Fixed Budget Best Arm Identification in Unimodal Bandits

Anonymous authors
Paper under double-blind review

Abstract

We consider the best arm identification problem in a fixed budget stochastic multi-armed bandit setting, where the arm mean rewards exhibit a unimodal structure. We establish that the probability of misidentifying the optimal arm within a budget of T is lower bounded as $\mathcal{O}\left(\exp\left\{-T/\bar{H}\right\}\right)$, where $\bar{H}$ depends on the sub-optimality gaps of arms in the neighborhood of the optimal arm. In contrast to the lower bound for the unstructured case, the error exponent in this bound does not depend on the number of arms K and is smaller by a factor $K \log K$, which captures the gain achievable by exploiting the unimodal structure. We then develop an algorithm named *Fixed Budget Best Arm Unimodal Bandits (FB-BAUB)* that exploits unimodality to achieve the gain. Specifically, we show that the error probability of FB-BAUB is upper bounded as $\mathcal{O}\left(\log_2 K \exp\left\{-T\Delta^2\right\}\right)$, where Δ is the gap between the neighboring arms and $\bar{H} \leq 2\Delta^{-2}$. We demonstrate that FB-BAUB outperforms the state-of-the-art algorithms through extensive simulations. Moreover, FB-BAUB is parameter-free and simple to implement.

1 Introduction

Multi-armed bandit (MAB) is a popular setup to study decision-making under uncertainty. It has been applied in drug trials, recommendation systems, auctions, communication networks, and the list is growing, see Bouneffouf & Rish (2019). In MAB, a policy is any strategy that sequentially selects arms based on past observations. The performance of a policy is evaluated using criteria such as cumulative regret, simple regret, or best arm identification (BAI), depending on the application, refer to Lattimore & Szepesvári (2020) for details. While the policies optimizing cumulative regret balance exploration and exploitation optimally, the policies optimizing simple regret, or BAI, focus on optimal exploration. In BAI, the goal is to minimize the sample complexity of identifying the best arm within a given tolerance on the error probability (fixed confidence setting) or to minimize the error probability of not identifying the best arm within a given number of rounds (fixed budget setting), see (Karnin et al., 2013; Carpentier & Locatelli, 2016; Atsidakou et al., 2022). We focus on the fixed budget BAI setting.

The classical MAB setup considers the unstructured case, where reward distributions of the arms are independent with no relation among their means. However, many practical problems exhibit structural properties such as smoothness, linearity, unimodality, and convexity on the mean rewards, which can be exploited to improve the performance of the MAB algorithms (Magureanu et al., 2014; Combes et al., 2017; Yu & Mannor, 2011; Cheshire et al., 2021). Our objective in this paper is to develop algorithms that exploit the unimodal structure of the arm means to improve learning performance.

In unimodality, the arms' means are increasing and then decreasing in the arms' indices, exhibiting a change in the monotonicity pattern only at the globally optimal arm. Many real-life applications of unimodal structure arise in network throughput, see Hashemi et al. (2018), sequential pricing, and bidding in online sponsored search auctions, see Yu & Mannor (2011). Unimodality is exploited in Combes & Proutiere (2014); Saber et al. (2020a) to improve the cumulative regret bounds, where it is shown that the regret bound is asymptotically optimal and does not depend on the number of arms. Unimodlity structure is also exploited in Blinn et al. (2021), where the transmitter identifies the best beam in the fixed confidence setting, which is aligned with the receiver's beam in the wireless network. However, in wireless networks where devices (e.g., base stations and mobiles) need to operate synchronously, fixed budget BAI is better suited as the devices can know exactly when to switch from exploration to exploitation and hence require less information exchange for synchronization. This simplifies the protocol design. To the best of our knowledge, the fixed budget BAI with unimodal structure has not been studied in the literature. Also, as noted in Atsidakou et al. (2022), fixed budget BAI is more challenging than their fixed confidence counterparts, and our work fills this gap.

We develop an algorithm named *Fixed Budget Best Arm in Unimodal Bandits (FB-BAUB)* to address the BAI problem with the unimodal structure. The algorithm is motivated by the *Line Search Eliminations (LSE)* algorithm, introduced in Yu & Mannor (2011) and works in phases. In each phase of FB-BAUB, a portion of the arms is eliminated ($\sim 33\%$) based on empirical means of the arms, thereby reducing the search space for the optimal arm in the subsequent phases. FB-BAUB achieves an error probability of the order of $\mathcal{O}(\log K \exp(-T\Delta^2))$ within budget T , where K is the number of arms and Δ is the minimum gap between the means of any two neighboring arms. In the unstructured case, the best known achievable error probability is $\mathcal{O}\left(\log K \exp\left(-\frac{T\Delta^2}{K\log K}\right)\right)$, see Audibert et al. (2010). As a result of the unimodal structure, we reduced the error exponent by factor $K\log K$. We show that this reduction is the best possible by establishing a lower bound. Thus, we quantify the gain achieved by exploiting the unimodal structure. In establishing the lower bound, we generalize the "flipping constructions" of the bandit instance from Carpentier & Locatelli (2016) to include distribution with unbounded support and adapt it to unimodal bandits.

In summary, our contributions are as follows:

- We establish a lower bound for the unimodal bandits in a fixed budget BAI setting.
- We develop a parameter-free algorithm, named FB-BAUB, and derived an upper bound on its error probability. We show that the error exponent in the bound does not depend on K. The lower bound shows that the error exponent of FB-BAUB is of the right order.
- We empirically validate the superior performance of FB-BAUB compared to the other state-of-the-art algorithms in the literature that exploit the unimodal structure.

Detailed proofs of all the statements are given in Appendix A.

2 Related Work

The BAI problem is well-studied in the unstructured bandits in both fixed confidence and fixed budget settings, see Mannor & Tsitsiklis (2004); Even-Dar et al. (2006); Kalyanakrishnan et al. (2012); Karnin et al. (2013); Kaufmann et al. (2016); Jamieson & Nowak (2014); Garivier & Kaufmann (2016); Atsidakou et al. (2022); Wang et al. (2021). The authors in Gabillon et al. (2012) develop a unifying approach to analyse both settings leading to a meta-algorithm that can be applied to both settings. The lower bounds for these settings for unstructured bandits are developed by Chen & Li (2015); Carpentier & Locatelli (2016); Audibert et al. (2010).

Improving Cumulative Regret: Several works exploit the structural properties of the bandits to minimize cumulative regret. Abbasi-Yadkori et al. (2011); Chu et al. (2011); Dani et al. (2008) exploit *linearity* of observed rewards. The authors in Cesa-Bianchi & Lugosi (2012); Combes et al. (2015) studied *Combinatorial* bandits with bandit feedback exploit the *combinatorial* structure of the arms to learn the best subset of arms.

Unimodal structure is exploited in Yu & Mannor (2011); Combes & Proutiere (2014) to improve the regret performance. Combes & Proutiere (2014) provide a lower bound on any policy exploiting unimodal structure and develop an Optimal Sampling for Unimodal Bandits (OSUB) algorithm with a matching upper bound. Saber et al. (2020b) developed algorithms that do not require force explorations as used in the OSUB. The *smoothness* of the rewards expressed in terms of *Lipschitz* conditions are studied by Magureanu et al. (2014); Valko et al. (2014); Hanawal et al. (2015). A generic framework for analysing the cumulative regret of bandits exhibiting structural properties that include linearity, smoothness, and unimodality property is given in Combes et al. (2017). All of the aforementioned works are in the cumulative regret minimization setting.

Improving Best Arm Identification: In the best arm identification setting, the authors in Soare et al. (2014); Jedra & Proutiere (2020); Azizi et al. (2022) exploit the linearity structures, and the authors in Kocák

& Garivier (2020) exploit the smoothness structures to improve performance of BAI algorithms. The authors in Wang et al. (2021) studied fixed confidence BAI problem and developed Frank-Wolfe-based Sampling (FWS) whose sample complexity matches the lower bounds for a wide class of pure exploration problems. They applied FWS to other structural bandits such as threshold, linear, and Lipschitz, but they did not consider unimodal bandits. The authors in Garivier et al. (2017) consider threshold bandit problems (TBP), where the means of the arms are monotonically increasing, and the goal is to identify the set of arms with means above a given threshold. The authors characterise the sample complexity of the TBP in the fixed confidence setting. The TBP is extended to include other structural properties such as unimodality and concavity in Cheshire et al. (2020), and problem-independent bounds on the simple regret are derived. The paper Cheshire et al. (2021) extended the analysis of the TBP with monotonicity and concavity to establish problem-dependent bounds.

Lower Bounds: The lower bounds for the best arm identification are explored in Audibert et al. (2010); Garivier & Kaufmann (2016), and Carpentier & Locatelli (2016). In the fixed budget setting, Audibert et al. (2010) established a first lower bound, and it was improved in Garivier & Kaufmann (2016) using 'flipping constructions'. Notably, these approaches assume a parametric form (Bernoulli and Gaussian) for the underlying distributions and establish a minimax bound. Further, refining the flipping argument, Carpentier & Locatelli (2016) derived a lower bound that matches with the upper bound of the Successive Reject algorithm introduced in Audibert et al. (2010), thus establishing a tighter lower bound. The case of non-parametric bandits is studied in Barrier et al. (2023), which concentrates on developing instance-independent bounds for fixed budget scenarios. However, these bounds hold only asymptotically. The above work deals with the unstructured case, whereas we deal with the structured (unimodal) bandits.

Our work is closer to Cheshire et al. (2020), Yu & Mannor (2011), and Carpentier & Locatelli (2016). Cheshire et al. (2020) focuses on TBPs to exploit the unimodal property and provide a guarantee on the simple regret, which is different from our setting. The LSE algorithm introduced in Yu & Mannor (2011) provides the PAC guarantees in the fixed confidence setting for a continuous set of arms. The algorithm runs in phases, and the arms played in each phase are selected based on the golden ratio. Though our algorithm adapts the basic ideas of LSE, it differs in how arms are selected and eliminated. Also, it does not require any problem-dependent information. The details are given in Subsection 5.1.

3 Problem Setup

3.1 Notations

We consider the stochastic multi-armed bandit setting where the learner explores a finite set $\mathcal{A} = \{1, 2, \dots, K\}$ of arms over a fixed horizon T. The reward sequence for arm $k \in \mathcal{A}$ corresponds to independently and identically (i.i.d.) samples drawn from an unknown distribution $p_k(\mu_k)$ with mean μ_k, i.e., $\mu_k = \mathbb{E}_{X \sim p_k(\mu_k)}[X]$. We assume $p_k(\mu_k)$ is β sub-Gaussian, where $\beta > 0$, $\forall k \in \mathcal{A}$.

Let ϵ_β denote the set of all bandits instances that are β sub-Gaussian with distinct arm means. For any instance $\boldsymbol{p}(\boldsymbol{\mu}) := \{p_1(\mu_1), \dots, p_K(\mu_K)\} \in \epsilon_\beta$ with mean rewards of the arms as $\boldsymbol{\mu} := \{\mu_1, \dots, \mu_K\}$, let $k^* := k^*\big(\boldsymbol{p}(\boldsymbol{\mu})\big) = \arg\max_{k \in \mathcal{A}} \mu_k$ denote the optimal arm, i.e., the arm with the highest mean. We denote ϵ_U as the set of bandits in ϵ_β satisfying unimodality, defined below.

Definition 1. *(Unimodality): A bandit instance $\boldsymbol{p}(\boldsymbol{\mu}) \in \epsilon_\beta$ is said to be unimodal iff $\mu_1 < \mu_2 < \cdots < \mu_{k^*}$ and $\mu_{k^*} > \mu_{k^*+1} > \cdots > \mu_K$.*

Let $\mathcal{S}$ be the set of the arm means satisfies the unimodal structure, i.e.,

$$\mathcal{S} = \big\{\boldsymbol{\mu} : \mu_1 < \mu_2 < \cdots < \mu_{k^*} > \mu_{k^*+1} > \cdots > \mu_K\big\}.$$

3.2 Learning setup

We consider the following interaction between the learner and the environment over fixed rounds $T > 0$. We refer to T as the fixed budget. For any round $1 \leq t \leq T$, the learner chooses an arm $k_t \in \mathcal{A}$ and observes a stochastic reward drawn from the distribution $p_{k_t}(\mu_{k_t})$. In each round t, the learner decides which arm to play based on the samples observed in the past. At the end of T, the learner returns an arm $k_T \in \mathcal{A}$. A policy π is any strategy that selects an arm in each round, given the past observations. For a given policy π, let k_T^π denote the arm output at the end of T budget. Let Π denote the set of all policies that output an arm within T budget on unimodal bandits instances.

Objective: The goal is to find a policy in Π that exploits the unimodal structure of the mean rewards and minimizes the probability that the arm output at the end of T budget is not the optimal arm. Specifically, our objective is given as follows:

$$\inf_{\pi \in \Pi} \sup_{\boldsymbol{\mu} \in \mathcal{S}} P_{\boldsymbol{p}(\boldsymbol{\mu})}\left(k_T^\pi \neq k^*\right), \tag{1}$$

where $\Pr(\cdot)$ is over the randomness of the reward and the policy. We refer to the BAI setup given in (1) as the *fixed budget BAI for unimodal bandits.*

3.3 Problem Dependent Complexity

For a given instance $\boldsymbol{p}(\boldsymbol{\mu}) \in \epsilon_U$ with arm means $\boldsymbol{\mu}$, we express the complexity as $\bar{H} := \bar{H}\big(\boldsymbol{p}(\boldsymbol{\mu})\big)$, and is given by

$$\bar{H} := \sum_{k \in \{k^*-1, k^*+1\}} \frac{1}{(\mu_{k^*} - \mu_k)^2}. \tag{2}$$

We call $\bar{H}$ complexity as the characterisation of the hardness of understanding the problem, as we will see later. Similar problem-dependent quantities are consider in Audibert et al. (2010); Carpentier & Locatelli (2016); Jamieson & Nowak (2014), that characterize the complexity of bandit problems, e.g.,

$$H_1 = \sum_{k \neq k^*} \frac{1}{(\mu_{k^*} - \mu_k)^2} \text{ and } H_2 = \sup_{k \neq k^*} \frac{k}{(\mu_{k^*} - \mu_{(k)})^2},$$

where $\mu_{(k)}$ denotes the k^{th} largest mean of the arms. Following Audibert et al. (2010), it is easy to show the following inequalities hold (see Appendix A.1 for proof)

$$\min(H_2, \bar{H}) \leq H_1 \leq 2\log(K) H_2. \tag{3}$$

We next consider the lower bound for fixed budget BAI with the unimodal structure.

4 Lower Bound for Fixed Budget BAI of Unimodal Bandits

A lower bound on the error probability for BAI in the fixed budget setting without assuming any structure is established in Audibert et al. (2010); Garivier & Kaufmann (2016) and Carpentier & Locatelli (2016) using different techniques for constructing bandit problems, and all of them have provided minimax lower bound for the unstructured bandits. More specifically, Audibert et al. (2010) constructs $K!$ bandit problems by permutation of the arms and proposes that for any bandit problem, there exists a permutation such that any algorithm will make an error with probability at least $\exp\left(-\frac{T}{H_2}\right)$. Garivier & Kaufmann (2016) proposed the 'flipping constructions' and showed that there exists a bandit problem such that any algorithm will make an error with probability at least $\exp\left(-\frac{T}{H_1}\right)$, where $H_2 < H_1$. Carpentier & Locatelli (2016) improved the flipping construction of Garivier & Kaufmann (2016) by providing further information to the algorithm. They proposed that there exists a bandit problem such that any algorithm will make an error with a probability of at least $\exp\left(-\frac{T}{\log_2(K) H_1}\right)$. The authors argue that in the fixed budget setting, unlike in the fixed confidence setting, there is an additional $\log(K)$ price to pay for adaptation to H_1 in the absence of knowledge over this quantity.

We adopt the lower bound proof of Carpentier & Locatelli (2016) for fixed budget BAI problems on unimodal instances. Carpentier & Locatelli (2016) provided a lower bound using a particular choice of Bernoulli rewards, whereas we do not assume any particular choice of bandit instances. Our only assumption is that an unimodal structure exists over the mean rewards, and we consider Gaussian distributions for our case. Below, we have given an overview of our construction.

Fix $\beta = 1$. Let $\boldsymbol{p}(\boldsymbol{\mu}) := \{p_k(\mu_k)\}_{k\in\mathcal{A}} \in \epsilon_U$ be a unimodal bandit instance such that $p_k(\mu_k) := N(\mu_k, 1)$, where $\mu_k \in [1/4, 1/2]$ for all $k \in \mathcal{A}$ and $\mu_{k^*} = 1/2$. Let $\boldsymbol{p}'(\boldsymbol{\mu}') := \{p'_k(\mu'_k)\}_{k\in\mathcal{A}}$ be another bandit instance where $p'_k(\mu'_k) := N(\mu'_k, 1)$ and $\mu'_k = 2\mu_{k^*} - \mu_k$ for all $k \in \mathcal{A}$. Using $\boldsymbol{p}(\boldsymbol{\mu})$ and $\boldsymbol{p}'(\boldsymbol{\mu}')$ we construct two more bandit instance $\boldsymbol{p}^{k^*-1}(\boldsymbol{\mu}^{k^*-1})$ and $\mathbf{p}^{k^*+1}(\boldsymbol{\mu}^{k^*+1})$ with means $\boldsymbol{\mu}^{k^*-1}$ and $\boldsymbol{\mu}^{k^*+1}$ as follows.

$$
\begin{aligned}
\mu_i^{k^*-1} &= \mu_i \ \ \forall i \neq k^* - 1 \text{ and } \mu_i^{k^*-1} = \mu'_i \text{ for } i = k^* - 1 \\
\mu_i^{k^*+1} &= \mu_i \ \ \forall i \neq k^* + 1 \text{ and } \mu_i^{k^*-1} = \mu'_i \text{ for } i = k^* + 1
\end{aligned}
$$

It is easy to note that both bandit instances $\mathbf{p}^{k^*-1}(\boldsymbol{\mu}^{k^*-1})$ and $\mathbf{p}^{k^*+1}(\boldsymbol{\mu}^{k^*+1})$ are unimodal with the optimal arms being $k^* - 1$ and $k^* + 1$, respectively. Recall the definition of $\bar{H}(\boldsymbol{p}(\boldsymbol{\mu}))$ given in (2). We define $\bar{h}$ as

$$
\bar{h} := \sum_{i\in\{k^*-1, k^*+1\}} \frac{1}{(\mu_{k^*} - \mu_i)^2 \bar{H}\Big(\mathbf{p}^i(\boldsymbol{\mu}^i)\Big)},
$$

where $\bar{H}\Big(\mathbf{p}^i(\boldsymbol{\mu}^i)\Big)$ for bandit instance $\mathbf{p}^i(\boldsymbol{\mu}^i)$ is defined as

$$
\bar{H}\Big(\mathbf{p}^i(\boldsymbol{\mu}^i)\Big) = \sum_{k\in\{i-1, i+1\}} \frac{1}{(\Delta_k^i)^2}, \text{ where } \Delta_k^i = \begin{cases} 2\mu_{k^*} - \mu_i - \mu_k, & \text{if } k \neq i, \\ \mu_{k^*} - \mu_i, & \text{if } k = i. \end{cases}
$$

Note that the authors in Carpentier & Locatelli (2016) have defined the quantity $h^* = \sum_{i\in\mathcal{A}, i\neq k^*} \frac{1}{(\mu_{k^*} - \mu_i)^2 \bar{H}(\mathbf{p}^i(\boldsymbol{\mu}^i))}$. However, for the unimodality structure of the mean rewards, we can define the quantity $\bar{h}$ on the neighbourhood of the optimal arm k^*. We now provide the lower bound of the fixed budget BAI problem for unimodal bandits, as stated in the following theorem.

Theorem 1. *For any unimodal bandit strategy that returns arm k_T after T budget, where $T \geq \max_{i\in\{k^*-1,k^*+1\}} \Big(\bar{H}\big(\mathbf{p}(\boldsymbol{\mu})\big), \bar{H}\big(\mathbf{p}^i(\boldsymbol{\mu}^i)\big)\bar{h}\Big) \frac{4\log(6TK)}{12}$, it holds that*

$$
\max_{i\in\{k^*-1,k^*+1\}} \left[P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(k_T \neq i) \exp\left(15 \frac{T}{\bar{H}\big(\mathbf{p}^i(\boldsymbol{\mu}^i)\big)} \right) \right] \geq \frac{1}{6}. \tag{4}
$$

Proof. The proof is given in Appendix A.2. □

From the above theorem, we conclude that the lower bound for the fixed budget unimodal bandit is $O\left(\frac{1}{6}\exp\left\{-\frac{15T}{\bar{H}(\boldsymbol{p}(\boldsymbol{\mu}))}\right\}\right)$. Note that the exponent does not depend on K, but only on the sub-optimal gaps of the neighbours of the optimal arm. Our focus on improving the scaling in K is motivated by the study of unimodal bandits in the regret setting, where the unimodal property helps improve the scaling with respect to K. Specifically, a similar observation is also made for unimodal bandits in the cumulative regret setting, see Combes & Proutiere (2014).

5 FB-BAUB Algorithm

We propose an algorithm for unimodal bandit instances in the fixed budget BAI setting. The algorithm is based on the *Line Search Elimination (LSE)* method developed in Yu & Mannor (2011)and we refer to it as

Fixed Budget Best Arm in Unimodal Bandits (FB-BAUB). It is a parameter-free algorithm that only needs to know K and T, and the arms are eliminated based on their empirical means.

FB-BAUB splits the total budget T into $L+1$ phases. Let N_l for phase $l = 1, 2, \dots, L+1$, denotes the number of samples in phase l, where $\sum_{l=1}^{L+1} N_l = T$. We have chosen N_l such that after the first two phases, the number of samples increases by a factor of 3/2 in each subsequent phase, which helps to distinguish between the empirical means of the remaining arms, which are likely to be closer. Thus, N_l is given by

$$N_l = \begin{cases} \frac{2^{L-2}}{3^{L-1}} T & \text{for } l = 1, 2 \\ \frac{2^{L-(l-1)}}{3^{L-(l-2)}} T & \text{for } l = 3, 4, \dots, L+1 \end{cases} \tag{5}$$

and satisfy the budget constraints, i.e,

$$2 \times \frac{2^{L-2}T}{3^{L-1}} + \sum_{l=3}^{L+1} \frac{2^{L-(l-1)}T}{3^{L-(l-2)}} = T. \tag{6}$$

The arms are sampled and eliminated in each phase, so only one arm survives after the $L+1$ phase.

The pseudo-code of FB-BAUB is given in ALGO 1. It works as follows: Let $\mathcal{B}_l$ denote the set of arms available in phase l and $j_l := |\mathcal{B}_l|$ is the number of arms in the set $\mathcal{B}_l$. In each phase $l = 1, 2, \dots L$, the algorithm selects four arms $S_l = \{k^M, k^A, k^B, k^N\} \in \mathcal{B}_l$, which include the first, last and the two middle arms uniformly spaced from them (lines 4-7). Each of the arms is sampled for $N_l/4$ number of times (line 8). At the end of the phase, their empirical means, denoted as $\hat{\mu}_k$ (line 9), are obtained as follows:

$$\hat{\mu}_k^l = \frac{1}{N_l/4} \sum_{s=1}^{N_l/4} X_{k,s}^l, \quad \forall k \in S_l, \tag{7}$$

where $X_{k,s}^l$ denotes the sth sample from the k^{th} arm in phase l. Based on these empirical means, we eliminate at most $1/3^{rd}$ of the number of arms from the remaining set[1]. Specifically, if the arms k^M or k^A have the highest empirical means, we eliminate all the arms succeeding k^B in the set $\mathcal{B}_l$ (line 12). Similarly, if the arms k^B or k^N have the highest empirical means, we eliminate all the arms preceding k^A in the set $\mathcal{B}_l$ (line 14). Fig. 1 gives a pictorial representation of the elimination of arms in two possible cases. The remaining set of arms is then transferred to the next phase. In phase $L+1$, we are left with three arms. Each is sampled $N_{L+1}/3$ times, and the one with the highest empirical mean is output as the optimal arm (lines 18-22).

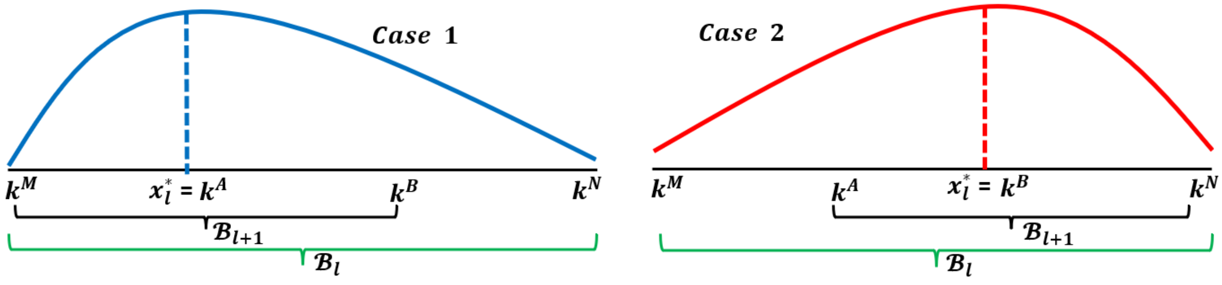

Figure 1: Different cases of elimination in phase l.

Remark 1. *Arms between k^M & k^A or k^B & k^N are eliminated in each phase, and the arms between k^A & k^B always survive.*

After phase $l = 1, 2, \dots, L$, $\lfloor \frac{2}{3} j_l \rfloor$ of the arms survive. For ease of exposition, we will drop the $\lfloor \rfloor$ function since this drop will influence only a few constants in the analysis. Thus, after the end of the L phases, there

---

[1] If the number of arms in a phase is not a multiple of 4, then less than $1/3^{rd}$ will be eliminated in that phase.

ALGO 1: Fixed Budget Best Arm in Unimodal Bandits (FB-BAUB)

```
Input: $T$ and $K$.
Initialise: $\mathcal{B}_1 = \mathcal{A}$, $j_1 \leftarrow |\mathcal{B}_1|$. Calculate $L$ from (8).
for $l = 1$ to $L$ do
    $k^M \leftarrow$ First arm of $\mathcal{B}_l$;
    $k^N \leftarrow$ Last arm of $\mathcal{B}_l$;
    $k^A \leftarrow \lceil j_l/3 \rceil^{th}$ arm of $\mathcal{B}_l$;
    $k^B \leftarrow \lfloor 2j_l/3 \rfloor^{th}$ arm of $\mathcal{B}_l$;
    Sample each arm in $S_l = \{k^M, k^A, k^B, k^N\}$ for $\frac{N_l}{4}$ number of times from (5)
    Obtain $\hat{\mu}^l_{k^M}$, $\hat{\mu}^l_{k^A}$, $\hat{\mu}^l_{k^B}$, $\hat{\mu}^l_{k^N}$ by (7).
    $x^*_l = \arg\max_{k \in S_l} \hat{\mu}_k$.
    if $x^*_l == \{k^M, k^A\}$ then
        $\mathcal{B}_{l+1} \leftarrow \{k \in \mathcal{B}_l : k^M \leq k \leq k^B\}$ Shrink to left
    else if $x^*_l == \{k^B, k^N\}$ then
        $\mathcal{B}_{l+1} \leftarrow \{k \in \mathcal{B}_l : k^A \leq k \leq k^N\}$ Shrink to right
    end
    $j_{l+1} \leftarrow |\mathcal{B}_{l+1}|$;
end for
for $l = L + 1$ do
    $\mathcal{B}_{L+1} = \{k^M, k^A, k^N\}$;
    Sample each arm in $\{k^M, k^A, k^N\}$ for $\frac{N_{L+1}}{3}$ no. of times and obtain $\hat{\mu}^l_{k^M}$, $\hat{\mu}^l_{k^A}$, $\hat{\mu}^l_{k^N}$.
    Obtain $\hat{k}_{L+1} = \arg\max_{k \in \mathcal{B}_{L+1}} \hat{\mu}_k$.
end for
Output: $k_T = \hat{k}_{L+1}$
```

will be three arms as

$$(2/3)^L K = 3 \implies L = \frac{\log_2 K/3}{\log_2 3/2}. \tag{8}$$

Therefore, FB-BAUB outputs the best arm as $\hat{k}_{L+1}$ (i.e., k_T) after exploring for T rounds.

5.1 Comparison between LSE and FB-BAUB

In Yu & Mannor (2011), LSE is developed for a continuous set of arms. However, they have also applied LSE for a finite set of arms in the fixed confidence setting. FB-BAUB and LSE differ primarily in three aspects.

1. *Elimination:* LSE eliminates about $1/\phi$ fraction of arms, where ϕ is the golden ratio. In contrast, FB-BAUB eliminates 1/3rd of the available arms in each phase.

2. *Input parameters:* LSE needs a sequence of parameters (ϵ_l, δ_l) for every phase as input, whereas no such input parameters are required in FB-BAUB. Hence, it is a parameter-free algorithm, which is highly desirable.

3. *Number of samples*: In the discrete case, LSE considers the instance where mean rewards of neighboring arms are separated at least by an amount D_L, i.e., $\Delta > D_L$ (Yu & Mannor (2011)[Assum 3.4]), and uses this information in deciding the arm plays in each phase (Combes & Proutiere (2014)[Prop. 5.4]). Whereas FB-BAUB does not require any problem-specific information and works as long means of the neighboring arms are separated, i.e., $\Delta > 0$.

4. *Arms selection policy*: LSE adds one new arm based on the golden ratio in each phase. Whereas FB-BAUB adds two new arms in each phase by uniformly dividing the space.

6 Performance Guarantee of FB-BAUB

This section provides the following theorem that gives the upper bound for the error probability of FB-BAUB.

Theorem 2. *Let $\mathbf{p}(\boldsymbol{\mu}) \in \epsilon_U$ follow β sub-Gaussian with arm means $\boldsymbol{\mu}$ and $\Delta = \min_{2 \le i \le K} |\mu_i - \mu_{i-1}| > 0$ denote the minimum gap between the means of any two neighboring arms. For any $T > K$, the error probability of FB-BAUB is bounded as*

$$P_{\boldsymbol{p}(\boldsymbol{\mu})}(\hat{k}_{L+1} \neq k^*) \leq 2\exp\left\{-\frac{TK}{32}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2\exp\left\{-\frac{TK}{72}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2\exp\left\{-\frac{T}{24}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2(L-2)\exp\left\{-\frac{T}{8}\left(\frac{\Delta}{\beta}\right)^2\right\}. \tag{9}$$

Proof. The proof is given in Appendix A.3. □

The first two terms in the upper bound correspond to the probability of eliminating the optimal arm k^* in phases 1 & 2. The 3rd term bounds the probability of eliminating the optimal arm in phase $L+1$, and the 4th term corresponds to the sum of the probabilities of eliminating the optimal arm in phases $l = 3, \ldots, L$. Note that, as $K > 1$, $\exp\left\{-TK\left(\frac{\Delta}{\beta}\right)^2\right\} < \exp\left\{-T\left(\frac{\Delta}{\beta}\right)^2\right\}$. As a result, the 3rd and 4th terms dominate the upper bound (9). Therefore, the error probability is of order $\mathcal{O}\left(\log_2 K \exp\left\{-T\Delta^2\right\}\right)$, where the error exponent term $\exp\left\{-T\Delta^2\right\}$ does not depend on K.

Comparison with Sequential Halving: For unstructured bandits, the error probability of *Sequential Halving* is upper bounded as $O\left(\log_2 K \exp\left\{-\frac{T\Delta^2}{K\log K}\right\}\right)$ and is optimal as it matches with the lower bound derived in Carpentier & Locatelli (2016) up to a multiplicative factor of $\log_2 K$. Note that the exponent term in the error bound of Sequential Halving has a $K\log_2 K$ factor. For unimodal bandits, the exponent term in the error bound of FB-BAUB does not depend on K is smaller by a factor of $K\log K$. As expected, the error probability for unimodal bandits should be smaller, and our analysis captures this gain by shaving off the factor $K\log_2 K$ in the error bound.

Comparision with LSE: For a finite set of arms, LSE assumes that the gap of the mean rewards of the neighboring arms is separated by at least $D_L > 0$ (Assumption 3.2 in Yu & Mannor (2011)). More specifically, LSE requires knowledge of D_L, and its sample complexity is expressed in terms of ϵ_l and δ_l for each phase l. However, FB-BAUB only considers that $\Delta > 0$, i.e., the arm means to be distinct. We do not need any assumptions on the minimum separation of the mean rewards of the neighboring arms, i.e., FB-BAUB need not know D_L. Therefore, FB-BAUB works when the arm means are distinct but arbitrarily close to each other, whereas LSE analysis requires the assumption that this separation is at least D_L. Moreover, we have adapted LSE to the fixed budget setting, where we have set N_l such that it satisfies the budget constraint. Furthermore, LSE gives a PAC bound in the fixed confidence setting, whereas we upper bound the error probability in the fixed budget setting.

We note that the error exponent of FB-BAUB differs from the optimal error exponent with respect to the complexity terms as $\bar{H}\big(\mathbf{p}(\boldsymbol{\mu})\big) \leq 2/\Delta^2$ and hence is not optimal with respect to the specific problem instance. It is an interesting open problem to develop an optimal algorithm in the fixed budget BAI setting with unimodality.

7 Simulation Results

In this section, we corroborate the theoretical guarantee of FB-BAUB by applying it to problem instances of varying difficulty. As no algorithm exists for the BAI in unimodal bandits in the fixed budget setting, we consider the following benchmark algorithms. The anonymized source code is available at https://anonymous.4open.science/r/FBBAUB-1BBD/README.md

Sequential Halving (Seq. Halv.) Karnin et al. (2013): This fixed budget BAI algorithm is for unstructured bandits but is shown to be optimal by Carpentier & Locatelli (2016). A comparison with this algorithm gives gains achieved by exploiting structure.

Successive Rejects (SR) Audibert et al. (2010): This fixed budget BAI algorithm for unstructured bandits is parameter-free and optimal up to a logarithmic term. SR outperforms UCB-E in Audibert et al. (2010); Shahrampour et al. (2017). Hence, we do not consider UCB-E for comparison. A comparison with this algorithm gives gains achieved by exploiting structure.

Linear Search Elimination (LSE) Yu & Mannor (2011): We consider the discrete variant of LSE proposed for a finite set of arms in the fixed confidence case and adopt it to the fixed budget setting. A comparison of FB-BAUB with LSE is pertinent as it is a well-known algorithm for unimodal bandits.

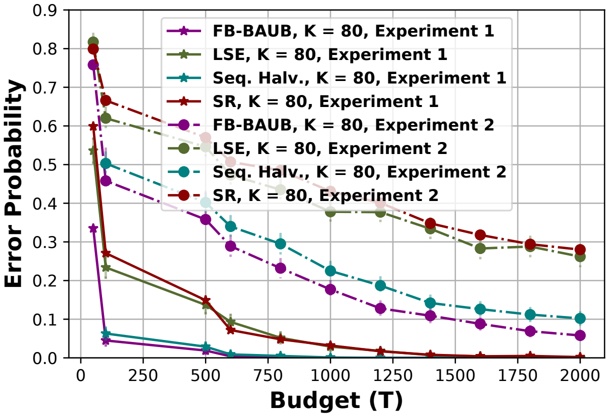

Figure 2: Error Probability vs T, $K = 80$.

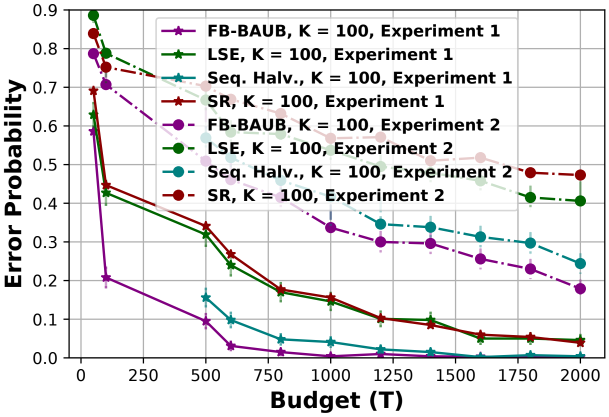

Figure 3: Error Probability vs T, $K = 100$.

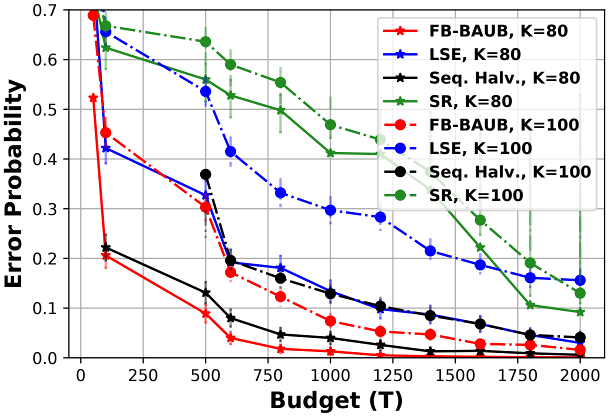

Figure 4: Exp. 3: Error Prob. vs T.

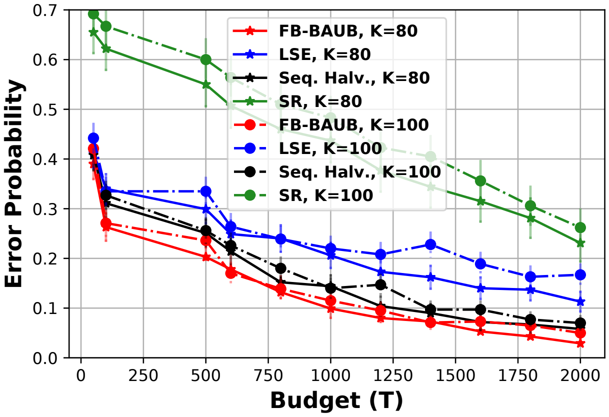

Figure 5: Exp. 4: Error Prob. vs T.

7.1 Comparison with pure exploration algorithms for unimodal structure

We consider two experimental setups to compare the performance of FB-BAUB with other benchmark algorithms. We consider K Gaussian arms with known variances $\sigma_i^2 = \sigma^2 = 5$, assuming that the mean of the best arm is $\mu_{k^*} = -252$. Our simulations are averaged over 1000 runs and are shown with confidence intervals.

Experiment 1: $\mu_1 = -312$, $\mu_{k^*} = -252$, $\mu_K = -311$, $\mu_{2:k^*-1} = \mu_1 + \frac{2(k-1)(\mu_{k^*}-\mu_1)}{k^*-1}$, and $\mu_{k^*+1:K} = \mu_{k^*} - \frac{2(k-k^*)(\mu_{k^*}-\mu_K)}{K-k^*-1}$.

Experiment 2: $\mu_1 = -312$, $\mu_{k^*} = -252$, $\mu_K = -311$, $\mu_{2:k^*-1} = \mu_1 + \frac{(k-1)(\mu_{k^*}-\mu_1)}{(k^*-1)}$, and $\mu_{k^*+1:K} = \mu_K - \frac{(k-k^*)(\mu_{k^*}-\mu_K)}{(K-k^*-1)}$.

Fig. 2 and Fig. 3 illustrate the performance of each algorithm for Exp. 1 and Exp. 2. We examine each setup for two values of $K = \{80, 100\}$. Exp. 1 and 2 are strictly unimodal in the sense that the optimal arm is lying in the interior. In Exp. 1, the means of the successive arms are well-separated compared to that in Exp. 2, i.e., Δ is lower for Exp. 2. As we decrease the gap between the means of the neighboring arms, the means of the neighboring arms of the optimal arm come close to each other. Hence, identifying the best arm becomes complicated, which increases the error probability. This makes Exp. 2 more challenging to identify the optimal arm compared to Exp. 1, and thereby the error probability is higher for Exp. 2 than Exp. 1. Moreover, the error probability increases as we increase the arm size.

SR has the worst error performance for each setup. For a small number of arms, both Seq. Halv. and FB-BAUB have comparable performance, but as the number of arms increases, FB-BAUB has a lesser error probability compared to Seq. Halv. as evident from the case of $K = 100$. Moreover, FB-BAUB can identify the best arm with a probability of more than 95% with a lesser budget than other state-of-the-art algorithms. More specifically, in Exp. 1 *FB-BAUB* can identify the best arm with a probability of more than 95% within 100 budget for 80 arms, while the other state-of-the-art algorithms need at least a 500 budget for executions. Hence, FB-BAUB outperforms the state-of-the-art algorithms.

We note that the minimum budget requirement (as a function of K) for LSE is much smaller than both FB-BAUB and Seq. Halv. for its feasible execution. However, as LSE samples for a fixed number of times for each of the arms in every phase, the number of samples it runs for arms neighboring k^* is much lesser, resulting in a higher error probability. Seq. Halv. needs at least $K \log_2(K)$ number of horizons to complete one phase and has samples for all arms in every phase. Note that for each setup, Seq. Halv. requires at least 100 and 300 rounds for $K = 80$ and $K = 100$, respectively, to complete their execution, and hence, their graph starts after those many rounds. Thereby, it has fewer rounds remaining to explore the best arm when the algorithm is executed in the neighbourhood of k^* compared to FB-BAUB. Thus, the minimum budget requirement for FB-BAUB as a function of K is much less than that of Seq. Halv. In addition, FB-BAUB has better error probability performance. This demonstrates the advantage of exploiting the unimodality of the reward function.

7.2 Comparison with pure exploration algorithms for monotone structure

We have considered two more experimental set-ups to compare FB-BAUB with the state-of-the-art algorithms. We consider K Gaussian arms with known variances $\sigma_i^2 = \sigma^2 = 0.5$, assuming that the mean of the best arm is $\mu_1 = 0.7$. Our simulations are averaged over 1000 runs and are shown with confidence intervals.

Experiment 3: $\mu_1 = 0.7$ and $\mu_{2:K} = \mu_1 - \frac{0.6(i-1)}{K-1}$.

Experiment 4: $\mu_1 = 0.7$ and $\mu_{2:K} = \mu_1 - 0.01\left(1 + \frac{4}{K}\right)^{i-2}$.

Fig. 4 and Fig. 5 illustrate the performance of each algorithm for Exp. 3 and Exp. 4, respectively, for two values of $K = \{80, 100\}$. Exp. 3 and Exp. 4 follow the monotone structure with the first arm as the optimal arm, similar to that considered in Shahrampour et al. (2017); Audibert et al. (2010). The error probability increases as we increase the arm size. In Exp. 3, the sub-optimal gap decreases in arithmetic progression, whereas in Exp. 4, the sub-optimal gap decreases in geometric progression. This makes Exp. 4 more challenging to find the best arm within a fixed budget, and thereby, the error probability of FB-BAUB is lower in Exp. 3 than in Exp. 4.

Furthermore, in Exp. 3, FB-BAUB can identify the optimal arm with a probability of more than 80% within 100 budget for 80 arms, and in Exp. 4, it identifies the optimal arm with a probability of more than 75% within 100 budget. Whereas, the other state-of-the-art algorithms need at least a 600 budget for execution. Hence, FB-BAUB outperforms the other state-of-the-art algorithms for Exp. 3 and 4.

8 Conclusion

We studied the fixed budget BAI problem with an unimodal structure on a finite set of arm means. We developed an algorithm named FB-BAUB to address the problem and derived an upper bound on its error probabilities. The algorithm works in phases and identifies the best arm with high probability. We established that the exponent in the error bound is independent of K in contrast to the unstructured bandits. We demonstrated that for any optimal algorithm, the error exponent should be independent of K by establishing a lower bound. Simulations validated the efficiency of FB-BAUB compared to state-of-the-art algorithms. FB-BAUB is parameter-free and easy to implement.

Many interesting research directions could be further investigated. The exponent in the upper bound on the error probability of FB-BAUB is optimal in T and K, but not in the problem-dependent complexity terms, which are characterized in terms of the smallest gap between any neighbouring arms. However, the lower bound is only dependent on the neighbours of the optimal arm. It is interesting to develop algorithms that are also optimal with respect to the complexity terms.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Alexia Atsidakou, Sumeet Katariya, Sujay Sanghavi, and Branislav Kveton. Bayesian Fixed-Budget Best-Arm Identification. *arXiv preprint arXiv:2211.08572*, 2022.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53, 2010.

MohammadJavad Azizi, Branislav Kveton, and Mohammad Ghavamzadeh. Fixed-Budget Best-Arm Identification in Structured Bandits. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2798–2804, 7 2022.

Antoine Barrier, Aurélien Garivier, and Gilles Stoltz. On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits. In *International Conference on Algorithmic Learning Theory*, pp. 136–181. PMLR, 2023.

Nathan Blinn, Jana Boerger, and Matthieu Bloch. mmWave Beam Steering with Hierarchical Optimal Sampling for Unimodal Bandits. In *ICC 2021-IEEE International Conference on Communications*, pp. 1–6. IEEE, 2021.

Djallel Bouneffouf and Irina Rish. A Survey on Practical Applications of Multi-Armed and Contextual Bandits. *CoRR*, abs/1904.10040, 2019. URL http://arxiv.org/abs/1904.10040.

Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pp. 590–604. PMLR, 2016.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012. JCSS Special Issue: Cloud Computing 2011.

Lijie Chen and Jian Li. On the Optimal Sample Complexity for Best Arm Identification, 2015.

James Cheshire, Pierre Ménard, and Alexandra Carpentier. The influence of shape constraints on the thresholding bandit problem. In *Conference on Learning Theory*, pp. 1228–1275. PMLR, 2020.

James Cheshire, Pierre, Ménard, and Alexandra Carpentier. Problem Dependent View on Structured Thresholding Bandit Problems. In *International Conference on Machine Learning*, pp. 1846–1854. PMLR, 2021.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Richard Combes and Alexandre Proutiere. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–521–I–529, 2014.

Richard Combes, Mazraeh Shahi Talebi, Sadegh Mohammad, Alexandre Proutiere, and Marc Lelarge. Combinatorial Bandits Revisited. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Conference on Learning Theory*, 2008.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *J. Mach. Learn. Res.*, 7, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

A. Garivier, P. Ménard, L. Rossi, and P. Menard. Thresholding Bandit for Dose-ranging: The Impact of Monotonicity. In *Arxiv*. PMLR, 2017. URL `https://doi.org/10.48550/arXiv.1711.04454`.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Manjesh Kumar Hanawal, Venkatesh Saligrama, Michal Valko, and Rémi Munos. Cheap Bandits. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

Morteza Hashemi, Ashutosh Sabharwal, C. Emre Koksal, and Ness B. Shroff. Efficient Beam Alignment in Millimeter Wave Systems Using Contextual Bandits. In *IEEE Conference on Computer Communications (INFOCOM)*, pp. 2393–2401, 2018.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC Subset Selection in Stochastic Multi-Armed Bandits. In *Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML)*, pp. 227–234, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246. PMLR, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17:1–42, 2016.

Tomáš Kocák and Aurélien Garivier. Best arm identification in spectral bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pp. 975–999. PMLR, 2014.

Shie Mannor and John N. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *J. Mach. Learn. Res.*, 5:623–648, dec 2004. ISSN 1532-4435.

Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Forced-exploration free strategies for unimodal bandits. *arXiv preprint arXiv:2006.16569*, 2020a.

Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Forced-exploration free Strategies for Unimodal Bandits. working paper or preprint, 2020b. URL `https://hal.archives-ouvertes.fr/hal-02883907`.

Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing*, 65(16):4281–4292, 2017.

Marta Soare, Alessandro Lazaric, and Remi Munos. Best-Arm Identification in Linear Bandits. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomas Kocak. Spectral Bandits for Smooth Graph ˇ Functions. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, 2014.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

Jia Yuan Yu and Shie Mannor. Unimodal Bandits. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 41–48, 2011.

A Appendix

A.1 Proof of Inequalities [Eq. (3)]

Proof. Note that in Audibert et al. (2010), the following inequality holds:

$$H_2 \leq H_1 \leq \log(2K) H_2 \leq 2\log(K) H_2 \tag{10}$$

By the definition of $\bar{H}$, we have

$$\bar{H} = \sum_{k\in\{k^*-1,k^*+1\}} \frac{1}{(\mu_{k^*}-\mu_k)^2} \leq \sum_{k\neq k^*} \frac{1}{(\mu_{k^*}-\mu_k)^2} = H_1 \quad \text{(By the definition of } H_1\text{)} \tag{11}$$

Combining (10) and (11) we obtain the given inequality in (3). □

A.2 Proof of Theorem 1

First, we state the following Theorem and Corollary to prove Theorem 1.

Theorem 3. *For any bandit strategy that returns the arm k_T after T budget, it holds that*

$$\max_{i\in\{k^*-1,k^*+1\}} P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(k_T \neq i) \geq \frac{1}{6}\exp\left(-12\frac{T}{\bar{H}(k^*)} - 2\sqrt{12\frac{T}{\bar{H}(k^*)}\log(6TK)}\right), \tag{12}$$

and also

$$\max_{i\in\{k^*-1,k^*+1\}} \left[P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(k_T \neq i)\exp\left(12\frac{T}{\bar{h}\bar{H}(i)} + 2\sqrt{12\frac{T}{\bar{h}\bar{H}(i)}\log(6TK)}\right)\right] \geq \frac{1}{6}. \tag{13}$$

Proof. The proof of this theorem follows the lines similar to Carpentier & Locatelli (2016)[Thm. 2] after applying the change of measure on the restricted set of arms.

Fix $\beta = 1$. Let $\mathbf{p}(\boldsymbol{\mu}) := \{p_k(\mu_k)\}_{k\in\mathcal{A}} \in \epsilon_U$ be a unimodal bandit instance such that $p_k(\mu_k) := N(\mu_k, 1)$, where $\mu_k \in [1/4, 1/2]$ for all $k \in \mathcal{A}$ and $\mu_{k^*} = 1/2$. We also consider $\mu_1 < \mu_2 < \cdots < \mu_{k^*-1} < \mu_{k^*} > \mu_{k^*+1} > \cdots > \mu_K$. We would like to find the lower bound of the probability that the learner fails to recommend the optimal arm when presented with instance $\boldsymbol{\mu}$, i.e., $P(k_T \neq k^*)$.

We define $d_k := \mu_{k^*} - \mu_k = \frac{1}{2} - \mu_k$, for any $1 \leq k \leq K$. Set $\Delta^i_k = d_i + d_k$, if $k \neq i$ and $\Delta^i_i = d_i$, for any $i \in \{k^* - 1, k^* + 1\}$ and any $k \in \{1, \ldots, K\}$. Note that $\{\Delta^i_k\}_k$ denotes the arm gaps of the bandit problem i.

We adapt the proof of Carpentier & Locatelli (2016)[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.

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

  We have considered K pairs of Gaussian arms where $\mathbf{p}(\boldsymbol{\mu}) := \{p_k(\mu_k)\}_{k\in\mathcal{A}} \in \epsilon_U$ be a unimodal bandit instance such that $p_k(\mu_k) := N(\mu_k, 1)$, where $\mu_k \in [1/4, 1/2]$ for all $k \in \mathcal{A}$ and $\mu_{k^*} = 1/2$, and $\mathbf{p}'(\boldsymbol{\mu}') := \{p'_k(\mu'_k)\}_{k\in\mathcal{A}}$ be another bandit instance where $p'_k(\mu'_k) := N(\mu'_k, 1)$ and $\mu'_k = 2\mu_{k^*} - \mu_k$ for all $k \in \mathcal{A}$.

  According to Carpentier & Locatelli (2016)[Thm. 2], we define K Gaussian bandit problem using "flipping constructions" where for the bandit problem $\mathbf{p}^i(\boldsymbol{\mu}^i) := p^i_1(\mu^i_1) \otimes p^i_2(\mu^i_2) \otimes \cdots \otimes p^i_K(\mu^i_K)$ with means $\boldsymbol{\mu}^i$, arm i is the optimal arm with distribution

  $$p^i_k(\mu^i_k) = \begin{cases} p_k(\mu_k), & \text{if } i \neq k \\ p'_k(\mu'_k), & \text{if } i = k. \end{cases}$$

  Hence, the bandit problem $\mathbf{p}^i(\boldsymbol{\mu}^i)$ 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 problem except for the bandit problems $\mathbf{p}^i(\boldsymbol{\mu}^i)$ where $i \in \{k^* - 1, k^*, k^* + 1\}$. For simplicity, we will denote bandit problem $\mathbf{p}^i(\boldsymbol{\mu}^i)$ as i. Note that the bandit problem $\mathbf{p}^{k^*}(\boldsymbol{\mu}^{k^*}) = \mathbf{p}(\boldsymbol{\mu})$.

  Hence we will focus on the three bandit problems $\mathbf{p}^i(\boldsymbol{\mu}^i)$ for $i \in \{k^* - 1, k^*, k^* + 1\}$ that follows unimodal structure. For $i \in [K]$, we use the notation $P_{\mathbf{p}^i(\boldsymbol{\mu}^i)}(.)$ and $E_{\mathbf{p}^i(\boldsymbol{\mu}^i)}(.)$ to denote the probability and expectation, respectively, with respect to the randomness of sampling for bandit problem i.

- **Step 2: Definition of high probability event and concentration of empirical KL divergences**

  For two distributions p and p' defined on $\mathbb{R}$ and p is absolutely continuous with respect to p', the Kullback Leibler (KL) divergence between distribution p and p', can be written as

  $$\mathrm{KL}(p, p') = \int_{\mathbb{R}} \log\left(\frac{dp(x)}{dp'(x)}\right) dp(x),$$

  Following the lines of proof of Atsidakou et al. (2022)[Thm. 9], for $k \in \{1, 2, \ldots, K\}$, the KL divergence between two Gaussian distributions $p_k(\mu_k)$ and $p'_k(\mu'_k)$ is given by

  $$\mathrm{KL}_k := \mathrm{KL}\Big(p_k(\mu_k), p'_k(\mu'_k)\Big) = \frac{(\mu_k - \mu'_k)^2}{2} = 2d_k^2. \tag{14}$$

  Let us consider $1 \leq t \leq T$. We define the quantity as

  $$\widehat{\mathrm{KL}}_{k,t} = \frac{1}{t}\sum_{s=1}^{t} \log\left(\frac{dp_k(\mu_k)}{dp'_k(\mu'_k)}(X_{k,s})\right) = \frac{1}{t}\sum_{s=1}^{t} 2(X_{k,s} - \mu_{k^*})d_k$$

  where $X_{k,s}$ are independently and identically (i.i.d.) distributed as $p^i_k(\mu^i_k)$ for $s \leq t$ and bandit problem i.

Note that

$$\mathbb{E}_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left[\widehat{KL}_{k,t}\right] = \begin{cases} 2(\mu_k - \mu_{k^*})d_k = -KL_k, k \neq i \\ 2(\mu_k + 2d_k - \mu_{k^*})d_k = KL_k, k = i \end{cases}$$

This implies that $\left|\widehat{KL}_{k,t}\right|$ is an unbiased estimator of KL_k.

Let us define an event as follows:

$$\zeta = \left\{\forall 1 \leq k \leq K, \forall 1 \leq t \leq T, \left|\widehat{\mathrm{KL}}_{k,t}\right| - \mathrm{KL}_k \leq 2d_k\sqrt{\frac{2\log(6TK)}{t}}\right\}. \tag{15}$$

According to Carpentier & Locatelli (2016)[Lemma 1], the concentration bound for $\left|\widehat{KL}_{k,t}\right|$ that holds for the bandit problem i where $i \in \{k^*-1, k^*, k^*+1\}$ is given by,

$$P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(\zeta) \geq \frac{5}{6}, \quad \text{for } i \in \{k^*-1, k^*, k^*+1\}. \tag{16}$$

- **Step 3: A change of measure**

  Let *Alg* denote the active strategy of the learner that returns some arm k_T at the end of the budget T. 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 $\sum_{1\leq k\leq K} T_k = T$.

  Let us write for any $0 \leq k \leq K$,

  $$t_k = \mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}\left[T_k\right] \quad \text{and} \quad \sum_{1\leq k\leq K} t_k = T.$$

  We recall the change of measure identity, refer Audibert et al. (2010), which states that for any measurable event ξ and for any $i \in \{k^* - 1k^* + 1\}$, we have

  $$P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(\xi) = \mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}\left[1\{\xi\}\exp\left(-T_i\widehat{KL}_{i,T_i}\right)\right] \tag{17}$$

  as the product distributions $\mathbf{p}^i(\boldsymbol{\mu}^i)$ and $\mathbf{p}^{k^*}(\boldsymbol{\mu}^{k^*})$ only differ in arm i and as the active strategy only explored the samples $\{X_{k,s}\}_{k\leq K, s\leq T_k}$

  We now consider the event ξ_i as the event where the algorithm outputs arm k^* at the end of T budget, where ζ holds, and where the number of times arm i was pulled is smaller than $6t_i$, i.e.,

  $$\xi_i = \left\{k_T = k^*\right\} \cap \left\{\zeta\right\} \cap \left\{T_i \leq 6t_i\right\}, \tag{18}$$

  for $i \in \{k^*-1, k^*+1\}$. Applying the event ξ_i as given by (18) in (17) we obtain,

  $$P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(\xi) = \mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}\left[1\{\xi_i\}\exp\left(-T_i\widehat{KL}_{i,T_i}\right)\right]$$

  Following the same lines of proof as given in Carpentier & Locatelli (2016)[Step 2, Thm. 2] and in Atsidakou et al. (2022)[Lemma 17] and applying (14), for $i \in \{k^*-1, k^*+1\}$, we get

  $$P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(\xi_i) \geq P_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}(\xi_i)\exp\left(-12t_i d_i^2 - 2\sqrt{12t_i d_i^2\log(6TK)}\right). \tag{19}$$

- **Step 4: Lower bound on $\mathbb{P}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}(\xi_i)$ for any reasonable algorithm**

  Let us assume that the probability that *Alg* makes a mistake on problem k^* is less than 1/2, i.e.,

  $$\mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}\left[k_T \neq k^*\right] \leq \frac{1}{2} \tag{20}$$

  Note that if *Alg* does not satisfy that, it performs badly on bandit problem k^*, and its probability of success is not larger than $\frac{1}{2}$ 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

  $$P_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}(T_k \geq 6t_k) \leq \frac{\mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}[T_k]}{6t_k} = \frac{1}{6}, \tag{21}$$

  since $\mathbb{E}_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}[T_k] = t_k$ for Algorithm *Alg*.

  Therefore, by combining (20), (21) and (16), it holds by an union bound that for any $i \in \{k^*-1, k^*+1\}$

  $$P_{\boldsymbol{p}^{k^*}(\boldsymbol{\mu}^{k^*})}(\xi_i) \geq 1 - \left(\frac{1}{6} + \frac{1}{2} + \frac{1}{6}\right) = \frac{1}{6}. \tag{22}$$

  We will now combine (22) and the fact that $P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \geq P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}(\xi_i)$ for $i \in \{k^*-1, k^*, k^*+1\}$ and by applying in (17), we obtain

  $$P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \geq \frac{1}{6}\exp\left(-12t_i d_i^2 - 2\sqrt{12t_i d_i^2 \log(6TK)}\right) \tag{23}$$

- **Step 5: Conclusions**

  We defined $\bar{H}(k^*) := \sum_{k \in \{k^*-1, k^*+1\}} \frac{1}{(\Delta_k^{k^*})^2}$. We also know that $\sum_{1 \leq k \leq K} t_k = T$. By combining these two facts, we can say that there exists $i \in \{k^*-1, k^*+1\}$ such that

  $$t_i \leq \frac{T}{\bar{H}(k^*) d_i^2}$$

  as the contraposition yields an immediate contradiction. For this i, it holds by (23) that

  $$\max_{i \in \{k^*-1, k^*+1\}} P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \geq \frac{1}{6}\exp\left(-12\frac{T}{\bar{H}(k^*)} - 2\sqrt{12\frac{T}{\bar{H}(k^*)}\log(6TK)}\right). \tag{24}$$

  This concludes the proof of the first part of the theorem.

  We have, $\mu_k = \frac{1}{2} - d_k$ such that $\mu_k \in [1/4, 1/2], \forall k \in \mathcal{A}$ and $\mu_{k^*} = \frac{1}{2}$. Note that $\boldsymbol{\mu}$ exhibits a unimodal structure. Since $\bar{h} = \sum_{i \in \{k^*-1, k^*+1\}} \frac{1}{d_i^2 \bar{H}(i)}$ and since $\sum_{1 \leq k \leq K} t_k = T$, then there exists $i \in \{k^*-1, k^*+1\}$ such that

  $$t_i \leq \frac{T}{\bar{h}\bar{H}(i) d_i^2}$$

  Therefore, for these $i \in \{k^*-1, k^*+1\}$ and by (23) we get

  $$\max_{i \in \{k^*-1, k^*+1\}} \left[P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \exp\left(12\frac{T}{\bar{h}\bar{H}(i)} + 2\sqrt{12\frac{T}{\bar{h}\bar{H}(i)}\log(6TK)}\right)\right] \geq \frac{1}{6}. \tag{25}$$

  This concludes the proof of the second part of the theorem. □

Corollary 1. *Assume that* $T \geq \max_{i \in \{k^*-1, k^*+1\}} \left(\bar{H}(k^*), \bar{H}(i)\bar{h}\right) \frac{4\log(6TK)}{12}$*. For any bandit strategy that returns the arm* $\hat{k}_T$ *at time* T*, it holds that*

$$\max_{i \in \{k^*-1, k^*+1\}} P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \geq \frac{1}{6} \exp\left(-24\frac{T}{\bar{H}(k^*)}\right),$$

and also

$$\max_{i \in \{k^*-1, k^*+1\}} \left[P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \exp\left(24\frac{T}{\bar{H}(i)\bar{h}}\right)\right] \geq \frac{1}{6}.$$

Proof. We assumed that

$$T \geq \max_{i \in \{k^*-1, k^*+1\}} \left(\bar{H}(k^*), \bar{H}(i)\bar{h}\right) \frac{4\log(6TK)}{12}. \tag{26}$$

- **Case 1:** Let us consider

  $$\max\left(\bar{H}(k^*), \bar{H}(i)\bar{h}\right) = \bar{H}(k^*) \tag{27}$$

  Applying (27) in (26) we obtain,

  $$\frac{12T}{\bar{H}(k^*)} \geq 2\sqrt{12\frac{T}{\bar{H}(k^*)}\log(6TK)} \tag{28}$$

  Applying (28) in (24) we get

  $$\max_{i \in \{k^*-1, k^*+1\}} P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \geq \frac{1}{6} \exp\left(-24\frac{T}{\bar{H}(k^*)}\right) \tag{29}$$

  This concludes the proof of the first part of the corollary.

- **Case 2:** Let us consider for each $i \in \{k^*-1, k^*+1\}$

  $$\max\left(\bar{H}(k^*), \bar{H}(i)\bar{h}\right) = \bar{H}(i)\bar{h} \tag{30}$$

  Applying (30) in (26) we obtain,

  $$\frac{12T}{\bar{h}\bar{H}(i)} \geq 2\sqrt{12\frac{T}{\bar{h}\bar{H}(i)}\log(6TK)} \tag{31}$$

  Applying (31) in (25) we get

  $$\max_{i \in \{k^*-1, k^*+1\}} \left[P_{\boldsymbol{p}^i(\boldsymbol{\mu}^i)}\left(k_T \neq i\right) \exp\left(24\frac{T}{\bar{H}(i)\bar{h}}\right)\right] \geq \frac{1}{6}. \tag{32}$$

  This concludes the proof of the second part of the corollary. □

We will now prove Theorem 1 using this corollary.

Proof. We have $\bar{h}$ defined as

$$\bar{h} = \sum_{i \in \{k^*-1, k^*+1\}} \frac{1}{d_i^2 \bar{H}(i)}$$

Let us denote

$$(I) := \frac{1}{d_{k^*-1}^2 \bar{H}(k^*-1)} \tag{33}$$

$$(II) := \frac{1}{d_{k^*+1}^2 \bar{H}(k^*+1)} \tag{34}$$

Therefore, $\bar{h}$ can be written as

$$\bar{h} = (I) + (II). \tag{35}$$

We will give an upper bound of (I) and (II).

Using the definition of $\bar{H}(k^*-1)$ (see 2), we get

$$d_{k^*-1}^2 \bar{H}(k^*-1) = d_{k^*-1}^2 \sum_{k \in \{k^*-2, k^*\}} \frac{1}{(d_{k^*-1} + d_k)^2}$$

Since $d_{k^*} = 0$ and $d_{k^*-2} \geq d_{k^*-1}$, we get

$$d_{k^*-1}^2 \bar{H}(k^*-1) \leq 1 + \frac{1}{4} = \frac{5}{4} \tag{36}$$

By applying (36) in (33) we get

$$(I) \leq \frac{4}{5} \tag{37}$$

Using the definition of $\bar{H}(k^*+1)$ ((see 2)) we get

$$d_{k^*+1}^2 \bar{H}(k^*+1) = d_{k^*+1}^2 \sum_{k \in \{k^*, k^*+2\}} \frac{1}{(d_{k^*+1} + d_k)^2}$$

Since $d_{k^*} = 0$ and $d_{k^*+2} \geq d_{k^*+1}$, we get

$$d_{k^*+1}^2 \bar{H}(k^*+1) \leq 1 + \frac{1}{4} = \frac{5}{4}. \tag{38}$$

By applying (38) in (34) we get

$$(II) \leq \frac{4}{5} \tag{39}$$

Applying (37) and (39) in (35), we get

$$\bar{h} \geq \frac{4}{5} + \frac{4}{5} = \frac{8}{5}.$$

Putting the value of $\bar{h}$ in Corollary, we get the required bound as given in (4). □

A.3 Proof of Theorem 2

Proof. We will upper bound the error probability as given by $P_{\boldsymbol{p}(\boldsymbol{\mu})}(\hat{k}_{L+1} \neq k^*)$. For simplicity of notation, we will drop $\boldsymbol{p}(\boldsymbol{\mu})$ and we refer to it as $P(\hat{k}_{L+1} \neq k^*)$. The FB-BAUB runs for T horizon in $L+1$ number of phases that satisfies (6), where $L = \frac{\log_2 K/3}{\log_2 3/2}$ and outputs the arm $\hat{k}_{L+1}$. We will now upper bound the probability of error as,

$$P(\hat{k}_{L+1} \neq k^*) \leq \sum_{l=1}^{L+1} P(k^* \text{ elim. in } l). \tag{40}$$

The best arm is eliminated in phase l in the following cases:

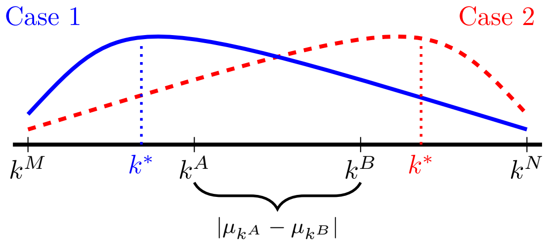

Figure 6: Different cases of elimination in any phase l. k^* will not get eliminated if it is in between arms k^A and k^B.

1. $k^* \in \{k^M, \dots, k^A\}$, and $\hat{\mu}^l_{k^B}$ or $\hat{\mu}^l_{k^N}$ is greater than both $\hat{\mu}^l_{x^M}$ and $\hat{\mu}^l_{k^A}$
2. $k^* \in \{k^B, \dots, k^N\}$, and $\hat{\mu}^l_{k^M}$ or $\hat{\mu}^l_{k^A}$ is greater than both $\hat{\mu}^l_{k^B}$ and $\hat{\mu}^l_{k^N}$

The two cases are illustrated in Fig. 6. From Remark 1, k^* will not get eliminated if $k^* \in \{k^A, \dots, k^B\}$. However, we will upper bound the error probability by assuming that k^* will always fall in the above two cases. Notice that Case 1 and Case 2 are symmetrical. Hence we can consider that k^* will always fall in either one of the cases. Without loss of generality, we consider Case 1.

$$\begin{aligned} P(k^* \text{ elim. in } l) &\leq P(\hat{\mu}^l_{k^B} > \hat{\mu}^l_{k^M} \text{and } \hat{\mu}^l_{k^A} | k^* \in \{k^M, \dots, k^A\}) \\ &\quad + P(\hat{\mu}^l_{k^N} > \hat{\mu}^l_{k^M} \text{and } \hat{\mu}^l_{k^A} | k^* \in \{k^M, \dots, k^A\}) \\ &\leq 2P(\hat{\mu}^l_{k^B} > \hat{\mu}^l_{k^M} \text{and } \hat{\mu}^l_{k^A} | k^* \in \{k^M, \dots, k^A\}), \end{aligned} \tag{41}$$

where the last inequality is due to the fact that, for Case 1, $\mu_{k^B} \geq \mu_{k^N}$ by unimodality. Now for Case 1, μ_{k^A} is always greater than μ_{k^B}, but μ_{k^M} may not be greater than μ_{k^B}. Then, we can further upper bound (41) as

$$P(k^* \text{ elim. in } l) \leq 2P(\hat{\mu}^l_{k^B} > \hat{\mu}^l_{k^A} | k^* \in \{k^M, .., k^A\}). \tag{42}$$

We will now apply Hoeffding's inequality Lattimore & Szepesvári (2020) in (42) as stated in the following Lemma.

Lemma 1 (Hoeffding's Inequality for Subgaussian Random Variables)**.** *If $X_1, \dots, X_m$ are m i.i.d samples drawn from β-Subgaussian then for any $i \in [m]$, then*

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

Thereby, applying Lemma 1 in (42), we have

$$P(\hat{\mu}^l_{k^B} > \hat{\mu}^l_{k^A}) \leq \exp\left\{-\frac{1}{2}\frac{N_l}{4}\left(\frac{\Delta^l_{A,B}}{\beta}\right)^2\right\}, \tag{43}$$

where $\Delta^l_{A,B} = \mu_{k^A} - \mu_{k^B}$ for phase l and is greater than 0 for Case 1. Using Δ, which is defined as $\Delta = \min_{2 \leq i \leq K-1} |\mu_i - \mu_{i-1}|$, and the fact that there are at least $\frac{j_l}{3}$ arms between k^A and k^B, for Case 1 we have, $\Delta^l_{A,B} \geq (j_l/3)\Delta$. Thus from (41) and (43) we have,

$$P(k^* \text{ elim. in } l) \leq 2\exp\left\{-\frac{N_l}{72}\left(j_l\frac{\Delta}{\beta}\right)^2\right\}. \tag{44}$$

Using $j_l = \left(\frac{2}{3}\right)^{l-1} K$ in (44), we can find the probability of the best arm getting eliminated in phase 1 and 2, phase $L+1$, and the rest of the phases separately. Using (8), we have

$$P(k^* \text{ elim. in } 1\&2) \leq 2\exp\left\{-\frac{TK}{32}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2\exp\left\{-\frac{TK}{72}\left(\frac{\Delta}{\beta}\right)^2\right\}. \tag{45}$$

For phase $L+1$, since the best arm is selected among three arms when each arm is sampled $T/9$ times, we have

$$P(k^* \text{ elim. in phase } L+1) \leq 2\exp\left\{-\frac{T}{24}\left(\frac{\Delta}{\beta}\right)^2\right\}. \tag{46}$$

From (44), the error probability for the remaining phases is

$$\begin{aligned} P(\text{best arm elim. in phase 3 to phase L}) &\leq 2\sum_{l=3}^{L} \exp\left\{-\frac{T}{8}\frac{K^2}{9}\left(\frac{2}{3}\right)^{2(l-1)}\frac{2^{L-l+1}}{3^{L-l+2}}\left(\frac{\Delta}{\beta}\right)^2\right\} \\ &= 2\sum_{l=3}^{L} \exp\left\{-\frac{TK}{24}\left(\frac{2}{3}\right)^{l}\left(\frac{\Delta}{\beta}\right)^2\right\} \\ &\leq 2(L-2)\exp\left\{-\frac{T}{8}\left(\frac{\Delta}{\beta}\right)^2\right\}. \end{aligned} \tag{47}$$

By (40), (45), (46) and (47), we obtain the upper bound as

$$\begin{aligned} P(\hat{k}_{L+1} \neq k^*) \leq\ & 2\exp\left\{-\frac{T}{24}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2\exp\left\{-\frac{TK}{32}\left(\frac{\Delta}{\beta}\right)^2\right) \\ & + 2\exp\left\{-\frac{TK}{72}\left(\frac{\Delta}{\beta}\right)^2\right\} + 2(L-2)\exp\left\{-\frac{T}{8}\left(\frac{\Delta}{\beta}\right)^2\right\}. \end{aligned} \tag{48}$$

□