) that exploits the unimodal structure of the amount of energy received, where the amount of the energy received can be approximated as an unimodal function of misalignment of the transmitter’s and receiver’s beam of the mmWave massive MIMO system. The algorithm is built based on the Optimal Sampling for Unimodal Bandits (OSUB) algorithm [16] by adding stopping criteria, where the stopping criteria are based on the threshold, which can be different under various environmental conditions. In [12], the transmitter scans the available beams where the receiver keeps omnidirectional beams of a mmWave MIMO system, and the authors developed Hierarchical Beam Alignment (HBA) algorithm that exploits the unimodal/multimodal structures of the beam signal strengths to narrow down on the optimal transmit beam. HBA has shown better performance for beam identification than UBA. The authors in [13] develop an algorithm named Adaptive Thompson sampling (ATS) that aims to maximize the cumulative rate obtained over a fixed duration using the Thompson sampling algorithm of mmWave MIMO system. Unlike the UBA and HBA, ATS is based on the discount factor to emphasize current reward and de-emphasize past reward. However, in reality, the discount factor is difficult to set and may result in selecting sub-optimal beams. The authors in [33] develop multiple novel Thompson Sampling (TS)-based algorithms that maximise the overall throughput by selecting a beam codebook that balances between beamwidth and beam alignment overhead. In [37], the authors proposed a two-stage approach for a Thompson sampling (TS)-based MAB algorithm. In the first stage, a group of data rates that could include the optimal modulation and coding scheme (MCS) is identified, and then, in the second stage, the optimal MCS is searched.
only in the reduced search space. However, the authors did not offer theoretical guarantees for their algorithm. As this paper is unrelated to our work, the challenge of grouping beams makes its applicability difficult in our context. The authors in [38] proposed Thompson sampling for unimodal bandits (TS-UB) algorithm, which makes decisions according to posterior distribution only in the beam’s neighborhood with the highest empirical mean estimate at each step instead of exploring the entire decision space. In [38], the TS-UB algorithm significantly improves the Unimodal Thompson Sampling (UTS) algorithm proposed by [38] for unimodal bandits. However, all these works, i.e., [7], [12], [13], [33], [37]–[39], optimize the exploration and exploitation based on regret minimization. However, due to continuous exploration, suboptimal beams can be used for data transfer, resulting in outages. Notably, strategies that minimise regret perform suboptimally when the objective is best beam identification in a pure-exploration setup [40]. This is because the algorithms for regret minimization setting discourage exploration beyond a reasonable point to optimize regret and, therefore, can lead to a suboptimal beam. Moreover, in the TBF problem, it is essential to choose the best beam during channel estimation phase with fixed time duration before allowing communication to begin between the transmitter and the receiver. Therefore, the better metric is the best beam identification in a pure-exploration setup, which is explored in [34], [35] and [36].

The authors of [34] propose a pure exploration strategy named Hierarchical Optimal Sampling of Unimodal Bandits (HOSUB) that exploits the benefits of hierarchical codebooks and the unimodality structure of the beam signal strengths to achieve fast beam steering of mmWave MISO systems. Simulations show better performance of HOSUB compared to HBA and a large reduction in computational complexity. However, the authors in [34] did not provide any theoretical guarantees on their proposed algorithm. The authors in [35] propose Two Phase Heteroscedastic Track-and-Stop (2PHTS) algorithm that exploits the correlation and heteroscedastic property among the beams of mmWave massive MISO system. The authors in [36] propose Successive Subtree Elimination (SSE) that exploits the benefits of hierarchical codebooks and the unimodality structure of the beam signal strengths of mmWave massive MISO. All the works [34]–[36] are based on fixed-confidence pure-exploration setting that uses the benefits of hierarchical codebooks and are computationally complex. On the other hand, the fixed-budget pure exploration strategies are more suitable for the transmit beamforming problem since the exploration can be completed during the channel estimation phase. Therefore, we focus on developing a strategy that identifies the best beam in a fixed-budget pure-exploration MAB setup [19], [41], by exploiting the unimodal structure of the RSS over the beams that eliminates the sub-optimal beams and narrow the beam search space quickly. We provide an algorithm that is simple to implement and has low computational complexity for mmWave massive MISO system. To our knowledge, this has not been studied in 5G networks with mmWave systems.

This paper is organized as follows. The system and the channel models for the mmWave MISO communication system are given in Sec. II. The communication scheme is formulated in III. The proposed algorithm for learning the best transmit beam towards the receiver is given in Sec. IV and its theoretical guarantee is provided in Sec. V. Numerical simulations of the proposed algorithm are provided in Sec. VI. Finally, Sec. VII concludes the paper.

II. SYSTEM AND CHANNEL MODELS

In this section, we discuss the system and channel model of the mmWave massive MISO system for which we propose our algorithm. We follow the setup and notation provided in [12]. To this end, let $\mathbb{R}$ and $\mathbb{C}$ denote the set of real and complex numbers, respectively.

A. System Model

We consider a mmWave massive MISO communication system model comprised of a transmitter and a receiver, as shown in Fig. 1. We assume that the transmitter is equipped with a single RF chain connected to $M$ antennas in the form of a linear array, where each antenna is equipped with an analogue phase shifter. The antennas are evenly spaced by a distance $D = \lambda / 2$, where $\lambda$ is the carrier wavelength. We assume that the receiver is equipped with a single antenna. We consider the far-field region where the transmitter and the receiver are sufficiently far apart.

B. Channel Model

We assume a block fading channel model, where each block has duration $T$. Thereby, we assume that the channel between the transmitter and the receiver remains constant during one block and changes only in the next block. For each block, we model the channel between the transmitter and the receiver using the well-known channel model provided in [15]. More specifically, we assume that the channel between the transmitter and the receiver consists of $Q$ paths, where one is the dominant LOS path and the rest $Q - 1$ are NLOS paths. Let $\theta_q$ denote the angle of arrival of the $q^{th}$ path of the channel [12], where $q = 0, 2, \ldots, Q - 1$. Then, the channel between the transmitter and the receiver can be modelled as a complex vector, denoted by $h \in \mathbb{C}^{M \times 1}$, and given by

$$h = g_0 a(v_0) + \sum_{q=1}^{Q-1} g_q a(v_q),$$

where $a(v_q) \in \mathbb{C}^{M \times 1}$ denotes the vector of complex sinusoids at spatial angle $v_q = \cos \theta_q$ and is given by

$$a(v_q) = \begin{bmatrix} 1, e^{j 2\pi D v_q}, e^{j 2\pi 2D v_q}, \ldots, e^{j 2\pi (M-1) v_q} \end{bmatrix}^T.$$
The magnitude squared of the received signal in (4) induced by the $k^{th}$ beam, $y_k$, is the RSS, and is denoted by $r_k$, which is given by

$$r_k = |y_k|^2 = \left| \sqrt{P}h^H b_k s + \zeta \right|^2 = \frac{P}{M}g_0a^H(v_0)a(w_k)s + \frac{\sum_{q=1}^{Q-1} P}{M}g_qa^H(v_q)a(w_k)s + \zeta^2,$$  

where $P$ denotes the transmit power and $\zeta$ denotes the additive white complex Gaussian noise (AWGN) with mean zero and variance $\sigma^2$ at the receiver.

The magnitude squared of the received signal in (4) induced by the $k^{th}$ beam, $y_k$, is the RSS, and is denoted by $r_k$, which is given by

$$r_k = |y_k|^2 = \left| \sqrt{P}h^H b_k s + \zeta \right|^2 = \frac{P}{M}g_0a^H(v_0)a(w_k)s + \frac{\sum_{q=1}^{Q-1} P}{M}g_qa^H(v_q)a(w_k)s + \zeta^2,$$  

where $P$ denotes the transmit power and $\zeta$ denotes the additive white complex Gaussian noise (AWGN) with mean zero and variance $\sigma^2$ at the receiver.

The expected value of the RSS induced by the $k^{th}$ beam, averaged over the noise fluctuation, is referred to as the mean RSS, and is denoted by $\mu_k$ and is given by $\mu_k = \mathbb{E}[r_k]$, where the expectation is considered with respect to the noise, $\eta$, for the $k^{th}$ beam being fixed. According to [12], the expression of $\mu_k$ can be simplified as

$$\mu_k = \frac{P}{M} \left| g_0a^H(v_0)a(w_k)s + \sum_{q=1}^{Q-1} g_qa^H(v_q)a(w_k)s \right|^2 + \sigma^2.$$

The power of the LOS component is around 10 dB higher than the total power of the NLOS components, according to channel measurement campaigns in [14], [44]. Therefore, since the NLOS paths are significantly weaker than the LOS path, the mean RSS can be approximated as

$$\mu_k \approx \frac{P}{M} \frac{g_0^2}{\sin^2(\pi D(w_k - v_0)/\lambda)} + \sigma^2.$$

Finally, as a result of (4), the SNR induced by the $k^{th}$ beam, denoted by $\gamma_k$, is given by

$$\gamma_k = \frac{r_k}{\sigma^2} = \left| h^H b_k \right|^2.$$

As can be seen from (8), the SNR depends on the selected beam $b_k$, which along with the power $P$, is the only parameter over which the transmitter has control over. A more optimal $b_k$ will result in a higher SNR and a more suboptimal $b_k$ will result in a lower SNR.

III. COMMUNICATION SCHEME AND CORRESPONDING ACHIEVABLE RATE

In the following section, we first define the unimodal structure of the mean RSS as a function of the available beams. Next, we define our objective of the transmit beamforming problem, and the corresponding achievable rates for each beam.

A. Unimodal Structure

For the beam $b_k \in \mathcal{B}$, the corresponding mean RSS, $\mu_k$, as given in (7), is a function of the angular misalignment $w_k - v_0$...
and $\mu_k$ has an unimodal property [12, Thm. 1] as a function of the beam indices, which is defined in the following Lemma.

**Lemma 1.** (Unimodality): The sequence of beams $b_1, b_2, \ldots, b_{K}$, as given by (3c), induce the following sequence of corresponding mean RSS $\mu_1, \mu_2, \ldots, \mu_K$, and this sequence satisfies the following inequality

\[ \mu_1 < \mu_2 < \ldots < \mu_{K^*} > \mu_{K^*+1} > \ldots > \mu_K. \]  

(9)

**Proof.** The proof is given in [12, Theorem 1].

**Remark 1.** The spatial angle of the LOS path $\theta_0$ depends on the angle of arrival $\theta_0$ of the channel’s LOS path, which is unknown to the transmitter. We assume that $\theta_0$ is a continuous and uniformly distributed random variable over the interval $[-1, 1]$. For a given beam $b_k$ with a corresponding fixed quantity $w_k$, if $\theta_0$ equals $\theta_k - m/M$ for an integer $m$, then $\mu_k = \sigma^2$ and the unimodal structure will not hold [45]. However, this scenario occurs with zero probability, as $\theta_0$ taking the specific value $\theta_k - m/M$ within the continuous interval $[-1, 1]$ has a zero probability. Consequently, in our case, the unimodal structure persists with a probability of one.

**Objective:** Our goal is to identify the optimal beam index $k^*$ in (9) by exploiting the unimodal structure of the mean RSS, given by (9), that maximises the mean RSS. Mathematically, we have the following objective

\[ k^* = \arg \max_{b_k \in B} \mu_k. \]

(10)

**B. Achievable Rate**

We assume $n$ transmitted symbols in each block. We divide the block into two sub-blocks. The first sub-block is used for channel estimation or, equivalently, beam selection, and has a duration of $\alpha T$. The other sub-block is used for data transmission with duration $(1 - \alpha)T$. As a result, in the channel estimation sub-block $\alpha n$ pilot symbols are transmitted, and in the data transmission sub-block $(1 - \alpha)n$ data symbols are transmitted.

The rate that can be achieved by the $k^{th}$ beam, $b_k$, using this scheme, is given by

\[ R_k = (1 - \alpha) \log \left( 1 + \frac{P}{\sigma^2} |h^H b_k|^2 \right) = (1 - \alpha) \log (1 + \gamma_k). \]

(11)

Note that there is a one-to-one mapping between mean RSS $\mu_k$, the rate $R_k$, and the SNR $\gamma_k$ for all $b_k \in B$. Thereby, the maximum rate achieved by the optimal beam $b_{k^*}$ is given by

\[ R^* = (1 - \alpha) \log \left( 1 + \frac{P}{\sigma^2} |h^H b_{k^*}|^2 \right). \]

**IV. PROBLEM FORMULATION AND PROPOSED ALGORITHM**

In this section, we first formulate our transmit beamforming problem that exploits the unimodal structure of the RSS as a function of the available beams. Next, we propose an algorithm that finds the optimal beam index by exploiting the unimodal structure of mean RSS within a fixed duration of the channel estimation phase.

**A. Problem Formulation**

We employ a fixed-budget pure exploration multi-armed bandit approach [19], [46] to address the aforementioned TBF problem. Note that the time duration $\alpha T$, or equivalently the number of pilot symbols $\alpha n$, is the budget. Following the terminology of multi-armed bandits, a policy is any strategy that selects the best beam based on the given past observations, which in this case are the noisy RSSs observed by the transmitter. During the channel estimation sub-block, the user sends $\alpha n$ pilot symbols to the transmitter. The transmitter uses the $\alpha n$ pilot symbols to select the optimal beam. Thereby, let $p_t^\pi$ be the $t^{th}$ pilot symbol sent by the receiver to the transmitter during the channel estimation sub-block, for $t = 1, 2, \ldots, \alpha n$. The transmitter selects a beam $b_k \in B$ by its policy and observes the noisy RSS, $r_{k,t}$ given by (5) for the $t^{th}$ pilot symbol, which is considered as the reward. The choice of $b_k$, can depend on the beams selected in the past and their associated RSS values. As the channel remains constant during one block, the reward values observed by the beam selected for each pilot symbol are independently and identically distributed (i.i.d.) across the beams and time due to the AWGN. The reward distribution is governed by noise fluctuations and follows a fixed distribution for a fixed beam within a duration of $T$. For simplicity, we assume that the reward follows a 1 sub-Gaussian distribution for all $b_k \in B$.

For a given policy $\pi$, let $k_{\alpha n}^\pi$ denote the index of the beam produced as output from the policy $\pi$, after using $\alpha n$ pilot symbols. Let $\Pi$ denote the set of all policies of pure exploration that output a beam within a fixed budget of $\alpha n$. Then the problem in (10) can be written in a probabilistic form as the optimal policy in $\Pi$ that minimizes the error probability defined as the probability that the output beam after using $\alpha n$ budget is not the optimal beam, expressed as

\[ \min_{\pi \in \Pi} \Pr(b_{k_{\alpha n}}^\pi \neq b_{k^*}) \]  

(12)  

s.t. \ (10),

where for each policy, $\Pr(\cdot)$ is calculated with respect to the samples induced by the policy.

**B. Proposed Algorithm**

The proposed algorithm is based on the Line Search Elimination (LSE) algorithm developed in [17]. The LSE algorithm [17] leverages the unimodal characteristics in a fixed-confidence setting for a discrete set of arms. However, the fixed-confidence and fixed-budget frameworks differ and do not share transferable bounds, refer to [47] for an additional discussion. Therefore, we have suitably modified the LSE algorithm to suit the fixed-budget setting scenario in the best beam identification setup. We refer to our proposed algorithm as Unimodal Bandit for Best Beam (UB3). It is parameter-free and only requires a priori knowledge of the set of beams $B$ and the number of pilot symbols $\alpha n$ used.

In the channel estimation sub-block, UB3 divides the number of received $\alpha n$ pilot symbols into $L + 1$ batches. The $l$-th batch contains $N_l$ number of pilot symbols, where $N_l$ is set to

\[ N_l = \begin{cases} 
\frac{2L-2}{L-1} \alpha n, & \text{for } l = 1, 2 \\
\frac{2L-(l-1)}{L-1} \alpha n, & \text{for } l = 3, 4, \ldots, L + 1.
\end{cases} \]

(13)
We choose $N_l$ as per (13) is in order for $N_l$ to satisfy the following

$$\sum_{i=1}^{L+1} N_l = 2 \times \frac{2L-2}{3L-1} \alpha n + \sum_{i=3}^{L+1} \frac{2L-(l-1)}{3L-(l-2)} \alpha n.$$  (14)

At the end of the $L+1$ batch, our algorithm outputs the beam $b_{k+1}$, which we declare as the optimal beam within the set of beams $B$. Note that we have chosen $N_l$ such that after the first two batches, the number of pilot observations increases by a factor of 3/2 in each subsequent batch, which helps to distinguish between the empirical means of the remaining beams. The pseudo-code of UB3 is given in ALGO 1 and it

ALGO 1: Unimodal Bandit for Best Beam (UB3)

1. Input: $\alpha n$ and $B$.
2. Initialise: $B_1 = B$, $j_1 \leftarrow |B_1|$. Calculate $L$ from (16).
3. for $l = 1$ to $L$ do
   
   ★★ Beam Selection ★★
   
   4. $b_{k,A} \leftarrow$ First beam of $B_l$; $b_{k,D} \leftarrow$ Last beam of $B_l$;
   
   5. $b_{k,B} \leftarrow \lfloor j_l/3 \rfloor$th beam of $B_l$; $b_{k,C} \leftarrow \lfloor 2j_l/3 \rfloor$th beam of $B_l$;
   
   ★★ Beam Sampling ★★
   
   6. Transmitter collects $\frac{N_l}{4}$ number of pilots for each beam in $S_l = \{b_{k,A}, b_{k,B}, b_{k,C}, b_{k,D}\}$ and observes noisy RSS values for each beam.
   
   7. Obtain $\hat{\mu}_k^A, \hat{\mu}_k^B, \hat{\mu}_k^C, \hat{\mu}_k^D$ by (15).
   
   8. $x_l^* = \arg\max_{b_k \in S_l} \hat{\mu}_k$.
   
   ★★ Beam Elimination ★★
   
   9. if $x_l^* = k^A$ or $x_l^* = k^B$ then
   
   10. $B_{l+1} \leftarrow \{b_k \in B_l : b_{k,A} \leq b_k \leq b_{k,A}\}$, i.e., shrink to left
   
   11. else if $x_l^* = k^C$ or $x_l^* = k^D$ then
   
   12. $B_{l+1} \leftarrow \{b_k \in B_l : b_{k,B} \leq b_k \leq b_{k,D}\}$, i.e., shrink to right
   
   13. end if
   
   14. $j_{l+1} \leftarrow |B_{l+1}|$.
   
   15. end for

   ★★ Beam Output ★★
   
   16. for $l = L + 1$ do
   
   17. $B_{L+1} = \{b_{k,A}, b_{k,B}, b_{k,C}\}$;
   
   18. Transmitter collects $\frac{N_l}{4}$ number of pilots for each beam in $B_{L+1}$ and observes noisy RSS values for each beam.
   
   19. Obtain $\hat{\mu}_k^A, \hat{\mu}_k^B, \hat{\mu}_k^C$.
   
   20. Obtain $\hat{k}_{L+1} = \arg\max_{b_k \in B_{L+1}} \hat{\mu}_k$.
   
   21. end for
   
   22. Output: $\hat{k}_\alpha = \hat{k}_{L+1}$

Fig. 3: Different cases of elimination in batch $l$.

beam, $b_k \in S_l$, for batch $l$, denoted as $\hat{\mu}_k^l$ (line 7), is obtained as

$$\hat{\mu}_k^l = \frac{1}{N_l/4} \sum_{i=1}^{N_l/4} r_{k,i}, \quad \forall b_k \in S_l,$$  (15)

where $r_{k,i}$ denotes the $t$th noisy RSS observed by the transmitter from the $k$th beam in batch $l$. Based on these empirical means in (15) for $b_k \in S_l$, the algorithm eliminates at most $1/3^d$ of the beams from the set $B_l$. Specifically, if the beams $b_{k,A}$ or $b_{k,B}$ have the highest empirical means, then all the beams succeeding $b_{k,C}$ in the set $B_l$ are eliminated (line 11, 12). Fig. 3 gives a pictorial representation of the elimination of beams in two possible cases. The remaining set of beams is then transferred to the next batch. In batch $L + 1$, we are left with three beams. For each beam, $N_{L+1}/3$ pilots are observed, and the beam with the highest empirical mean is returned as the output of the algorithm (lines 16-22).

Remark 2. In every batch, UB3 narrows down its set of available beams for the next batch where the optimal beam $b_{k^*}$ lies with high probability, and the unimodal structure ensures that the other beams have a low probability of having $b_{k^*}$, and therefore, UB3 eliminates those beams. Also note that the beams between $b_{k,A}$ and $b_{k,B}$ or the beams between $b_{k,C}$ and $b_{k,D}$ are eliminated in each batch, and the beams between $b_{k,A}$ and $b_{k,D}$ always survive.

After batch $l = 1, 2, \ldots, L$, only $\lfloor 2/3 j_l \rfloor$ of the beams survive. For ease of exposition, we will drop the $\lfloor \cdot \rfloor$ function since this drop will influence only a few constants in the analysis. Thus, after the end of the $L$ batches, there will be three beams, i.e., $(2/3)^L K = 3$. Therefore, we get

$$L = \left( \log_2 K/3 \right)/\left( \log_2 3/2 \right).$$  (16)

Hence, UB3 outputs the best beam index as $\hat{k}_{L+1}$ (i.e., $\hat{k}_\alpha$) after exploring for $\alpha n$ rounds.

Remark 3. UB3 and LSE differ notably in three aspects. Firstly, in LSE, the beams played in a batch are selected based on the golden ratio $\phi$. LSE discards about a $1/\phi$ fraction of the beams.
beams in each batch, while UB3 eliminates a fixed 1/3rd of the available beams in each batch. Note that 1/\phi > 1/3. Hence, as compared to LSE, UB3 eliminates beams less aggressively. Secondly, LSE requires a sequence of parameters (\epsilon, \delta_i) for every batch as input. Moreover, unlike LSE, UB3 operates as a parameter-free algorithm, which is highly desirable. LSE is applicable to problems where the mean rewards of the neighboring beams are separated at least by an amount \Delta_L, i.e., \Delta > \Delta_L (\{17, Assum 3.4\}), and uses this information in deciding the beam plays in each batch (\{17, Prop. 5.4\}). UB3 is an improvement over LSE since it does not impose constraints on problem instances, such as requiring a minimum separation of at least \Delta_L, and works as long as \Delta > 0. Lastly, in each batch, LSE adds one new beam based on the golden ratio, while UB3 adds two beams by uniformly dividing the space. UB3 uniformly increases the number of samples by a constant factor (3/2) in each batch. In our case, carefully selecting the batch duration l is crucial to meet the budget constraint.

V. Bounds on the Error Probability

In the following, we find an upper bound on the error probability of UB3. Next, we find a lower bound on the error probability of not identifying the optimal beam for any fixed-budget pure exploration bandits with unimodal structure.

A. Upper Bound on the Error Probability of Algorithm UB3

**Theorem 1.** Let the output of UB3 be \( b^\ast_{L+1} \) after \( L+1 \) number of batches, where \( L \) is given by (16). Let \( \Delta = \min_{2 \leq i \leq K} |\mu_{b_{i-1}} - \mu_{b_{i-1}}| > 0 \) denote the minimum gap between the means of the RSS of any two neighboring beams. Then for any \( \alpha n > K \), the probability that the beam \( b^\ast_{L+1} \) provided by UB3 is not the best beam is bounded as

\[
Pr(b^\ast_{L+1} \neq b_{k^\ast}) \leq 2 \left( \frac{\log_2 K / 3}{\log_2 3/2} - 1 \right) \exp \left( -\frac{\alpha n}{72} \Delta^2 \right) + 4 \exp \left( -\frac{\alpha n}{162} \Delta^2 \right).
\]

(17)

**Proof.** The proof is given in Appendix A. □

Note that, as \( K > 1, \exp \left( -\frac{\alpha n}{72} \Delta^2 \right) < \exp \left( -\frac{\alpha n}{72} \Delta^2 \right). \) As a result, the first term is dominant in the upper bound in (17), and the error probability is thus of order \( O \left( \log_2 K \exp \left( -\alpha n \Delta^2 \right) \right) \). For unstructured bandits, the error probability of Sequential Halving algorithm is upper bounded as \( O \left( \log_2 K \exp \left( -\frac{\alpha n \Delta^2}{K \log_2(K)} \right) \right) \), which matches with the lower bound up to a multiplicative factor of \( \log_2(K) \) [47]. Note that the exponent term in the error bound of Sequential Halving has \( K \log_2(K) \) factor in the denominator, which does not appear in the exponent of the error bound of UB3 for unimodal bandits. Therefore, the error probability for unimodal bandits is smaller than that for non-unimodal bandits, as expected, and our analysis captures this gain by shaving off the factor \( K \log_2(K) \) in the error bound exponent.

Note that for the case \( \Delta = 1/K \), the exponent term in the error probability bound of UB3 is dependent on \( K \). However, this is a special case of the suboptimal gap. In general, the gap between the beams is considered for any arbitrary mean values, independent of the number of beams. Hence, we believe that the correct interpretation is that our result will hold for any arbitrary mean values, and the exponent term of the error bound is independent of the number of beams, in general.

One can show that the reduction of factor \( K \log_2(K) \) in the error exponent is the best one can achieve for the unimodal bandits. We next consider the lower bound for fixed-budget pure exploration with the unimodal structure, confirming that the error exponent should be independent of the number of beams for any optimal algorithm for \( \alpha n > K \).

B. Lower Bound for Pure Exploration Unimodal Bandit

A lower bound on the error probability for the fixed budget MAB without assuming any structure is established in [47]. We adapt the proof to include the unimodal structure to derive a lower bound on the error probability. Note that for the lower bound, we have considered \( K \) arms, i.e., \( K \) beams. The proof uses bandit instances derived using "flipping construction" [2]. Below, we have given an overview of our constructions.

Let \( \nu := \{\nu_k\}_{1 \leq k \leq K} \) be a unimodal bandit instance such that \( \nu_k := Ber(p_k) \), where \( p_k \in [1/4, 1/2] \) for all \( 1 \leq k \leq K \) and \( p_{k^\ast} = 1/2 \). We denote \( p := \{p_k\}_{1 \leq k \leq K} \) as the set of means of \( K \) arms. Let \( \nu' := \{\nu'_k\}_{1 \leq k \leq K} \) be another bandit instance where \( \nu'_k := Ber(p'_k) \) and \( p'_k = 1 - p_k \) for all \( 1 \leq k \leq K \). Using \( \nu \) and \( \nu' \) we construct two more bandit instances, \( \nu_{k^\ast - 1} \) and \( \nu_{k^\ast + 1} \), with means \( p_{k^\ast - 1} \) and \( p_{k^\ast + 1} \), respectively, as follows

\[
p_{k^\ast - 1} = p_i \quad \forall i \neq k^\ast - 1 \quad \text{and} \quad p_{k^\ast - 1} = p'_i \quad \forall i = k^\ast - 1
\]

\[
p_{k^\ast + 1} = p_i \quad \forall i \neq k^\ast + 1 \quad \text{and} \quad p_{k^\ast + 1} = p'_i \quad \forall i = k^\ast + 1
\]

It is easy to note that both bandit instances \( \nu_{k^\ast - 1} \) and \( \nu_{k^\ast + 1} \) are unimodal with the optimal arms being \( k^\ast - 1 \) and \( k^\ast + 1 \), respectively. For simplicity, we denote the bandit instance \( \nu' \) as \( i \). Furthermore, we define \( d_k := p_{k^\ast} - p_k = 1/2 - p_k \), for any \( 1 \leq k \leq K \). Set \( \Delta_k = d_k + d_{k^\ast}, \) if \( k \neq i \) and \( \Delta_k = d_{k^\ast}, \) for any \( i \in \{k^\ast - 1, k^\ast + 1\} \) and any \( k \in \{1, \ldots, K\} \). Note that \( \{\Delta_k\}_{k} \) denotes the beam gaps of the bandit problem \( i \).

We now define the following quantities for bandit instances \( i \) for any \( i \in \{k^\ast - 1, k^\ast + 1\} \) as

\[
\bar{H}(i) := \sum_{k \in \{i-1, i+1\}} \frac{1}{(\Delta_k)^2} \quad \text{and} \quad \bar{H} := \sum_{i \in \{k^\ast - 1, k^\ast + 1\}} \frac{1}{d_i^2} \bar{H}(i).
\]

The following theorem gives the lower bound.

**Theorem 2.** For any bandit strategy that returns the beam \( \hat{k}_{\alpha n} \) after \( \alpha n \) budget, it holds that

\[
\max_{i \in \{k^\ast - 1, k^\ast + 1\}} P_i(\hat{k}_{\alpha n} \neq i) \geq \frac{1}{6} \exp \left( -\frac{\alpha n}{H(k^\ast)} \right) - 2\sqrt{\alpha n \log(6\alpha nK)}.
\]

(18)

2The authors in [41] have first introduced the concept of "flipping constructions". They constructed \( K \) Bernoulli bandit problems using "flipping constructions", where for the bandit problem \( i, \) any arm \( k, \) where \( k = 1, 2, \ldots, K, \) has the distribution \( \nu'_k, \) i.e., all the arms have the distribution \( \nu'_k \) except arm \( i, \) which has the distribution \( \nu'_k \). More specifically, in the bandit problem \( i, \) arm \( i \) has the best mean. The authors in [47] improved on the flipping construction of [41] by providing further information to the algorithm.
\[
\max_{i \in \{k^*-1, k^*+1\}} \left[ P_i(\hat{k}_\text{on} \neq i) \times \exp \left( \frac{60 \alpha_n}{h \tilde{H}(i)} + 2\sqrt{\alpha_n \log(6\alpha_n K)} \right) \right] \geq \frac{1}{6}. \tag{19}
\]

**Proof.** The proof of this theorem follows lines similar to [47, Thm. 2] after applying the change of measure rule to the restricted beams set. The proof is given in Appendix B. ■

**Corollary 1.** Assume that \( \alpha_n \geq \max_{i \in \{k^*-1, k^*+1\}} \left( \tilde{H}(k^*), \tilde{H}(i) \right)^2 \left( \frac{4 \log(6\alpha_n K)}{(60)^2} \right). \) For any bandit strategy that returns beam \( \hat{k}_\text{on} \) after \( \alpha_n \) budget, it holds that
\[
\max_{i \in \{k^*-1, k^*+1\}} \left[ P_i(\hat{k}_\text{on} \neq i) \times \exp \left( \frac{120 \alpha_n}{H(k^*)} \right) \right] \geq \frac{1}{6},
\]
and also
\[
\max_{i \in \{k^*-1, k^*+1\}} \left[ P_i(\hat{k}_\text{on} \neq i) \times \exp \left( \frac{120 \alpha_n}{H(i)} \right) \right] \geq \frac{1}{6}.
\]

**Proof.** The proof is given in Appendix C. ■

We can establish a lower bound using this corollary.

**Theorem 3.** For any unimodal bandit strategy that returns beam \( \hat{k}_\text{on} \) after \( \alpha_n \) budget, the following holds
\[
\max_{i \in \{k^*-1, k^*+1\}} \left[ P_i(\hat{k}_\text{on} \neq i) \times \exp \left( \frac{75 \alpha_n}{H(i)} \right) \right] \geq \frac{1}{6} \tag{20}
\]

**Proof.** The proof is given in Appendix D. ■

From the above theorem, following the same interpretation as in [47], we conclude that for at least one bandit problem \( \nu \) characterized by a complexity \( \tilde{H}(\nu) \), any bandit strategy will misidentify the best arm with probability of at least order \( O \left( \frac{1}{6} \exp \left( -75 \frac{\alpha_n}{H(i)} \right) \right) \). Note that the error exponent does not depend on \( K \), but only depends on the suboptimal gaps of the neighbours of the optimal arm. In contrast, for the unstructured case, the authors in [47] proved that the lower bound of the error probability of the fixed-budget for best arm identification is of order \( \exp \left( -\frac{\alpha_n \log K}{H(\nu)} \right) \). Note that for fixed \( \tilde{H} \), as we increase the number of beams \( K \), the error bound of the unstructured case will become worse, whereas for the unimodal case, the error bound will not be affected. Therefore, the unimodal structure improves the scaling of the error probability bound w.r.t. \( K \). We are motivated by the study of unimodal bandits in the cumulative regret setting, where the unimodal property helps improve the scaling with respect to \( K \). Using the approach of [47], we could establish similar properties for the unimodal bandits in the fixed budget setting, which has not yet been addressed in the existing literature. We note that the exponent in the error probability of \( UB3 \) differs from the optimal error exponent with respect to the complexity terms as \( \tilde{H}(\nu) \leq 2/\Delta^2 \) and hence is not optimal with respect to the specific problem instance parameters. Therefore, substantial room exists for refinement and enhancement of this lower bound, representing a promising avenue for future work.

**VI. NUMERICAL SIMULATIONS**

In this section, we corroborate our theoretical results using simulations. We first describe the simulation setup with the parameters used and present the results in the following subsections.

**A. Simulation Parameters**

We have set the carrier frequency \( f \) at 30 GHz, the signal bandwidth is 200 KHz, and the noise power spectrum density at the HMT is \(-174 \) dBm/Hz. The transmit power is set to \( P = 35 \) dBm. The number of antennas at the transmitter is set to \( M = 50 \) antennas. The number of beams \( K \) varies from 64 to 128. We have considered one LOS path and 2 NLOS paths. For the LOS path, we have the free space path loss(FSPL) model [48] as
\[
FSPL(\text{in dB}) = 10 \log_{10} \left( \frac{4\pi d^2}{\lambda} \right), \tag{21}
\]
where \( \xi, d, \) and \( \lambda \) represent the path loss exponent, transmission distance (in meters), and wavelength (in meters), respectively. For each NLOS path, we assume that the NLOS path loss suffers around 10 dB more than the LOS path loss [49]. The channel parameters for the simulations are given in Table I. Simulation results are averaged over 1000 channel realizations.

We compare the performance of \( UB3 \) with the following algorithms:

- **Sequential Halving (Seq. Halv.)** [50]: This algorithm is used for pure exploitation in non-unimodal bandits for fixed \( \alpha_n \) pilot symbols. The algorithm was proved to be optimal [47], and hence, a comparison would give the idea about how the additional information of unimodality would improve the performance.

- **Linear Search Elimination (LSE)** [17]: Although this algorithm was proposed for continuous beam unimodal bandit problems, we have considered the algorithm for fixed \( \alpha_n \) pilot symbols and for discrete beams. A comparison of \( UB3 \) with LSE is pertinent as it is a well-known algorithm for unimodal bandits.

- **Hierarchical Beam Alignment (HBA)** [12]: This algorithm was shown to have good performance for regret minimization when compared to existing algorithms, considering the prior knowledge of channel fluctuations. In [12], authors develop the HBA algorithm that outperforms other state-of-the-art algorithms such as UCB, UBA and HOO. As a result, we have considered HBA as our exploration and exploitation tradeoff-based benchmark. The algorithm parameters are kept at \( \rho_1 = 3, \gamma = 0.5 \). We have adapted HBA algorithm to a fixed-budget setting.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier frequency ( f )</td>
<td>30 GHz</td>
</tr>
<tr>
<td>Wavelength ( \lambda )</td>
<td>1 cm</td>
</tr>
<tr>
<td>( \sigma^2 ) (Noise power for 200 KHz)</td>
<td>(-121 ) dBm</td>
</tr>
<tr>
<td>Number of beams ( K )</td>
<td>{64, 128}</td>
</tr>
<tr>
<td>HBA parameters ( (\rho_1, \gamma) )</td>
<td>(3, 0.5)</td>
</tr>
<tr>
<td>Distance considered ( d )</td>
<td>{200, 400} m</td>
</tr>
<tr>
<td>Path loss exponent ( \xi )</td>
<td>1.74</td>
</tr>
<tr>
<td>Total number of transmitted symbols ( n )</td>
<td>3000</td>
</tr>
</tbody>
</table>

**TABLE I:** Parameters for simulation
HBA scales quadratically in $\alpha n$, whereas it is of $O(\alpha n)$ in UB3.

B. Comparison With Other Pure Exploration Algorithms

In Fig. 4, we have numerically evaluated the convergence of the upper bound of UB3 to its actual error probability for beam sizes of $K = \{64, 128\}$ and for distances of $d = 200$ and 400 m, given by (17). As the number of pilots increases, both the error probability and the upper bound of the error probability of UB3 converge to each other, and both converge to zero.

In Fig. 5, we have compared the error probability for UB3 algorithm with LSE and Seq. Halv. for beam sizes of $K = \{64, 128\}$ and for distances of $d = 200$ and 400 m. We see that LSE, which has an equal number of sampling for the beams in all phases, has the worst error probability performance. For a small number of beams, both Seq. Halv. and UB3 have comparable performance, but as the number of beams increases, UB3 has a smaller error probability compared to Seq. Halv.

UB3 can identify the best beam with a probability of more than 97% using 300 pilots in the channel estimation phase for 64 beams. In contrast, the other state-of-the-art algorithms identify the best beam with a probability of 85%. Note that for Seq. Halv. require at least 400 and 900 pilots for $K = 64$ and 128 beams, respectively, to complete their execution and hence their graph starts after that many pilot symbols $\alpha n$.

Note that even though the minimum budget requirement (as a function of $K$) for LSE is much smaller than both UB3 and Seq. Halv., the number of samples it runs for beams neighbouring to $k^*$ is much lesser, resulting in more error probability. Seq. Halv. needs at least $K \log_2(K)$ batches to complete one phase and has samples for all beams in every phase. Hence, it has fewer pilots remaining when the algorithm is executed in the neighbourhood of $k^*$ compared to UB3. Thus, the minimum duration requirement for UB3 as a function of $K$ is much

- **Hierarchical Optimal Sampling of Unimodal Bandits (HOSUB) [34]:** This algorithm exploits the benefits of hierarchical codebooks and the unimodality of RSS to achieve the best beam in the fixed $\alpha n$ pilot symbols.

- **Thompson Sampling for Unimodal Bandits (TS-UB) [38]:** This TS-based algorithm exploits the unimodal structure of the mean RSS and makes decisions based on the posterior distribution within the neighborhood of the beam with the highest empirical mean at each step instead of exploring the entire decision space. In [38], TS-UB outperforms other TS-based benchmarks, such as TS, UTS. Therefore, we have considered TS-UB as TS-based exploration and exploitation tradeoff-based benchmark.

We note that the computational complexity of HBA scales quadratically in $\alpha n$, whereas it is of $O(\alpha n)$ in UB3.

In Fig. 5, we have compared the error probability for UB3 algorithm with LSE and Seq. Halv. for beam sizes of $K = \{64, 128\}$ and for distances of $d = 200$ and 400 m. We see that LSE, which has an equal number of sampling for the beams in all phases, has the worst error probability performance. For a small number of beams, both Seq. Halv. and UB3 have comparable performance, but as the number of beams increases, UB3 has a smaller error probability compared to Seq. Halv. UB3 can identify the best beam with a probability of more than 97% using 300 pilots in the channel estimation phase for 64 beams. In contrast, the other state-of-the-art algorithms identify the best beam with a probability of 85%. Note that for Seq. Halv. require at least 400 and 900 pilots for $K = 64$ and 128 beams, respectively, to complete their execution and hence their graph starts after that many pilot symbols $\alpha n$.

Note that even though the minimum budget requirement (as a function of $K$) for LSE is much smaller than both UB3 and Seq. Halv., the number of samples it runs for beams neighbouring to $k^*$ is much lesser, resulting in more error probability. Seq. Halv. needs at least $K \log_2(K)$ batches to complete one phase and has samples for all beams in every phase. Hence, it has fewer pilots remaining when the algorithm is executed in the neighbourhood of $k^*$ compared to UB3. Thus, the minimum duration requirement for UB3 as a function of $K$ is much
We also observe that initially, as we increase the number of pilots, the benchmark algorithms get sufficient pilots to exploit the unimodal structure of the RSS of the beams. We developed an algorithm named UB3 that identified the best beam with high probability within a fixed time duration. We gave an upper bound on the error probability of UB3. We also provided the lower bound for fixed-budget pure exploration with the unimodal structure, confirming that the error exponent should be independent of the number of beams for any optimal algorithm. Simulations validated the efficiency of UB3 which can identify the best beam using a smaller number of explorations.

UB3 works well when the RSS of beams satisfies the unimodal property. However, when RSS has a multimodal structure, then UB3 can be adapted using backtracking ideas proposed in [51]. In backtracking, eliminated beams are revisited to check if it is done by mistake and thus will not be stuck in a sub-optimal set of beams. It is interesting to evaluate the UB3 algorithms with backtracking on multimodal function and establish its performance guarantees as a future work.

APPENDIX

In this section, we will provide proof of the main results.

A. Proof of Theorem 1

Proof. UB3 splits the \( \alpha n \) pilot symbols in \( L + 1 \) batches that satisfies (14), where \( L = \frac{\log_2 K}{3} \) and outputs the beam \( b_{k_{L+1}} \).

We will now upper bound the error probability as,

\[
\Pr(b_{k_{L+1}} \neq b_{k^*}) \leq \sum_{l=1}^{L+1} \Pr(b_{k^*} \text{ elim. in } l). \tag{22}
\]

The best beam is eliminated in batch \( l \) in the following cases:

1) \( b_{k^*} \in \{b_{kA}, \ldots, b_{kB} - 1\} \), and \( \hat{\mu}_{b_{LC}}^l \) or \( \hat{\mu}_{b_{LD}}^l \) is greater than both \( \hat{\mu}_{b_{kA}}^l \) and \( \hat{\mu}_{b_{kB}}^l \).

2) \( b_{k^*} \in \{b_{kC} + 1, \ldots, b_{kD}\} \), and \( \hat{\mu}_{b_{kA}}^l \) or \( \hat{\mu}_{b_{kB}}^l \) is greater than both \( \hat{\mu}_{b_{LC}}^l \) and \( \hat{\mu}_{b_{LD}}^l \).

The two cases are illustrated in Fig. 7. From Remark 2, \( b_{k^*} \) will not get eliminated if \( b_{k^*} \in \{b_{kB}, \ldots, b_{kC}\} \), so this case is not favourable here. Further note that Case 1 and Case 2 are symmetrical. Hence we can consider that \( b_{k^*} \) will always fall in either one of the cases. Without loss of generality, we consider Case 1.

\[
\Pr(b_{k^*} \text{ elim. in } l) \\
\leq \Pr(\hat{\mu}_{b_{kC}}^l > \hat{\mu}_{b_{kA}}^l \text{ and } \hat{\mu}_{b_{kB}}^l b_{k^*} \in \{b_{kA}, \ldots, b_{kB} - 1\}) \\
+ \Pr(\hat{\mu}_{b_{kD}}^l > \hat{\mu}_{b_{kA}}^l \text{ and } \hat{\mu}_{b_{kB}}^l b_{k^*} \in \{b_{kA}, \ldots, b_{kB} - 1\})
\]
Case 2

Fig. 7: Different cases of elimination in any batch. $\Delta_{B,C}$

$1. \text{Using } \Delta = \min_{j} \{ \mu_{b_{k,A}} - \mu_{b_{k,B}} \} | b_{k,*} \in \{ b_{k,A}, \ldots, b_{k,B} - 1 \} \}$, 

where the last inequality is due to the fact that, for Case 1, $\mu_{b_{k,C}} \geq \mu_{b_{k,D}}$ by unimodality. Now for Case 1, $\mu_{b_{k,B}}$ is always greater than $\mu_{b_{k,C}}$, but $\mu_{b_{k,A}}$ may not be greater than $\mu_{b_{k,C}}$. Then, we can further upper bound (23) as

$$\Pr(b_{k,*} \text{ elim. in } l) \leq 2 \Pr(\hat{\mu}_{b_{k,C}} > \hat{\mu}_{b_{k,B}} | b_{k,*} \in \{ b_{k,A}, \ldots, b_{k,B} - 1 \}).$$

We will now apply Hoeffding’s inequality in (24) as stated in the following lemma.

**Lemma 2** (Hoeffding’s Inequality for Subgaussian Random Variables [52]). If $X_1, \ldots, X_m$ are $m$ i.i.d samples drawn from $\beta$-Subgaussian then for any $i \in [m]$, then

$$P(X_i \geq \mu + \epsilon) \leq \exp \left( -\frac{\epsilon^2}{2\beta^2} \right)$$

and

$$P \left( \frac{1}{m} \sum_{i \in [m]} X_i \geq \mu + \epsilon \right) \leq \exp \left( -\frac{m\epsilon^2}{2\beta^2} \right).$$

Thereby, applying Lemma 2 in (24), we have

$$\Pr(\hat{\mu}_{b_{k,C}} > \hat{\mu}_{b_{k,B}}) \leq \exp \left\{ -\frac{N_l}{2} \left( \Delta_{B,C} \right)^2 \right\},$$

where $\Delta_{B,C} = |\hat{\mu}_{b_{k,B}} - \hat{\mu}_{b_{k,C}}|$ which is greater than 0 for Case 1. Using $\Delta$, which is defined as $\Delta = \min_{2 \leq i \leq K} |\mu_{b_{i}} - \mu_{b_{i-1}}|$, and the fact that there are at least $\frac{3}{4}$ beams between $b_{k,B}$ and $b_{k,C}$, for Case 1 we have, $\Delta_{B,C} \geq (j_l/3)\Delta$. Thus from (23) and (25) we have,

$$\Pr(b_{k,*} \text{ elim. in } l) \leq 2 \exp \left\{ -\frac{N_l}{72} \left( j_l \Delta \right)^2 \right\}. \quad (26)$$

Using $j_l = (\frac{3}{4})^l K$ in (26), we can find the error probability in batches 1 and 2, batch $L + 1$, and the rest of the batches separately. Using (13) and (16), we have

$$\Pr(b_{k,*} \text{ elim. in } 1\&2) \leq 2 \exp \left\{ -\frac{\alpha n K^2}{72} \Delta^2 \right\} + 2 \exp \left\{ -\frac{\alpha n K^2}{162} \Delta^2 \right\} \leq 4 \exp \left\{ -\frac{\alpha n K^2}{162} \Delta^2 \right\}.$$  

(27)

For batch $L + 1$, since the best beam is selected among three beams, so eliminating $b_{k,*}$ in batch $L + 1$ can happen if the empirical mean of the second best beam or the third best beam is greater than the empirical mean for $b_{k,*}$, where each beam is sampled $\alpha n/9$ times. Therefore, we have

$$\Pr(b_{k,*} \text{ elim. in batch } L + 1) \leq 2 \exp \left\{ -\frac{\alpha n}{18} \Delta^2 \right\}.$$ 

(28)

From (26), the error probability for the remaining batches is

$$\Pr(b_{k,*} \text{ elim. in batch } 3 \text{ to batch } L) \leq 2 \exp \left\{ -\frac{\alpha n}{48} \left( \frac{2}{3} \right)^L \Delta^2 \right\}.$$ 

(29)

By (22), (27), (28) and (29), we obtain the upper bound as given in (17).

**B. Proof of Theorem 2**

**Proof.** The proof of this theorem follows the lines similar to [47, Thm. 2] to include the unimodal structure to derive a lower bound by applying the change of measure rule to on the restricted set of arms after applying the change of measure on the restricted set of arms.

We denote $p := \{ p_k \}_{1 \leq k \leq K}$ as the set of means of $K$ arms. We will find the lower bound of the probability that the learner fails to recommend the optimal arm when presented with instance $p$, i.e., $P(k_{\text{ann}} \neq k^*)$. To this end, we first define a set of bandit instances as follows.

- **Step 1: The bandit problems that satisfy the unimodal structure**

  We have considered $K$ pairs of Bernoulli arms where $\nu := \{ \nu_k \}_{1 \leq k \leq K}$ be a unimodal bandit instance such that $\nu_k := \text{Ber}(p_k)$, where $p_k \in [1/4, 1/2]$ for all $1 \leq k \leq K$, $p_{k^* - 1} = 1/2$, and $p_1 < p_2 < \cdots < p_{k^*-1} < p_{k^*} > p_{k^*+1} > \cdots > p_K$. We consider $\nu' := \{ \nu'_k \}_{1 \leq k \leq K}$ be another bandit instance where $\nu'_k := \text{Ber}(p'_k)$ and $p'_k = 1 - p_k$ for all $1 \leq k \leq K$. We define the $K$ Bernoulli bandit problem using “flipping constructions” [47, Thm. 2] where for the bandit problem $\nu' := \nu'_1 \otimes \nu'_2 \otimes \cdots \otimes \nu'_K$ with means $\nu'$, arm $i$ is the optimal arm with distribution

  $$\nu'_i = \begin{cases} 
  \nu_k, & \text{if } i \neq k \\
  \nu'_k, & \text{if } i = k.
  \end{cases}$$

  Hence, the bandit problem $\nu'$ does not follow unimodal structure if $i \notin \{ k^*-1, k^*, k^*+1 \}$. Therefore, by flipping the distributions for all other arms will result in a non-unimodal bandit problems except for the bandit problems $\nu'$ where $i \in \{ k^*-1, k^*, k^*+1 \}$. For simplicity, we will denote bandit problem $\nu'$ as $i$. Note that the bandit problem $p_{k^*} = p$. We define $d_k := p_{k^*} - p_k = \frac{1}{2} - p_k$, for any $1 \leq k \leq K$. Set $\Delta_k = \Delta_k + d_k$, if $k \neq i$ and $\Delta_i = d_i$, for any $i \in \{ k^*-1, k^*, k^*+1 \}$ and any $k \in \{ 1, \ldots, K \}$. Note that $\{ \Delta_k \}$ denotes the arm gaps of the bandit problem $i$. Hence we will focus on the three bandit problems $p_i$ for...
We recall the change of measure identity, refer [19], which write for any $\alpha n K$, we use the notation $P_i(\cdot)$ and $E_i(\cdot)$ to denote the probability and expectation, respectively, with respect to $\nu$, the Kullback–Leibler (KL) divergence between distribution $\nu$ and $\nu'$, can be written as
\[
KL(\nu, \nu') = \int \log \left( \frac{d\nu(x)}{d\nu'(x)} \right) d\nu(x).
\]
For $k \in \{1, 2, \ldots, K\}$, the KL divergence between two Bernoulli distributions $\nu_k$ of parameters $p_k$ and $\nu'_k$ of parameters $1 - p_k$ is given by
\[
KL_k := KL(\nu_k, \nu'_k) = (1 - 2p_k) \log \left( \frac{1 - p_k}{p_k} \right).
\]
Since $p_k \in \left[ \frac{1}{4}, \frac{1}{2} \right]$, the following inequality holds:
\[
KL_k \leq 10d_k^2. \tag{30}
\]
Let us consider $1 \leq t \leq \alpha n$. We define the quantity as
\[
KL_{k,t} = \frac{1}{t} \sum_{s=1}^{t} \log \left( \frac{d\nu_s}{d\nu_k}(X_{k,s}) \right)
\]
\[
= \frac{1}{t} \sum_{s=1}^{t} \left\{ 1 \{X_{k,s} = 1\} \log \left( \frac{p_i}{1 - p_i} \right) + 1 \{X_{k,s} = 0\} \log \left( \frac{1 - p_i}{p_i} \right) \right\},
\]
where $X_{k,s}$ are i.i.d. samples from $\nu'_k$ for $s \leq t$ and bandit problem $i$.

Let us define an event as follows:
\[
\zeta = \left\{ \forall 1 \leq k \leq K, \forall 1 \leq t \leq \alpha n, \left| KL_{k,t} - KL_k \right| \leq 2 \sqrt{\frac{\log(nK)}{t}} \right\}. \tag{31}
\]
According to [47], Lemma 1], the concentration bound for $KL_{k,t}$ that holds for the bandit problem $i$ where $i \in \{k^* - 1, k^*, k^* + 1\}$ is given by,
\[
P_i(\zeta) \geq \frac{5}{6}, \quad \text{for } i \in \{k^* - 1, k^*, k^* + 1\}. \tag{32}
\]

Step 3: A change of measure Let Alg denote the active strategy of the learner that returns some arm $\hat{k}_\alpha$ after using $\alpha n$ pilot symbols. Let $\{T_k\}_{1 \leq k \leq K}$ denote the numbers of samples collected by Alg on each arm of the bandits and they are stochastic in nature. Note that according to the definition of the fixed budget setting we have $\sum_{1 \leq k \leq K} T_k = \alpha n$. Let us write for any $0 \leq k \leq K$,
\[
t_k = E_{k^*}[T_k] \quad \text{and} \quad \sum_{1 \leq k \leq K} t_k = \alpha n.
\]
We recall the change of measure identity, refer [19], which states that for any measurable event $\xi$ and for any $i \in \{k^* - 1, k^* + 1\}$, we have
\[
P_i(\xi) = E_{k^*}\left[ 1\{\xi\} \exp \left( -T_iKL_{i,T_i} \right) \right], \tag{33}
\]
as the product distributions $p^i$ and $p^{k^*}$ only differ in arm $i$ and as the active strategy only explored the samples $\{X_{k,s}\}_{1 \leq k \leq K, s \leq T_k}$.

We consider the event $\xi_i$ as the event where the algorithm outputs arm $k^*$ at the end of $\alpha n$ budget, where $\zeta$ holds, and where the number of times arm $i$ was pulled is smaller than $6t_i$, i.e., for $i \in \{k^* - 1, k^* + 1\}$ we define
\[
\xi_i = \left\{ \hat{k}_\alpha = k^* \right\} \cap \left\{ \zeta \right\} \cap \left\{ T_i \leq 6t_i \right\}. \tag{34}
\]
Applying the event $\xi_i$ as given by (34) in (33) we obtain,
\[
P_i(\xi_i) = E_{k^*}\left[ 1\{\xi_i\} \exp \left( -T_iKL_{i,T_i} \right) \right].
\]
Following the same lines of proof of [47][Step 2, Thm 2], for $i \in \{k^* - 1, k^* + 1\}$ we get
\[
P_i(\xi_i) \geq \exp \left( -6t_i KL_{i,T_i} - 2\sqrt{\alpha n \log(6\alpha n K)} \right) P_{k^*}(\xi_i) \tag{35}
\]

Step 4: Lower bound on $P_{k^*}(\xi_i)$ for any reasonable algorithm We assume that the probability that Alg makes a mistake on problem $k^*$ is less than $1/2$, i.e.,
\[
E_{k^*}[\hat{k}_\alpha \neq k^*] \leq 1/2. \tag{36}
\]
Note that if Alg does not satisfy that, it performs badly on bandit problem $k^*$ and its probability of success is not larger than $1/2$ uniformly on the three bandit problems we defined for $\{k^* - 1, k^*, k^* + 1\}$. For any $1 \leq k \leq K$, $k \neq k^*$ it holds by Markov’s inequality that
\[
P_{k^*}(T_k \geq 6k) \leq \frac{E_{k^*}[T_k]}{6k} = \frac{1}{6}, \tag{37}
\]
since $E_{k^*}[T_k] = t_k$ for Algorithm Alg.

By combining (36), (37) and (32), it holds by an union bound that for any $i \in \{k^* - 1, k^* + 1\}$
\[
P_{k^*}(\xi_i) \geq 1 - \left( \frac{1}{6} + \frac{1}{2} + \frac{1}{6} \right) = \frac{1}{6}. \tag{38}
\]
We will now combine (38) and the fact that $P_i(\hat{k}_\alpha \neq i) \geq P_i(\xi_i)$ for $i \in \{k^* - 1, k^* + 1\}$ and by applying in (33), we obtain
\[
P_i(\hat{k}_\alpha \neq i) \geq \frac{1}{6} \exp \left( -6t_i KL_{i,T_i} - 2\sqrt{\alpha n \log(6\alpha n K)} \right). \tag{39}
\]
By applying (30) in (39), we get,
\[
P_i(\hat{k}_\alpha \neq i) \geq \frac{1}{6} \exp \left( -60t_i d_i^2 - 2\sqrt{\alpha n \log(6\alpha n K)} \right). \tag{40}
\]

Step 5: Conclusions We defined $H(k^*) := \sum_{k \in \{k^* - 1, k^* + 1\}} \frac{1}{(1 + T_k)^2}$. We also know that $\sum_{1 \leq k \leq K} t_k = \alpha n$. By combining these two facts, we can say that there exists
Proof. of Corollary 1

Case 1: Let us consider
\[
\max_{i \in \{k^*-1, k^*+1\}} \left( \tilde{H}(k^*), \tilde{H}(i) \right) = \tilde{H}(i) h
\]  
Applying (44) in (43) we obtain,
\[
\frac{\alpha n}{\tilde{H}(i) h} \geq \frac{4 \log(6nK)}{(60)^2}
\]
\[
\frac{60\alpha n}{\tilde{H}(k^*)} \geq 2 \sqrt{N \log(6nK)}
\]
Applying (45) in (41) we get
\[
\max_{i \in \{k^*-1, k^*+1\}} P_i(\tilde{k}_{an} \neq i) \geq \frac{1}{6} \exp \left( -120 \frac{\alpha n}{\tilde{H}(i) h} \right).
\]
This concludes the proof of first part of the corollary.

Case 2: Let us consider for each \(i \in \{k^*-1, k^*+1\}\)
\[
\max_{i \in \{k^*-1, k^*+1\}} \left( \tilde{H}(k^*), \tilde{H}(i) \right) = \tilde{H}(k^*)
\]
Applying (47) in (43) we obtain,
\[
\frac{\alpha n}{\tilde{H}(k^*)} \geq \frac{4 \log(6nK)}{(60)^2}
\]
\[
\frac{60\alpha n}{\tilde{H}(k^*)} \geq 2 \sqrt{\alpha n \log(6nK)}
\]
Applying (48) in (42) we get
\[
\max_{i \in \{k^*-1, k^*+1\}} P_i(\tilde{k}_{an} \neq i) \exp \left( 120 \frac{\alpha n}{\tilde{H}(i) h} \right) \geq \frac{1}{6}.
\]
This concludes the proof of second part of corollary. ■

D. Proof of Theorem 3

Proof. We have, \(p_k = \frac{1}{2} - d_k\) such that \(p_k \in [1/4, 1/2]\) and follows unimodality and \(p_k^* = \frac{1}{2}^\ast\). Upper bounding \(\bar{h}\), we have
\[
\tilde{h} = \frac{1}{d_{k^*-1}^2} \tilde{H}(k^*-1) + \frac{1}{d_{k^*+1}^2 \tilde{H}(k^*+1)} = (I) + (II).
\]
We will upper bound (I) and (II).
\[
d_{k^*-1}^2 \tilde{H}(k^*-1) = d_{k^*-1}^2 \sum_{k \in \{k^*-2, k^*\}} (d_{k^*-1} + d_k)^2.
\]
\[
d_{k^*+1}^2 \tilde{H}(k^*+1) = d_{k^*+1}^2 \sum_{k \in \{k^*, k^*+2\}} (d_{k^*+1} + d_k)^2.
\]
Since \(d_{k^*} = 0\), \(d_{k^*-2} \geq d_{k^*-1}\), and \(d_{k^*+2} \geq d_{k^*+1}\), we get
\[
d_{k^*-1}^2 \tilde{H}(k^*-1) \leq 1 + \frac{1}{4} = \frac{5}{4}\]
\[
d_{k^*+1}^2 \tilde{H}(k^*+1) \leq 1 + \frac{1}{4} = \frac{5}{4}\]
By (51) and (52) we get
\[
\tilde{h} \geq \frac{4}{5} + \frac{5}{5} = \frac{8}{5}.
\]
Putting the value of \(\tilde{h}\) in Corollary we get
\[
\max_{i \in \{k^*-1, k^*+1\}} P_i(\tilde{k}_{an} \neq i) \exp \left( 75 \frac{\alpha n}{\tilde{H}(i) h} \right) \geq \frac{1}{6}.
\]

REFERENCES


