The Influence of Shape Constraints on the Thresholding Bandit Problem

James Cheshire, Pierre Ménard, Alexandra Carpentier

▶ **To cite this version:**

James Cheshire, Pierre Ménard, Alexandra Carpentier. The Influence of Shape Constraints on the Thresholding Bandit Problem. COLT 2020 - Thirty Third Conference on Learning Theory, Jul 2020, Graz / Virtual, Austria. pp.1228-1275. hal-03001947v2

HAL Id: hal-03001947
https://hal.science/hal-03001947v2

Submitted on 22 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Proceedings of Machine Learning Research vol 125:1–48, 2020 33rd Annual Conference on Learning Theory

The Influence of Shape Constraints on the Thresholding Bandit Problem

James Cheshire JAMES.CHESHIRE@OVGU.DE
Otto von Guericke University Magdeburg

Pierre Menard PIERRE.MENARD@INRIA.FR
Centre Inria Lille - Nord Europe

Alexandra Carpentier ALEXANDRA.CARPENTIER@OVGU.DE
Otto von Guericke University Magdeburg

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We investigate the stochastic *Thresholding Bandit problem* (*TBP*) under several *shape constraints*. On top of (i) the vanilla, unstructured *TBP*, we consider the case where (ii) the sequence of arm's means $(\mu_k)_k$ is monotonically increasing *MTBP*, (iii) the case where $(\mu_k)_k$ is unimodal *UTBP* and (iv) the case where $(\mu_k)_k$ is concave *CTBP*. In the *TBP* problem the aim is to output, at the end of the sequential game, the set of arms whose means are above a given threshold. The regret is the highest gap between a misclassified arm and the threshold. In the fixed budget setting, we provide *problem independent* minimax rates for the expected regret in all settings, as well as associated algorithms. We prove that the minimax rates for the regret are (i) $\sqrt{\log(K)K/T}$ for *TBP*, (ii) $\sqrt{\log(K)/T}$ for *MTBP*, (iii) $\sqrt{K/T}$ for *UTBP* and (iv) $\sqrt{\log\log K/T}$ for *CTBP*, where K is the number of arms and T is the budget. These rates demonstrate that *the dependence on* K of the minimax regret varies significantly depending on the shape constraint. This highlights the fact that the shape constraints modify fundamentally the nature of the *TBP* problem to the other.

1. Introduction

Stochastic multi-armed bandit problems consider situations in which a learner faces multiple unknown probability distributions, or "arms", and has to sequentially sample these arms.

In this paper, we focus on the Thresholding Bandit Problem (*TBP*), a *Combinatorial Pure Exploration (CPE)* bandit setting introduced by Chen et al. (2014). The learner is presented with $[K] = \{1, \dots, K\}$ arms, each following an unknown distribution ν_k with unknown mean μ_k. Given a budget $T > 0$, the learner samples the arms sequentially for a total of T times and then aims at predicting the set of arms whose mean is above a given threshold $\tau \in \mathbb{R}$. The performance of the learner is measured through the *expected simple regret* which in this setting is the expected maximal gap between τ and the mean of a misclassified arm. Note that our problem is in fact akin to estimating in a sequential setting a given level-set of a discrete function under shape constraints.

In this paper we will be interested only in the *problem independent* case, and want to characterise the *minimax-order* of the expected simple regret on various sets of bandit problems. In particular we study the influence of various *shape constraints on the sequence*

© 2020 J. Cheshire, P. Menard & A. Carpentier.

of means of the arms, on the *TBP* problem, i.e. see how classical shape constraints influence the minimax rate of the expected simple regret. We will consider four shape constraints.

Vanilla, unstructured case *TBP* First we consider the vanilla case where we only assume that the distributions of the arms are supported in $[0, 1]$. We will refer to this case as the unstructured problem, (*TBP*). The fixed confidence version of *TBP* was studied in Chen et al. (2014, 2016) -see also e.g. Even-Dar et al. (2002); Chen and Li (2015); Simchowitz et al. (2017); Garivier and Kaufmann (2016) for papers in the related best arm identification and TOP-M setting[1] in the fixed confidence case. The fixed budget version of *TBP* was studied in Chen et al. (2014); Locatelli et al. (2016); Mukherjee et al. (2017), Zhong et al. (2017) - but also see e.g. Bubeck et al. (2009); Audibert and Bubeck (2010); Gabillon et al. (2012); Carpentier and Locatelli (2016) for papers in the related best arm identification and TOP-M setting in the fixed budget case. These papers almost exclusively concern the *problem dependent regime*, which is not the focus of this paper, and the adaptation of their rate to the problem independent case is sub-optimal, see the discussion under Theorem 1 for a more thorough comparison to this literature, and Appendix H for details.
In this paper, we prove that the minimax-optimal order of the expected simple regret in *TBP* is $\sqrt{\frac{K \log K}{T}}$. While a simple uniform-sampling strategy attains this bound, the lower bound is more interesting, in particular the presence of the $\sqrt{\log K}$ term. See the discussion following Theorem 1. For a discussion on the performance of the uniform-sampling strategy in the *problem dependent* regime, see Appendix H.

Monotone constraint, *MTBP*. We then consider the problem where on top of assuming that the distributions are supported in $[0, 1]$, we assume that the sequence of means $(\mu_k)_k$ is monotone - this is problem *MTBP*. This specific instance of the *TBP* is introduced within the context of drug dosing in Garivier et al. (2017). In this paper, the authors provide an algorithm for the fixed confidence setting that is optimal from a *problem dependent* point of view. The shape constraint on the means of the arms implies that the *MTBP* is related to *noisy binary search*, i.e. inserting an element into its correct place within an ordered list when only noisy labels of the elements are observed, see Feige et al. (1994). In the noiseless case, an effective approach due to the shape constraint is to conduct a binary search - and the classification of the arms can therefore be performed in just $O(\log(K))$ steps, while K steps are needed in the noiseless *TBP*. It is therefore clear that *MTBP* is radically different from *TBP*, even in the noiseless case. In the noisy case, the learner has to sample many times each arm in order to get a reliable decision at each step. While a simple naive strategy, although sufficient in Xu et al. (2019), is to do *noisy binary search* where at each step the learner simply samples about $O(T/\log(K))$ times each arm, there are clear hints from the literature that in the *MTBP* this is not going to be optimal. For the related yet different problem of noisy binary search, Feige et al. (1994), Ben-Or and Hassidim (2008) and Emamjomeh-Zadeh et al. (2016) solve this issue by introducing a noisy binary search *with corrections* - see also Nowak (2009), Karp and Kleinberg (2007). However, all these papers consider the problem of noisy binary search in settings with more structural assumptions and where the objective is more related to a fixed confidence setting, their results are therefore not directly applicable to our setting. See the discussion under

1. In the TOP-M setting, the objective of the learner is to output the M arms with highest means. A popular version of it it is the TOP-1 problem where the aim is to find the arm that realises the maximum.

Theorem 2 for a more thorough comparison to this literature and see Appendix H for details. In this paper, we prove that the minimax-optimal order of expected simple regret in *MTBP* is $\sqrt{\log(K)/T}$. Interestingly and as highlighted in this paragraph, this rate is much smaller than the minimax rate over *TBP*. This reflects the fact that the monotone shape constraint makes the problem much simpler than *TBP*, and closer to noisy binary search. Further discussion on the comparison between the *TBP* and *MTBP*, specifically the difference coming from the monotone assumption, can be found in Appendix H and see the algorithm `Explore` and the associated text in Section 4 for more intuition on the link to noisy binary search. Discussion on the performance of our algorithms for the *MTBP* in the *problem dependent* regime can also be found in Appendix H.

Unimodal constraint, *UTBP*. We also consider the problem where on top of assuming that the distributions are supported in $[0, 1]$, we assume that the sequence of means $(\mu_k)_k$ is unimodal - this is problem *UTBP*. It has not been considered to the best of our knowledge. However similar problems have been studied such that identifying the best arm or minimizing the cumulative regret Combes and Proutiere (2014a,b); Paladino et al. (2017); Yu and Mannor (2011). Paladino et al. (2017); Combes and Proutiere (2014b) focus on the *problem dependent regime*, and are not transferable - at least to the best of our knowledge - to the problem independent setting. Yu and Mannor (2011); Combes and Proutiere (2014a) are closer to our problem as it focuses on the problem independent regime. However, they consider the $\mathcal{X}$-armed setting (continuous set of arms e.g. in $[0, 1]$) setting and assume Hölder type regularity assumption around the maximum, which prevents jumps in the mean vector. These results therefore do not apply to our setting, where of course jumps are bound to happen as we are in the discrete setting. See the discussion under Theorem 3 for a more thorough comparison to this literature.
In this paper, we prove that the minimax-optimal order of the expected simple regret in *UTBP* is of order $\sqrt{K/T}$. This is interesting in contrast to the rate of *MTBP*. Monotone bandit problems are much easier than unimodal bandit problems - which can be written as a combination of a non-decreasing bandit problem, and a non-increasing bandit problem. This is however not very surprising, as finding the maximum of the unimodal bandit problem - i.e. the points where the non-increasing and non-decreasing bandit problems merge - is difficult.

Concave constraint, *CTBP*. Finally we consider the problem where on top of assuming that the distributions are supported in $[0, 1]$, we assume that the sequence of means $(\mu_k)_k$ is concave - this is problem *CTBP*. To the best of our knowledge this setting has not yet been consider in the literature. However, two related problems have been considered: the problem of estimating a concave function, and the problem of optimising a concave function - for both problems, mostly in the continuous setting, which renders a comparison with our setting delicate. The problem of estimating a concave function has been thoroughly studied in the noiseless setting, and also in the noisy setting, see e.g. Simchowitz et al. (2018), where the setting of a continuous set of arms is considered, under Hölder smoothness assumptions. The problem of optimising a convex function in noise without access to its derivative - namely zeroth order noisy optimisation - has also been extensively studied. See e.g. Nemirovski and Yudin. (1983)[Chapter 9], and Wang et al. (2017); Agarwal et al. (2011); Liang et al. (2014) to name a few, all of them in a continuous setting and in dimension d. The focus of this

Results	Unstructured *TBP*	Monotone *TBP*	Unimodal *TBP*	Concave *TBP*
Regret	$\sqrt{\frac{K \log K}{T}}$	$\sqrt{\frac{\log K}{T}}$	$\sqrt{\frac{K}{T}}$	$\sqrt{\frac{\operatorname{loglog} K}{T}}$

Table 1: Order of the minimax expected simple regret for the thresholding bandit problem, in the case of all four structural assumptions on the means of the arms considered in this paper. All results are given up to universal multiplicative constants.

literature is however very different than ours, as the main difficulty under their assumption is to obtain a good dependence in the dimension d, and in this setting logarithmic factors are not very relevant. See the discussion under Theorem 4 for a more thorough comparison to this literature.

In this paper, we prove that the minimax-optimal order of the expected simple regret in *CTBP* is $\sqrt{\operatorname{loglog}(K)/T}$. This is interesting in contrast to rate in the case of *UTBP*. Concave bandit problems are much easier than unimodal bandit problems. Also, if we compare with *MTBP*, we have that concave bandit problems are also much easier than monotone bandit problems, which is perhaps surprising - in particular the fact that the dependence in K is much smaller.

Organisation of the paper Our results are summarized in Table 1. See also Appendix A for an adaptation of these results in the $\mathcal{X}$-armed bandit setting. In Section 2 we define the setting and the *TBP*, *MTBP*, *CTBP* and *UTBP* problems. Minimax rates for the expected regret for all cases are given in Section 3. In Section 4 we describe algorithms attaining the minimax rates of Section 3, again for all cases. The Appendix contains the proofs for all results, as well as formulation of the upper and lower bounds leading to the minimax rates in a broader setting, transposition of some results in the fixed confidence setting, and also some additional discussions and remarks.

2. Problem formulation

The learner is presented with a K-armed bandit problem $\underline{\nu} = \{\nu_1, \dots, \nu_K\}$, with $K \geq 3$, where ν_k is the unknown distribution of arm k. Let $\tau \in \mathbb{R}$ be a fixed threshold known to the learner. We aim to devise an algorithm which classifies arms as above or below threshold τ. That is, the learner aims at finding the vector $Q \in \{-1, 1\}^K$ that encodes the true classification, i.e. $Q_k = 2\mathbb{1}_{\{\mu_k \geq \tau\}} - 1$ with the convention $Q_k = 1$ if arm k is above the threshold and $Q_k = -1$ otherwise.

The *fixed budget* bandit sequential learning setting goes as follows: the learner has a budget $T > 0$ and at each round $t \leq T$, the learner pulls an arm $k_t \in [1, K]$ and observes a sample $Y_t \sim \nu_{k_t}$, conditionally independent from the past. After interacting with the bandit problem and expending their budget, the learner outputs a vector $\hat{Q} \in \{-1, 1\}^K$ and the aim is that it matches the unknown vector Q as well as possible.

That is, the *fixed budget* objective of the learner following the strategy π is then to minimize the expected simple regret of this classification for $\hat{Q} := \hat{Q}^\pi$:

$$\bar{R}_T^{\underline{\nu},\pi} = \mathbb{E}_{\underline{\nu}}\left[\max_{\{k \in [K]:\ \hat{Q}_k^\pi \neq Q_k\}} \Delta_k\right],$$

where $\Delta_k := |\tau - \mu_k|$ is the gap of arm k, and where $\mathbb{E}_{\underline{\nu}}$ is defined as the expectation on problem $\underline{\nu}$ and $\mathbb{P}_{\underline{\nu}}$ the probability. We also write for the simple regret as a random variable $R_T^{\underline{\nu},\pi} = \max_{\{k\in[K]:\ \hat{Q}_k^\pi \neq Q_k\}} \Delta_k$. When it is clear from the context we will remove the dependence on the bandit problem $\underline{\nu}$ and/or the strategy π. We now present several sets of bandit problems that correspond to our four shape constraints.

Vanilla, unstructured case *TBP* We assume that the distribution of all the arms ν_k are supported in $[0,1]$. We denote by μ_k the mean or arm k. Let $\mathcal{B} := \mathcal{B}(K)$ be the set of such problems.

Monotone case *MTBP* We denote by $\mathcal{B}_m$ the set of bandit problems,

$$\mathcal{B}_m := \{\nu \in \mathcal{B} :\ \mu_1 \leq \mu_2 \leq \ldots \leq \mu_K\} ,$$

where the learner is given the additional information that the sequence of means $(\mu_k)_{k\in[K]}$ is a monotonically increasing sequence.

Unimodal case *UTBP* We will denote by $\mathcal{B}_u$ the set of bandit problems,

$$\mathcal{B}_u := \{\nu \in \mathcal{B} :\ \exists k^* \in [K] \text{s.t.} \forall l \leq k^*, \mu_{l-1} \leq \mu_l \text{ and } \forall l \geq k^*, \mu_l \geq \mu_{l+1}\} ,$$

where the learner is given the additional information that the sequence of means $(\mu_k)_{k\in[K]}$ is unimodal.

Concave case *CTBP* We will denote by $\mathcal{B}_c$ the set of bandit problems,

$$\mathcal{B}_c := \left\{\nu \in \mathcal{B} : \forall 1 < k < K-1, \frac{1}{2}\mu_{k-1} + \frac{1}{2}\mu_{k+1} \leq \mu_k\right\} ,$$

where the learner is given the additional information that the sequence of means $(\mu_k)_{k\in[K]}$ is concave.

Minimax expected regret Consider a set of bandit problems $\tilde{\mathcal{B}}$ - e.g. $\mathcal{B}_u, \mathcal{B}_m, \mathcal{B}_c, \mathcal{B}$. The minimax optimal expected regret on $\tilde{B}$ is then

$$\bar{R}_T^*(\tilde{\mathcal{B}}) := \inf_{\pi \text{ strategy}} \sup_{\underline{\nu}\in\tilde{\mathcal{B}}} \bar{R}_T^{\underline{\nu},\pi} .$$

3. Minimax expected regret for *TBP*, *MTBP*, *UTBP*, *CTBP*

In this section we present all minimax rates on the expected regret in the case of all four shape constraints. Algorithms achieving these mini-max rates are described in Section 4. For two positive sequences of real numbers $(a_n)_n, (b_n)_n$, we write $a_n \asymp b_n$ if there exists two *universal constants*[2] $0 < c < C$ such that $ca_n \leq b_n \leq Ca_n$.

Theorem 1 provides the minimax rate of the *TBP*. The proof can be found in Appendix C, i.e. Proposition 8, and Proposition 10.

Theorem 1 *It holds that*

$$\bar{R}_T^*(\mathcal{B}) \asymp \sqrt{\frac{K \log K}{T}}.$$

The algorithm `Uniform` *described in Sections 4 (see also Appendix C) attains this rate.*

2. In particular, independent of T, K.

It is difficult to compare this result to state of the art literature as existing papers consider almost exclusively the *problem dependent regime*, and often the fixed confidence setting. One can however deduce from Locatelli et al. (2016) an upper bound of order $\sqrt{K \log(K \log T/\delta)/T}$, and from Chen et al. (2016) a lower bound of order $\sqrt{K/T}$, which are both slightly sub-optimal.

Theorem 2 provides the minimax rate of the *MTBP*. The proof can be found in Appendix D, i.e. Proposition 11, and Corollary 13.

Theorem 2 *It holds that*

$$\bar{R}_T^*(\mathcal{B}_m) \asymp \sqrt{\frac{\log K}{T}}.$$

The algorithm MTB *described in Section 4 attains this rate.*

The literature that achieves results closest to this theorem is the noisy binary search literature cited in the introduction. The results that are most comparable to ours are the ones in Karp and Kleinberg (2007). They consider the special case where all arms follow a Bernoulli distribution with parameter p_k and $p_1 < ... < p_K$, and the aim is to find a i such that p_i is close to $1/2$. In the *fixed confidence setting*, they prove that the naive binary search approach is not optimal and propose an involved exponential weight algorithm, as well as a random walk binary search, for solving the problem. They prove that for a fixed $\varepsilon, \delta > 0$, the algorithm returns all arms above threshold with probability larger than $1 - \delta$, and tolerance ε, in an expected number of pulls less than a multiplicative constant *that depends on δ in a non-specified way* times $\log_2(K)/\varepsilon^2$. They prove that this is optimal up to a constant depending on δ. In the paper Ben-Or and Hassidim (2008) they refine the dependence on δ in a slightly different setting - where one has a fixed error probability. They prove that *up to terms that are negligible with respect to* $\log(K)/\varepsilon^2$, a lower bound in the expected stopping time is of order $(1-\delta)\log(K)/\varepsilon^2$. Even after a non-trivial transposition effort from their setting to ours, these results would still provide sub-optimal bounds in our setting as we consider the *expected* simple regret - and a sharper dependence in their δ would be absolutely necessary here in all regimes to get our results.

Theorem 3 provides the minimax rate of the *UTBP*. The proof can be found in Appendix E, i.e. Proposition 24, and Proposition 25.

Theorem 3 *It holds that*

$$\bar{R}_T^*(\mathcal{B}_u) \asymp \sqrt{\frac{K}{T}}.$$

The algorithm UTB *described in Section 4 attains this rate.*

Most related papers consider the problem dependent setting. However the papers Yu and Mannor (2011); Combes and Proutiere (2014a) consider the problem independent regime, in the $\mathcal{X}$-armed setting and in both cases under additional shape constraint assumptions inducing that the maximum is not too "peaky" and isolated. They prove that the minimax simple regret for the TOP-1 problem is of order $\sqrt{\log(T)/T}$.
This seems to contradict our results, to which a direct corollary is that the minimax expected regret for finding a given level set of a β-Hölder, unimodal function in $[0,1]$ is $n^{-\frac{\beta}{2\beta+1}}$, see Appendix A. This might seem unintuitive when compared to their result where the rate is much faster. But is not, as the assumption that both papers make essentially imply

that the set of arms that are ε-close to the arm with highest mean decays in a regular way, which implies that a binary search will provide good results in this case - unlike in our setting. Therefore their setting is closer in essence to the *MTBP* problem than to the *TBP* problem, as binary-search type methods work well there as highlighted in Combes and Proutiere (2014a). And interestingly, a direct corollary to Theorem 2 for `MTB` is that the minimax expected regret for finding a given level set of a β-Hölder, monotone function in $[0,1]$ is $\sqrt{\log(T)/T}$, see Appendix A, which is very much aligned with the findings in Combes and Proutiere (2014a).

Theorem 4 provides the minimax rate of the *CTBP*. The proof can be found in Appendix F, i.e. Proposition 26, and Proposition 27.

Theorem 4 *It holds that*

$$\bar{R}_T^*(\mathcal{B}_c) \asymp \sqrt{\frac{\operatorname{loglog} K}{T}}.$$

The algorithm `CTB` described in Section 4 attains this rate.

As stated in the introduction, the closest literature to our setting is that which concerns sequential estimation of a convex function and noisy convex zeroth order optimisation. Since this literature deals with the continuous case, let us first remark that a straightforward[3] corollary of Theorem 4 is that in the case where the arms are in $[0,1]$ and where f is $\beta-$Hölder for some $\beta > 0$, the minimax expected regret according to our definition (but in this continuous setting) is $\sqrt{\operatorname{loglog}(T)/T}$, see Appendix A for details.
In Simchowitz et al. (2018), the authors present the problem of estimating a convex function by constructing a net of points that is more refined in areas where the function varies more, i.e. by adapting a quadrature method to the noisy setting. Under an assumption on the modulus of continuity, that is essentially equivalent to assuming that the function is $\beta-$Hölder for some $\beta > 0$, the authors provide results in the fixed confidence setting. If one inverses their bounds to go to the fixed budget setting, their results hint toward a lower bound on estimating the convex function in l_∞ norm of order $\sqrt{\log(T)/T}$ and an upper bound of order $\log(T)/\sqrt{T}$. The fact that the logarithmic dependency is much worse in their setting than in ours highlights that the problem of estimating entirely the convex function is more difficult than the problem of estimating a single level set.
In Nemirovski and Yudin. (1983)[Chapter 9], and Wang et al. (2017); Agarwal et al. (2011); Liang et al. (2014) the authors consider continuous zeroth order noisy convex optimisation, and focus mainly on reducing the exponent for the dimension d - in this setting the minimax precision for estimating the minimum of the function is conjectured to be $d^{3/2}\text{poly}(\log(T))/\sqrt{T}$ where the $\text{poly}(\log(T))$ term is not really investigated, as the problem is already very difficult as it is. We on the other hand consider mainly $d = 1$ and aim at obtaining optimal logarithmic terms.

4. Minimax optimal algorithms

In this section we present algorithms that match minimax regret rates in Section 3 up to multiplicative constants for *TBP*, *MTBP*, *UTBP* and *CTBP*.

3. By simply discretising the space in $K^{1/\beta}$ bins and applying the method on these bins.

4.1. Unstructured case *TBP*

Given an unstructured problem $\underline{\nu} \in \mathcal{B}$ we consider the algorithm `Uniform` which samples uniformly across the arms. That is each arm in $[K]$ is sampled $\lfloor T/K \rfloor$ times. The learner then classifies each arm according to its sample mean, see Algorithm 6 in Appendix C.

Surprisingly the naive `Uniform` algorithm is optimal in the unstructured case with respect to the lower bound of Theorem 1. See the proof of Proposition 10 in Appendix C. This contrasts with the related TOP-1 bandit problem where the minimax regret rate is $\sqrt{K/T}$, see Bubeck et al. (2009); Audibert and Bubeck (2009) for hints toward this. This is not very surprising as in the TOP-1 problem we are interested in finding one arm only, namely the arm with highest mean, while in our problem we search for *all arms above threshold* and for this we pay an additional $\sqrt{\log K}$.

4.2. Monotone case *MTBP*

In this section we fix a problem $\underline{\nu} \in \mathcal{B}_m$. We also assume, in this section, without loss of generality that $\tau \in [\mu_1, \mu_K]$. Indeed, we can always add two deterministic arms 0 and $K+1$ with respective means $\mu_0 = -\infty$ and $\mu_{K+1} = +\infty$. While we assume that the distributions of the original K arms are supported in $[0,1]$ the addition of two such arms will not invalidate our proofs, see Appendix D.

We introduce the `MTB` (Monotone Thresholding Bandits) algorithm, composed of two sub-algorithms, `Explore` and `Choose`. The first, `Explore`, performs a random walk on the set of arms $[K]$ seen as a binary tree, the algorithm `Choose` then selects, among the visited arms, the one that will be chosen as the threshold for the classification. That is, we choose an arm $\hat{k}$ which leads to the estimator $\hat{Q}$, where $\hat{Q}: \ \hat{Q}[k] = -1 \ \forall k < \hat{k}, \ \hat{Q}[k] = 1 \ \forall k \geq \hat{k}$.

Binary Tree We associate to each problem $\underline{\nu} \in \mathcal{B}_m$ a binary tree. Precisely we consider a binary tree with nodes of the form $v = \{L, M, R\}$ where $\{L, M, R\}$ are indexes of arms and we note respectively $v(l) = L, v(r) = R, v(m) = M$. The tree is built recursively as follows: the root is $\texttt{root} = \{1, \lfloor (1+K)/2 \rfloor, K\}$, and for a node $v = \{L, M, R\}$ with $L, M, R \in \{1, \dots, K\}$ the left child of v is $L(v) = \{L, M_l, M\}$ and the right child is $R(v) = \{M, M_r, R\}$ with $M_l = \lfloor (L+M)/2 \rfloor$ and $M_r = \lfloor (M+R)/2 \rfloor$ as the middle index between. The leaves of the tree will be the nodes $\{v = \{L, M, R\} : R = L+1\}$. If a node v is a leaf we set $R(v) = L(v) = \emptyset$. We consider the tree up to maximum depth $H = \lfloor \log_2(K) \rfloor + 1$. We note $P\big(l(v)\big) = P\big(r(v)\big)$ the parent of the two children and let $|v|$ denote the depth of node v in the tree, with $|\texttt{root}| = 0$. We adopt the convention $P(\texttt{root}) = \texttt{root}$. In order to predict the right classification we want to find the arm whose mean is the one just above the threshold τ. Finding this arm is equivalent to inserting the threshold into the (sorted) list of means, which can be done with a binary search in the aforementioned binary tree. But in our setting we only have access to estimates of the means which can be very unreliable if the mean is close to the threshold. Because of this there is a high chance we will make a mistake on some step of the binary search. For this reason we must allow `Explore` to backtrack and this is why `Explore` performs a binary search *with corrections*. Then `Choose` selects among the visited arms the most promising one.

`Explore` **algorithm** We first define the following integers,

$$T_1 := \lceil 6 \log(K) \rceil \qquad T_2 := \left\lfloor \frac{T}{3T_1} \right\rfloor .$$

The algorithm Explore is then essentially a random walk on said binary tree moving one step per iteration for a total of T_1 steps. Let $v_1 = \texttt{root}$ and for $t < T_1$ let v_t denote the current node, the algorithm samples arms $\{v_t(k) : k \in \{l, m, r\}\}$ each T_2 times. Let the sample mean of arm $v_t(k)$ be denoted $\hat{\mu}_{k,t}$. Explore will use these estimates to decide which node to explore next. If an error is detected - i.e. the interval between left and rightmost sample mean do does not contain the threshold, then the algorithm backtracks to the parent of the current node, otherwise Explore acts as the deterministic binary search for inserting the threshold τ in the sorted list of means. More specifically, if there is an anomaly, $\tau \notin [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$, then the next node is the parent $v_{t+1} = P(v_t)$, otherwise if $\tau \in [\hat{\mu}_{l,t}, \hat{\mu}_{m,t}]$ the the next node is the left child $v_{t+1} = L(v_t)$ and if $\tau \in [\hat{\mu}_{m,t}, \hat{\mu}_{r,t}]$ the next node is the right child $v_{t+1} = R(v_t)$. If at time t, $\tau \in [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$ and the node v_t is a leaf then $v_{t+1} = v_t$. See Algorithm Explore for details.

Algorithm 1 Explore

Initialization: $v_1 = \texttt{root}$
for $t = 1 : T_1$ **do**
  sample T_2 times each arm in v_t
  if $\tau \notin [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$ **then**
    $v_{t+1} = P(v_t)$
    else if $R(v_t) = L(v_t) = \emptyset$ **then**
      $v_{t+1} = v_t$
    else if $\hat{\mu}_{m,t} \leq \tau \leq \hat{\mu}_{r,t}$ **then**
      $v_{t+1} = R(v_t)$
    else if $\hat{\mu}_{l,t} \leq \tau \leq \hat{\mu}_{m,t}$ **then**
      $v_{t+1} = L(v_t)$
  end
end

Choose algorithm Algorithm Choose takes the history of algorithm Explore, namely the sequence of empirical means $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \leq T_1}$ and visited nodes $(v_t)_{t \leq T_1}$, as the input. In addition it takes as input a parameter $\varepsilon > 0$. The action of Choose is to then identify the set of arms among those sampled whose empirical means satisfy one or more of the following:

- their empirical mean is within ε of τ,
- their empirical mean is less than τ and the empirical mean of the right hand adjacent arm is greater than τ.

Here we recognize the set of arms that may lead to a classification with simple regret smaller than ε if the estimates are correct. The algorithm Choose then orders this set by ascending arm index and returns the median, see Algorithm 2.

Algorithm 2 Choose

Input: ε, $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \leq T_1}$, $(v_t)_{t \leq T_1}$
Initialization: $S_1 = [\,]$
for $t = 1 : T_1$ **do**
  $S_{t+1} = S_t$
  if $\{\exists k \in \{l, m, r\} : |\hat{\mu}_{k,t} - \tau| \leq \varepsilon\} \vee \{k = v_t(r) = v_t(l) + 1;\ \hat{\mu}_{l,t} + \varepsilon < \tau \leq \hat{\mu}_{r,t} - \varepsilon\}$ **then**
    append $v_t(k)$ to the list S_{t+1}
  end
end
order the list S_{T_1+1} by ascending arm index
return Median(S_{T_1+1}).

Remark 5 *Note that for any time $t \leq T_1$ we append at most one arm to the list S_{t+1}. If at time t there are multiple candidates the choice is made at random.*

MTB algorithm The algorithm first runs Explore. We fix a constant $\varepsilon_0 = \sqrt{2\log(48)/T_2}$, and compute the parameter $\hat{\varepsilon}$ with the history of algorithm Explore,

$$\hat{\varepsilon} = \begin{cases} 2\varepsilon_0 & \text{if } \exists (t,k):\ k = v_t(l) = v_t(r) - 1;\ \hat{\mu}_{l,t} \leq \tau \leq \hat{\mu}_{r,t} \\ \max\left(2\varepsilon_0, \min\limits_{t \leq T_1, k \in \{l,m,r\}} |\hat{\mu}_{k,t} - \mu_{v_t(k)}|\right) & \text{else} \end{cases}.$$

Then MTB runs the algorithm Choose with parameter $\hat{\varepsilon}$. Note that $\hat{\varepsilon}$ is the smallest parameter greater than $2\varepsilon_0$ such that the list S_{T_1+1} is non empty. This choice will become clear in the proof of Theorem 2 in Appendix D. Morally it allows to select a majority of "good" arms (i.e that provide a low regret classification $\hat{Q}$) in S_{T_1+1} such that the median $\hat{k}$ is also a "good" arm, see Algorithm 3.

The MTB algorithm will achieve the minimax rate on expected simple regret given in Theorem 2, see the proof of Theorem 2, in Appendix D, for details.

Algorithm 3 MTB

run algorithm Explore
- Output: $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \leq T_1}$, $(v_t)_{t \leq T_1}$

run algorithm Choose
- Input: $\hat{\varepsilon}$, $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \leq T_1}$, $(v_t)_{t \leq T_1}$
- Output: arm index $\hat{k}$

return $(\hat{k}, \hat{Q})$: $\quad \hat{Q}_k = 2\mathbb{1}_{\{k \geq \hat{k}\}} - 1$

Remark 6 (Adaptation of MTB to a non-increasing sequence, DEC-MTB) *MTB is applied for a monotone non-decreasing sequence $(\mu_k)_k$, and it is easy to adapt it to a monotone non-increasing sequence $(\mu_k)_k$. In this case, we transform the label of arm i into $K - i$, and apply MTB to the newly labeled problem - where the mean sequence in now non-decreasing. We refer to this modification as DEC-MTB.*

4.3. Unimodal case *UTBP*

We now turn to the algorithm for the unimodal case, UTB (Unimodal Thresholding Bandits) algorithm. This algorithm is based on the algorithm MTB, and on any black-box algorithm that is minimax-optimal for TOP-1 simple regret on $\mathcal{B}$, as described in Bubeck et al. (2009). We name such an algorithm SR; it takes no parameter and returns an arm $\hat{m}$. Since SR is minimax optimal for the TOP-1 simple regret, we have on any problem $\nu \in \mathcal{B}$ with means $(\mu_k)_k$ and maximal mean μ^*, that if SR is run for T times, then

$$\mathbb{E}_\nu[\mu^* - \mu_{\hat{m}}] \leq c_{SR}\sqrt{\frac{K}{T}},$$

where $c_{SR} > 0$ is a universal constant. Note that taking MOSS from Audibert and Bubeck (2009) and modifying it so that it outputs $\hat{m}$ as being sampled at random according to the proportion of times that each arm was sampled by MOSS, is minimax-optimal algorithm for the TOP-1 problem.

The idea of UTB is to start by running SR on a fraction of the budget, and take its output $\hat{m}$. Then we run respectively MTB on $\{1, \dots, \hat{m}\}$, and DEC-MTB on $\{\hat{m}, \dots, K\}$ on a fraction of the budget. They respectively return $\hat{l}, \hat{r}$. We then use the last fraction of the budget to sample all arms in $\{\hat{l}, \hat{r}, \hat{m}, \hat{l} - 1, \hat{r} + 1\}$ and compute the respective empirical means $\hat{\mu}_k$ for k being one of these arms. If $\hat{l}, \hat{r}$ seem either close enough to the threshold, or seem above while the adjacent arm seems below, we predict $\{\hat{l}, \dots, \hat{r}\}$ as the set of arms above threshold. Otherwise we return the empty set, see Algorithm 4.

This intuitively makes sense as $\hat{m}$ is an estimator of the maximum k^* of the mean sequence and unimodality implies that $(\mu_k)_{k\leq k^*}$ is non-decreasing, and that $(\mu_k)_{k\geq k^*}$ is non-increasing. So $\hat{l}, \hat{r}$ are estimators of the points where the mean sequence crosses the threshold, respectively on the left and on the right of the estimator of the maximum. The last step - where we compute empirical means and check based on them if the outputs seem reasonable - is a checking step for making sure that the output of SR is not so close to threshold (or flawed), that the outputs of MTB and DEC-MTB are completely flawed.

Algorithm 4 UTB

Initialization: $\hat{m}$ = output of SR with budget $\lfloor T/4 \rfloor$,
$\hat{l} = \hat{k}$ output of MTB with arms $\{1, \dots, \hat{m}\}$, threshold τ, budget $\lfloor T/8 \rfloor$,
$\hat{r} = \hat{k}$ output of DEC-MTB with arms $\{\hat{m}, \dots, K\}$, threshold τ, budget $\lfloor T/8 \rfloor$,
Sample $\hat{m}, \hat{l}, \hat{r}, \hat{l}-1, \hat{r}+1$ each $\lfloor T/10 \rfloor$ times
if $(\{\hat{\mu}_{\hat{l}-1} < \tau < \hat{\mu}_{\hat{l}}\} \vee \{|\hat{\mu}_{\hat{l}} - \tau| \leq \hat{\mu}_{\hat{m}} - \tau\}) \wedge (\{\hat{\mu}_{\hat{r}} < \tau < \hat{\mu}_{\hat{r}+1}\}) \vee \{|\hat{\mu}_{\hat{r}} - \tau| \leq \hat{\mu}_{\hat{m}} - \tau\})$ **then**
  $\hat{S} = \{\hat{l}, \dots, \hat{r}\}$
  else
    $\hat{S} = \emptyset$
  end
end
return $\hat{Q}$: $\hat{Q}_k = 2\mathbb{1}_{\{k \in \hat{S}\}} - 1$

4.4. Concave case *CTBP*

In this section, we present the CTB algorithm, which is based on several applications of MTB. We first define the following *log-sets*. Consider two integers $l \leq r$ and the associated set $\{l, l+1, \dots, r\}$. We write $\mathcal{S}^{\log}_{l,r} = \{l, l+1, l+2, l+2^2, \dots, (l+2^a) \wedge \lfloor (l+r)/2 \rfloor\}$, where a is the smallest integer such that $l + 2^a \leq r \leq l + 2^{a+1}$.

Algorithm CTB proceeds in phases. At phase i an interval $\{l_i, \dots, r_i\}$ is refined from both ends by applying MTB and DEC-MTB. Algorithm CTB makes sure that with high probability, the regret of $\{l_i, \dots, r_i\}$, is bounded by $\varepsilon_i = (7/8)^i$. A very important idea of CTB is that it does not apply MTB and DEC-MTB on $\{l_i, \dots, r_i\}$ but thanks to the *concavity* only on the *log-sets associated to* $\{l_i, \dots, r_i\}$. I.e. we will apply MTB on $\mathcal{S}^{\log}_{l_i,r_i}$ and DEC-MTB on $-\mathcal{S}^{\log}_{-r_i,-l_i}$. This allows us to have much shorter phases as the two log-sets contain about $\log(r_i - l_i)$ arms, instead of $r_i - l_i$ arms.

We now describe formally CTB. The algorithm CTB consists of two sub-routines, an iterative application of MTB and then a decision rule based on the collected samples. These routines are respectively the **for loop** and **if statement** in the CTB algorithm.

Iterative application of MTB. For $\tilde{M} > 0$ and $i < M$ we set

$$\delta_i^{(\tilde{M})} = 2^{i-\tilde{M}} \qquad \varepsilon_i = \left(1 - \frac{1}{8}\right)^i \qquad \tau_i = \tau - \frac{3}{4}\varepsilon_i, \qquad T_2^{(i)}(\tilde{M}) = \left\lfloor \frac{2^{14} \log\log K}{\varepsilon_i^2} \log\left(\frac{1}{\delta_i^2}\right) \right\rfloor,$$

and let M be the largest integer such that $6\sum_{i \leq \tilde{M}} T_2^{(i)}(\tilde{M}) \leq T$. In what follows we write

$$\delta_i := \delta_i^{(M)}, \qquad T_2^{(i)} = T_2^{(i)}(M).$$

CTB proceeds in M phases and at each it updates a set of three arms $l_i \leq m_i \leq r_i$ - where m_i is at the middle between l_i and r_i. It first samples all these arms - as well as $l_i - 1, r_i + 1$ - $T_2^{(i)}$ times, and these samples are used to compute empirical means $\hat{\mu}_{p,i}$ for $p \in \{m, l, r, l-1, r+1\}$ - corresponding respectively to the arms $\{m_i, l_i, r_i, l_i - 1, r_i + 1\}$. It then runs respectively MTB on $\mathcal{S}^{\log}_{l_i,r_i}$ and DEC-MTB on $-\mathcal{S}^{\log}_{-r_i,-l_i}$, both with threshold τ_i and budget $T_2^{(i)}$. These routines output l_{i+1}, r_{i+1}, and we define m_{i+1} as the middle between these arms.

Decision rule The second sub routine of CTB is a decision rule between all l_i, r_i, for finding the right scale, based on the arms and empirical means collected in the previous routine. It takes the l_i, r_i that are as close as possible to arms m_i far from threshold, but that are close to threshold - and it outputs a set $\hat{S}$. Finally CTB classifies this set as being above threshold. Set

$$\mathcal{I}_m = \{m_i : \hat{\mu}_{m,i} \geq \tau + 2\varepsilon_i\}, \quad \text{and}$$

$$\mathcal{I}_l = \{l_i : \hat{\mu}_{l,i} \geq \tau - 2\varepsilon_i, \quad \hat{\mu}_{l-1,i} \leq \tau - \frac{\varepsilon_i}{4}\}, \text{and} \quad \mathcal{I}_r = \{r_i : \hat{\mu}_{r,i} \geq \tau - 2\varepsilon_i, \hat{\mu}_{r+1,i} \leq \tau - \frac{\varepsilon_i}{4}\}.$$

Algorithm 5 CTB

Initialization: $l_0 = 1, r_0 = K, m_0 = \lfloor \frac{l_0 + r_0}{2} \rfloor$
for $i = 1 : M$ **do**
  sample arms $l_i, l_i - 1, r_i, r_i + 1$ and m_i each $T_2^{(i)}$ times.
  l_{t+1} = output $\hat{k}$ of MTB with arms $\mathcal{S}^{\log}_{l_i,r_i}$, threshold τ_i, budget $T_2^{(i)}$
  r_{t+1} = output $\hat{k}$ of DEC-MTB with arms $-S^{\log}_{-r_i,-l_i}$, threshold τ_i, budget $T_2^{(i)}$
  $m_{i+1} = \lfloor \frac{l_{i+1} + r_{i+1}}{2} \rfloor$
end
if $\mathcal{I}_m = \emptyset$ **then**
  Set $\hat{S} = \emptyset$
  else
    Set $\hat{l} = \max\{k \in \mathcal{I}_l, k \leq \min_i \mathcal{I}_m\}$
    Set $\hat{r} = \min\{k \in \mathcal{I}_r, k \geq \max_i \mathcal{I}_m\}$
    Set $\hat{S} = \{\hat{l}, \dots, \hat{r}\}$
  end
end
return $\hat{Q}$: $\quad \hat{Q}_k = 2\mathbb{1}_{\{k \in \hat{S}\}} - 1$

Acknowledgements. The work of J. Cheshire is supported by the Deutsche Forschungsgemeinschaft (DFG) DFG - 314838170, GRK 2297 MathCoRe. The work of P. Ménard is supported by the European CHISTERA project DELTA , partially supported by the SFI Sachsen-Anhalt for the project RE-BCI and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18. The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG - 314838170, GRK 2297 MathCoRe, by the DFG GRK 2433 DAEDALUS (384950143/GRK2433), by the DFG CRC 1294 'Data Assimilation', Project A03, and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the SFI Sachsen-Anhalt for the project RE-BCI.

References

Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. 2009.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. 2010.

Michael Ben-Or and Avinatan Hassidim. The bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 221–230. IEEE, 2008.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

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

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1659–1667, 2016.

Richard Combes and Alexandre Proutiere. Unimodal bandits without smoothness. *arXiv preprint arXiv:1406.7447*, 2014a.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529, 2014b.

Ehsan Emamjomeh-Zadeh, David Kempe, and Vikrant Singhal. Deterministic and probabilistic binary search in graphs. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 519–532. ACM, 2016.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

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*, pages 3212–3220, 2012.

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

Aurélien Garivier, Pierre Ménard, Laurent Rossi, and Pierre Menard. Thresholding bandit for dose-ranging: The impact of monotonicity. *arXiv preprint arXiv:1711.04454*, 2017.

Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano's inequality for random variables. *arXiv preprint arXiv:1702.05985*, 2017.

Richard M Karp and Robert Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 881–890. Society for Industrial and Applied Mathematics, 2007.

Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. On zeroth-order stochastic convex optimization via random walks. *arXiv preprint arXiv:1402.2667*, 2014.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. *arXiv preprint arXiv:1605.08671*, 2016.

Subhojyoti Mukherjee, Naveen Kolar Purushothama, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented ucb. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2515–2521. AAAI Press, 2017.

A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. *Wiley, New York*, 1983.

Robert Nowak. The geometry of generalized binary search. *arXic preprint arXiv:0910.4397*, 2009.

Stefano Paladino, Francesco Trovo, Marcello Restelli, and Nicola Gatti. Unimodal thompson sampling for graph-structured arms. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. *arXiv preprint arXiv:1702.05186*, 2017.

Max Simchowitz, Kevin Jamieson, Jordan W Suchow, and Thomas L Griffiths. Adaptive sampling for convex regression. *arXiv preprint arXiv:1808.04523*, 2018.

Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. *arXiv preprint arXiv:1710.10551*, 2017.

Yichong Xu, Xi Chen, Aarti Singh, and Artur Dubrawski. Thresholding bandit problem with both duels and pulls. *arXic preprint arXiv:1910.06368v1*, 2019.

Jia Yuan Yu and Shie Mannor. Unimodal bandits. 2011.

Jie Zhong, Yijun Huang, and Ji Liu. Asynchronous parallel empirical variance guided algorithms for the thresholding bandit problem. *arXic preprint arXiv:1704.04567*, 2017.

Contents

Appendix A. Adaptation to the β-Hölder continuous case

In this section we explain how our results can be adapted in a very simple way to the case where the arms are not $\{1, \dots, K\}$ but the continuous set $[0, 1]$, and where the mean sequence $(\mu_k)_{k\in[0,1]}$ is now a function. We assume, on top of the fact that the distributions are supported in $[0, 1]$, that the mean function μ is β-Hölder for some constant $\beta > 0$, i.e. in the case $\beta \leq 1$ and a constant $L > 0$ such that $\forall x, y \in [0, 1]$, $|\mu_x - \mu_y| \leq L|x - y|^\beta$. In this case, straightforward corollaries of our results imply the minimax regret rates in Table A.

In oder to get these results, it is sufficient to divide $[0, 1]$ in M intervals of same size and adapt the results as usually done in the non-parametric litterature (by controlling the bias). We need to choose (i) M as $\left(\frac{T}{\log T}\right)^{\frac{1}{2\beta+1}}$ in *TBP*, (ii) M as $T^{1/\beta}$ in *MTBP*, (iii) M as $T^{\frac{1}{2\beta+1}}$ in *UTBP*, and (iii) M as $T^{1/\beta}$ in *CTBP*.

Interestingly, the rates of *MTBP* and *CTBP* *do not depend on* β - but note that β plays a role in the multiplicative constants in front of the rate, i.e. the smaller β, the larger the constant. On the other hand the rates in *TBP* and *UTBP* depend on β. Note that this is a phenomenon *specific to the* 1*-dimensional case*. Indeed, finding the level set of a monotone and of a convex function in dimension d is typically done at a much slower rate, depending on β and d.

Our results	Unstructured *TBP*	Monotone *MTBP*	Unimodal *UTBP*	Convex *CTBP*
K-arms	$\sqrt{\frac{K \log K}{T}}$	$\sqrt{\frac{\log K \vee 1}{T}}$	$\sqrt{\frac{K}{T}}$	$\sqrt{\frac{\log\log K \vee 1}{T}}$
β-Hölder	$\left(\frac{\log T}{T}\right)^{\frac{\beta}{2\beta+1}}$	$\sqrt{\frac{\log T \vee 1}{T}}$	$\left(\frac{1}{T}\right)^{\frac{\beta}{2\beta+1}}$	$\sqrt{\frac{\log\log T \vee 1}{T}}$

Table 2: Order of the minimax expected regret for the thresholding bandit problem, in the case of all four structural assumptions on the means of the arms considered in this paper. All results are given up to universal multiplicative constants. The first line concerns the $K-$armed setting of the main paper, and the second line concerns the $\mathcal{X}$-armed setting where the set of arms is $[0, 1]$ and where the function is β-Hölder (on top of the shape constraints).

Appendix B. Extension to σ^2-sub-Gaussian for *TBP* and *MTBP*

While in the main text for simplicity we only consider distributions bounded on the $[0, 1]$ interval all proofs relating to the *TBP* and *MTBP* given in the appendix will extend to the sub Gaussian case. The lower bound for the *CTBP* will also extend to the sub Gaussian case. That is we redefine the setting as follows: the learner is presented with a K-armed bandit problem $\underline{\nu} = \{\nu_1, \dots, \nu_K\}$, where ν_k is the unknown distribution of arm k. Let $\sigma^2 > 0$, all arms are assumed to be σ^2-sub-Gaussian as described in the following definition, we write μ_k for the mean of arm k.

Definition 7 (σ^2-sub-Gaussian) *A distribution ν of mean μ is said to be σ^2-sub-Gaussian if for all $t \in \mathbb{R}$ we have,*

$$\mathbb{E}_{X\sim\nu}\left[e^{t(X-\mu)}\right] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right).$$

In particular the Gaussian distributions with variance smaller than σ^2 and the distributions with absolute values bounded by σ are σ^2-sub-Gaussian.

The only adaptation that has to be made to accomodate this case in the MTB algorithm is to define

$$\varepsilon_0 = \sqrt{\frac{2\sigma^2 \log(48)}{T_2}}.$$

Appendix C. Proof of Theorem 1

In the proof of all results in this section, we assume that the more general sub-Gaussian assumption described in Section B is satisfied - and not necessarily that the distributions of all arms are bounded on the $[0,1]$ interval. We explain in the proof how the lower bound can be straightforwardly adapted to distributions supported in $[0,1]$.

We denote the Kullback-Leibler divergence between two Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ (with the usual conventions) by

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

Algorithm 6 Uniform

for $k = 1 : K$ **do**
  Sample arm k a total of $\lfloor \frac{T}{K} \rfloor$ times.
  Compute $\hat{\mu}_k$ the sample mean of arm k.
end
return

$$\hat{Q}: \quad \hat{Q}_k = \begin{cases} -1 & \text{if } \hat{\mu}_k < \tau \\ 1 & \text{if } \hat{\mu}_k \geq \tau \end{cases}$$

During this section we will prove Theorem 1 by first demonstrating a lower bound on expected regret across $\mathcal{B}$ and then showing that the Uniform algorithm achieves said lower bound. We first prove the following proposition to establish a lower bound.

Proposition 8 *For any $T \geq 1$ and any strategy π, there exists a unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, such that*

$$\bar{R}_T^{\pi,\underline{\nu}} \geq \frac{3}{4}\sqrt{\frac{\sigma^2 \max\left(2, \log(K)\right)K}{8T}}.$$

Proof Without loss of generality we can assume that $\tau = 0$. Fix some positive real number $0 < \varepsilon < 1$. And consider the family of Gaussian bandit problems indexed by an vertex of the unite hyper-cube of dimension K, id est $Q \in \{-1,1\}^K$

$$\underline{\nu}^Q = \left(\mathcal{N}(Q_1\varepsilon, \sigma^2), \ldots, \mathcal{N}(Q_K\varepsilon, \sigma^2)\right),$$

and note that if we wish to consider distributions supported in $[0,1]$ we can consider instead $\tau = 1/2$ and

$$\underline{\nu}^Q = \big(\mathcal{B}(1/2 + Q_1\varepsilon), \dots, \mathcal{B}(1/2 + Q_K\varepsilon)\big)\,,$$

up to minor adaptations of the constants, and to considering $\tau = 1/2$. Note that all these bandit problems belong to the set of unstructured bandit problems, $\underline{\nu}^Q \in \mathcal{B}$. The regret in the bandit problem $\underline{\nu}^Q$ of the strategy π can be rewritten as follows

$$\begin{aligned}\bar{R}_T^{\underline{\nu}^Q,\pi} &= \varepsilon\, \mathbb{E}_Q \max_k \mathbb{1}_{\{\hat{Q}_k \neq Q_k\}} \\ &= \varepsilon(1 - \mathbb{E}_Q\, \mathbb{1}_{\{\hat{Q}=Q\}})\,,\end{aligned}$$

where we denote by $\mathbb{E}_Q$ the expectation under the bandit problem $\underline{\nu}^Q$. We will provide a minimax lower bound on the regret by using the classic Fano inequality. We first lower bound the minimax expected regret in the problem $\underline{\nu}^Q$ by the Bayesian regret with a uniform distribution over the bandit problems $\underline{\nu}^Q$,

$$\max_Q \bar{R}_T^{\underline{\nu}^Q,\pi} \geq \varepsilon \left(1 - \frac{1}{2^K} \sum_Q \mathbb{E}_Q\, \mathbb{1}_{\{\hat{Q}=Q\}}\right)\,. \tag{1}$$

Let Q^k be the transformation of Q that flip the sign of the coordinate k,

$$Q_a^k = \begin{cases} Q_a \text{ If } a \neq k, \\ -Q_a \text{ If } a = k\,. \end{cases}$$

Thanks to the contraction and the convexity of the relative entropy, see Gerchinovitz et al. (2017), we have

$$\mathrm{kl}\left(\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\, \mathbb{1}_{\{\hat{Q}=Q^k\}}, \underbrace{\frac{1}{K}\sum_{k=1}^K \mathbb{E}_Q\, \mathbb{1}_{\{\hat{Q}=Q^k\}}}_{\leq 1/K}\right) \leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\big[N_k(T)\big]\frac{\varepsilon^2}{2\sigma^2}\,,$$

where $N_k(T) = \sum_{t=1}^T \mathbb{1}_{\{k_t=k\}}$ denotes the number of times in total arm k is sampled. Then using a refined Pinsker inequality (see Gerchinovitz et al. (2017)) $\mathrm{kl}(x,y) \geq (x-y)^2 \max\big(2, \log(1/y)\big)$, we obtain

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\, \mathbb{1}_{\{\hat{Q}=Q^k\}} \leq \frac{1}{K} + \sqrt{\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\big[N_k(T)\big]\frac{\varepsilon^2}{2\sigma^2 \max\big(2, \log(K)\big)}}\,. \tag{2}$$

Therefore thanks to the concavity of the square root, we can average over all the bandit problems $\underline{\nu}^Q$

$$\frac{1}{2^K}\sum_Q \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\, \mathbb{1}_{\{\hat{Q}=Q^k\}} \leq \frac{1}{K} + \sqrt{\frac{1}{2^K}\sum_Q \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\big[N_k(T)\big]\frac{\varepsilon^2}{2\sigma^2 \max\big(2, \log(K)\big)}}\,.$$

Now it remains to remark that by symmetry

$$\sum_{Q}\sum_{k=1}^{K}\mathbb{E}_{Q^k}\,\mathbb{1}_{\{\hat{Q}=Q^k\}} = \sum_{Q'}\sum_{k=1}^{K}\mathbb{E}_{Q'}\,\mathbb{1}_{\{\hat{Q}=Q'\}} = K\sum_{Q}\mathbb{E}_{Q}\,\mathbb{1}_{\{\hat{Q}=Q\}}\,,$$
$$\sum_{Q}\sum_{k=1}^{K}\mathbb{E}_{Q^k}\big[N_k(T)\big] = \sum_{Q'}\sum_{k=1}^{K}\mathbb{E}_{Q'}\big[N_k(T)\big] = \sum_{Q}T\,.$$

Hence from (2) we get

$$\frac{1}{2^K}\sum_{Q}\mathbb{E}_{Q}\,\mathbb{1}_{\{\hat{Q}=Q\}} \leq \frac{1}{K} + \sqrt{\frac{T\varepsilon^2}{2K\sigma^2\max\big(2,\log(K)\big)}}\,,$$

and then from (1) we obtain

$$\max_{Q}\bar{R}_T^{\nu^Q,\pi} \geq \varepsilon\left(\frac{1}{2} - \sqrt{\frac{T\varepsilon^2}{2K\sigma^2\max\big(2,\log(K)\big)}}\right)\,.$$

Choosing $\varepsilon = \sqrt{K\sigma^2\max\big(2,\log(K)\big)/(8T)}$ allows us to conclude. ■

We next prove the following proposition to establish a upper bound on the regret of the Uniform algorithm with high probability,

Proposition 9 *For any unstructured bandit problem $\underline{\nu}\in\mathcal{B}$, any $T\geq K$, any $0<\delta<1$, Uniform satisfies*

$$\mathbb{P}_{\underline{\nu}}\left(R_T^{\mathit{Uniform},\underline{\nu}} \geq \sqrt{\frac{4\sigma^2K}{T}\log\left(\frac{2K}{\delta}\right)}\right) \leq \delta\,.$$

Proof

During the execution of the Uniform algorithm $\forall k\in\{1,...,K\}$ arm k is sampled $\lfloor T/K\rfloor$ times with sample mean $\hat{\mu}_k$. Let $\delta>0$ and consider the event,

$$\xi := \left\{\forall\, k\leq K,\ |\hat{\mu}_k-\mu_k| \leq \sqrt{\frac{4\sigma^2K}{T}\log\left(\frac{2K}{\delta}\right)}\right\}\,.$$

Thanks to the Hoeffding inequality and an union bound this event occurs with probability greater than $1-\delta$. As under the event ξ,

$$\hat{\mu}_k \in \left[\hat{\mu}_k - \sqrt{\frac{4\sigma^2K}{T}\log\left(\frac{2K}{\delta}\right)}, \hat{\mu}_k + \sqrt{\frac{4\sigma^2K}{T}\log\left(\frac{2K}{\delta}\right)}\right]\,,$$

and the returning classification is

$$\hat{Q}: \quad \hat{Q}_k = \begin{cases} -1 & \text{if } \hat{\mu}_k<\tau \\ 1 & \text{if } \hat{\mu}_k\geq\tau \end{cases}\,,$$

we have with probability at least $1-\delta$

$$R_T = \max_{\{k\in[K]:\ \hat{Q}_k \neq Q_k\}} \Delta_k \leq \sqrt{\frac{4\sigma^2 K}{T}\log\left(\frac{2K}{\delta}\right)}\,.$$

■

We are now able to demonstrate a bound on the expected regret of the Uniform algorithm.

Proposition 10 *For any unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, and any $T \geq K$, Uniform satisfies*

$$\bar{R}_T^{\mathit{Uniform},\underline{\nu}} \leq 7\sqrt{\frac{\sigma^2 \log(2K)K}{T}}\,.$$

Proof By application of Theorem 9, for $\varepsilon > 0$ we have,

$$\mathbb{P}\left(R_T \geq \varepsilon\right) \leq 2K \exp\left(-\varepsilon^2 \frac{T}{4\sigma^2 K}\right).$$

Hence for $\varepsilon_0 = \sqrt{4\sigma^2 \log(2K)K/T}$ integrating these probabilities we obtain an upper bound on the expected simple regret

$$\begin{aligned}
\bar{R}_T &\leq \sqrt{2}\varepsilon_0 + \int_{\sqrt{2}\varepsilon_0}^{+\infty} \exp\left(-(\varepsilon^2 - \varepsilon_0^2)\frac{T}{2\sigma^2 K}\right) \mathrm{d}\varepsilon \\
&\leq \sqrt{2}\varepsilon_0 + \int_{0}^{+\infty} \exp\left(-\varepsilon^2 \frac{T}{8\sigma^2 K}\right) \mathrm{d}\varepsilon \\
&= \sqrt{\frac{8\sigma^2 \log(2K)K}{T}} + \sqrt{\frac{2\pi\sigma^2 K}{T}} \\
&\leq 7\sqrt{\frac{\sigma^2 \log(2K)K}{T}}\,.
\end{aligned}$$

■

Setting $\sigma = 1$, Theorem 1 follows directly from a combination of Propositions 10 and 8.

Appendix D. Proof of Theorem 2

In the proofs of all results in this section, we assume that the more general sub-Gaussian assumption described in Section B is satisfied - and not necessarily that the distributions of all arms are bounded on the $[0,1]$ interval. In this case, we remind that we redefine ε_0 as in Section B. Also, we explain in the proof of the lower bound how it is possible to straightforwardly adapt the proof to the case where the distributions are supported in $[0,1]$.

During this section we will prove Theorem 2 by first demonstrating a lower bound upon expected regret in the *MTBP* setting, Proposition 11. We will then go on to provide an upper bound on the regret of the MTB with high probability, Proposition 12 which will be used to finally prove Corollary 13 which provides a optimal bound for the MTB in expected regret. Setting $\sigma = 1$ Theorem 2 will then follow directly from Proposition 11 and Corollary 13.

Proposition 11 *For any $T \geq 1$ and any strategy π, there exists a structured bandit problem $\underline{\nu} \in \mathcal{B}_m$, such that*

$$\bar{R}_T^{\pi,\underline{\nu}} \geq \frac{1}{8}\sqrt{\frac{\sigma^2 \max\big(2, \log(K)\big)}{8T}}\,.$$

Proof We will proceed as in the proof of Proposition 8. Fix some positive real number $0 < \varepsilon < 1$. Without loss of generality we can assume that $\tau = \varepsilon/2$. And consider the family of Gaussian bandit problems $\underline{\nu}^k$ indexed by $k \in \{0, \dots, K\}$, such that for all $k \in \{0, \dots, K\}$, $l \in [K]$,

$$\nu_l^k = \begin{cases} \mathcal{N}(0, \sigma^2) & \text{if } k < l \\ \mathcal{N}(\varepsilon, \sigma^2) & \text{else} \end{cases}\,.$$

Note that if we wish to consider distributions supported in $[0, 1]$ we can consider instead $\tau = 1/2 + \varepsilon/2$ and

$$\nu_l^k = \begin{cases} \mathcal{B}(1/2) & \text{if } k < l \\ \mathcal{B}(1/2 + \varepsilon) & \text{else} \end{cases}\,.$$

up to minor adaptations of the constants, and to considering $\tau = 1/2$.

Note that all these bandit problems belong to the set of structured bandit problems, $\underline{\nu}^k \in \mathcal{B}$. Following the same steps as in the proof of Proposition 8 one can lower bound the maximum of the expected regrets over all the bandit problems introduced above,

$$\max_{k \in [K]} \bar{R}_T^{\underline{\nu}^k,\pi} \geq \frac{\varepsilon}{2}\left(1 - \frac{1}{K}\sum_{k=1}^K \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q} = Q^k\}}\right)\,,$$

where we denote by $\mathbb{E}^k$ the expectation and by Q^k the true classification in the problem $\underline{\nu}^k$. Thanks to the contraction and the convexity of the relative entropy we have

$$\begin{aligned} \mathrm{kl}\left(\frac{1}{K}\sum_{k=1}^K \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q} = Q^k\}}, \underbrace{\frac{1}{K}\sum_{k=1}^K \mathbb{E}_0\, \mathbb{1}_{\{\hat{Q} = Q^k\}}}_{\leq 1/K}\right) &\leq \frac{1}{K}\sum_{k=1}^K \sum_{l=k}^K \mathbb{E}_k\big[N_l(T)\big]\frac{\varepsilon^2}{2\sigma^2} \\ &\leq \frac{T\varepsilon^2}{2\sigma^2}\,. \end{aligned}$$

Then using a refined Pinsker inequality $\mathrm{kl}(x, y) \geq (x - y)^2 \max\big(2, \log(1/y)\big)$, we obtain

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q} = Q^k\}} \leq \frac{1}{K} + \sqrt{\frac{T\varepsilon^2}{2\sigma^2 \max\big(2, \log(K)\big)}}\,.$$

Hence combining the last three inequalities we get

$$\max_{k \in [K]} \bar{R}_T^{\underline{\nu}^k,\pi} \geq \frac{\varepsilon}{2}\left(\frac{1}{2} - \sqrt{\frac{T\varepsilon^2}{2\sigma^2 \max\big(2, \log(K)\big)}}\right)\,.$$

Choosing $\varepsilon = \sqrt{\sigma^2 \max\big(2, \log(K)\big)/(8T)}$ allows us to conclude. ■

We next prove the following to proposition to establish an upper bound on the simple regret of the MTB algorithm with high probability and then prove Corollary 13 to establish an upper bound on the expected regret of the MTB algorithm. For Proposition 12 we consider a more general set of problems, given $\varepsilon > 0$, define,

$$\mathcal{B}_m^{*,\varepsilon} := \{\mathcal{B} : (\min(|\mu_i - \tau|, \varepsilon)\,\mathrm{sign}(\mu_i - \tau))_{k \le K} \text{ is an increasing sequence}\}\,.$$

Note that for all $\varepsilon > 0$, $\mathcal{B}_m \subset \mathcal{B}_m^{*,\varepsilon}$, hence all results will hold also in the unaltered monotone setting.

Proposition 12 *For any $\varepsilon > \varepsilon_0$ and any problem $\underline{\nu} \in \mathcal{B}_m^{*,\varepsilon}$, and any $T > 6\log(K)$, the MTB Algorithm will achieve the following bound on simple regret,*

$$\mathbb{P}_{\underline{\nu}}(R_T^{MTB,\underline{\nu}} \ge \varepsilon) \le \min\left(\exp\left(-\frac{3\log(K)}{4}\right),\ 72\log(K)\exp\left(-\frac{T\varepsilon^2}{216\sigma^2\log(K)}\right)\right).$$

Corollary 13 *For any problem $\underline{\nu} \in \mathcal{B}_m$ and any $T \ge 12\log(K)$, the MTB algorithm will achieve the following bound on expected regret,*

$$\bar{R}_T^{MTB,\underline{\nu}} \le 80\sqrt{\frac{\sigma^2\log(K)}{T}}\,.$$

The proof of Proposition 12 and Corollary 13 is structured in several steps which we will first summarise. For a level $\varepsilon > 0$ we define a set of "good nodes" containing "ε-good arms", those which when outputted will achieve the bound $R_T < 2\varepsilon$. In Proposition 16 we prove these nodes form a "consecutive tree", see Definition 15. At time t we say we have a "favourable event" if all sampled empirical means are within ε of the true mean, In this case we say the algorithm makes a "good decision", see (10). In Lemma 19 we prove that on every good decision we move towards the set of good arms or remain within them. Lemma 20 then shows that provided we make enough good decisions the number of good arms in S is large. We can then bound the probability of making a high proportion of good decisions, see Lemma 21, to give an upper bound on regret. This in combination with a second upper bound, Lemma 23, will give our result.

Step 0: Definitions and Lemmas We will use the following definitions.

Definition 14 *We define the subtree $ST(v)$ of a node v recursively as follows: $v \in ST(v)$ and*

$$\forall\, q \in ST(v),\ L(q), R(q) \in ST(v)\,.$$

Definition 15 *A consecutive tree U with root u_{root} is a set of nodes such that $u_{root} \in U$ and*

$$\forall v \in U : v \ne u_{root},\ P(v) \in U.$$

with the additional condition,

$$\texttt{root} \in U \Rightarrow u_{\texttt{root}} = \texttt{root}$$

where `root` *is the root of the entire binary tree.*

We define Z^ε, the set of ε-good nodes, as the union of the two sets

$$Z_1^\varepsilon := \{v : \exists k \in \{l, m, r\} : |\mu_{v(k)} - \tau| \leq \varepsilon\} \,, \tag{3}$$

$$Z_2^\varepsilon := \{v : v(r) = v(l) + 1;\ \mu_{v(l)} \leq \tau \leq \mu_{v(r)}\} \backslash Z_1^\varepsilon \,, \tag{4}$$

that is

$$Z^\varepsilon := Z_1^\varepsilon \ \cup \ Z_2^\varepsilon \,.$$

It is important to note that

$$Z_2^\varepsilon \neq \emptyset \Rightarrow |Z^\varepsilon| = 1 \,. \tag{5}$$

Proposition 16 *Z^ε is a consecutive tree with root $z^\varepsilon_{\text{root}}$ the unique element $v \in Z^\varepsilon$, such that $P(v) \notin Z^\varepsilon$.*

Proof If $Z_2^\varepsilon \neq \emptyset$ by (5) we have $|Z^\varepsilon| = 1$ and the proposition is trivially verified. Hence we assume $Z^\varepsilon = Z_1^\varepsilon$. Consider $v \in Z^\varepsilon$, such that $P(v) \notin Z^\varepsilon$, there is at least one such node. We first prove that v is unique. As $v \in Z^\varepsilon = Z_1^\varepsilon$ we know that

$$\exists k \in \{l, m, r\} : \left|\mu_{v(k)} - \tau\right| \leq \varepsilon \,. \tag{6}$$

Now since $v(l), v(r) \in P(v)$ and $P(v) \notin Z^\varepsilon$, it follows that, thanks to (6),

$$\forall k \in \{l, r\} : \left|\mu_{v(k)} - \tau\right| > \varepsilon \qquad |\mu_{v(m)} - \tau| \leq \varepsilon \,.$$

For node $q \neq v$ satisfying the same properties, assume that $v(m) < q(m)$ without loss of generality. With this assumption we have,

$$v(r) \leq v(m) \leq q(l) \leq q(m) \,,$$

however, as the sequence $(\min(|\mu_i - \tau|, \varepsilon) \operatorname{sign}(\mu_i - \tau))_{k \leq K}$ is increasing we must have $|\mu_{v(r)} - \tau| \leq \varepsilon$ and $|\mu_{q(l)} - \tau| \leq \varepsilon$, a contradiction. Hence $v = q$, and thus v is unique which implies $\forall q \in Z^\varepsilon :\ q \neq v,\ P(q) \in Z^\varepsilon$. ■

At time t we define w_t^ε as the node of maximum depth whose subtree contains both v_t and an "ε-good node" belonging to Z^ε. Formally, for $t \leq T_1$,

$$w_t^\varepsilon := \underset{\{w : ST(w) \cap Z^\varepsilon \neq \emptyset \ \&\ v_t \in ST(w)\}}{\arg\max} |w| \,.$$

Lemma 17 *The node w_t^ε is unique and*

$$w_t^\varepsilon = \underset{\{w: ST(w)\cap Z^\varepsilon \neq \emptyset \ \&\ v_t \in ST(w)\}}{\arg\min} \left(|v_t| - |w| + (|z^\varepsilon_{root}| - |w|)^+\right) \ . \tag{7}$$

Proof

At time t consider, a node q_t^ε which also satisfies 7, giving

$$|v_t| - |w_t^\varepsilon| + (|z^\varepsilon_{\texttt{root}}| - |w_t^\varepsilon|)^+ = |v_t| - |q_t^\varepsilon| + (|z^\varepsilon_{\texttt{root}}| - |q_t^\varepsilon|)^+ \ .$$

As $v_t \in ST(w_t^\varepsilon)$ and $v_t \in ST(q_t^\varepsilon)$ we can assume without loss of generality $q_t^\varepsilon \in ST(w_t^\varepsilon)$ with $|q_t^\varepsilon| \geq |w_t^\varepsilon|$. Thus,

$$|v_t| - |q_t^\varepsilon| \leq |v_t| - |w_t^\varepsilon| \ ,$$

and therefore,

$$(|z^\varepsilon_{root}| - |q_t^\varepsilon|)^+ \geq (|z^\varepsilon_{root}| - |w_t^\varepsilon|)^+ \ ,$$

which implies, $|q_t^\varepsilon| \geq |w_t^\varepsilon|$, therefore $|q_t^\varepsilon| = |w_t^\varepsilon|$ and as $q_t^\varepsilon \in ST(w_t^\varepsilon)$, we have $q_t^\varepsilon = w_t^\varepsilon$. ■

For $t \leq T_1$ we define D_t^ε as the distance from v_t to Z^ε, it is taken as the length of the path running from v_t up to w_t^ε and then down to an ε-good node in Z^ε. Formally, we have

$$D_t^\varepsilon := |v_t| - |w_t^\varepsilon| + (|z^\varepsilon_{\texttt{root}}| - |w_t^\varepsilon|)^+ .$$

Note the following properties of D_t^ε and w_t^ε,

$$ST(v_t) \cap Z^\varepsilon \neq \emptyset \Rightarrow v_t = w_t^\varepsilon \ , \tag{8}$$

$$D_t = 0 \Rightarrow v_t = w_t^\varepsilon \text{ And } w_t^\varepsilon, v_t \in Z^\varepsilon \ . \tag{9}$$

Let S_t^ε denote the list produced by an execution of algorithm `Choose` with parameter $\varepsilon \geq \varepsilon_0$. We define W_ε as the set of ε-good arms

$$W_\varepsilon := \left\{k \in [K] : \Delta_k \leq 3\varepsilon \text{ OR } \mu_{k-1} < \tau < \mu_k\right\},$$

and at time t the counter G_t^ε, tracking the number of 3ε-good arms in $S_t^{2\varepsilon}$,

$$G_t^\varepsilon := \left|\left\{k \in S_t^{2\varepsilon} : \ k \in W_{3\varepsilon}\right\}\right| . \tag{10}$$

Note that if $\hat{k}$ belongs to this set then we suffer at most a regret of 3ε. We define also the favorable event where the estimates the means are close to the true ones for all the arms in v_t,

$$\xi_t^\varepsilon := \left\{\forall k \in \{l, m, r\}, \left|\hat{\mu}_{k,t} - \mu_{v_t(k)}\right| \leq \varepsilon\right\} . \tag{11}$$

Step 2: Actions of the algorithm on all iterations After any execution of algorithm `Explore` and subsequent execution of algorithm `Choose` with parameter ε, note the following,

- for $t \leq T_1$, v_t and v_{t+1} are separated by at most one edge, i.e.

$$v_{t+1} \in \{L(v_t), R(v_t), P(v_t)\}\,, \tag{12}$$

- for $t \leq T_1$,

$$|S_t^{2\varepsilon}| \leq |S_{t+1}^{2\varepsilon}| \leq |S_t^{2\varepsilon}| + 1\,. \tag{13}$$

Lemma 18 *On execution of algorithm `Explore` and algorithm `Choose` with parameter $\varepsilon > 0$ for all $t \leq T_1$ we have the following,*

$$D_{t+1}^{\varepsilon} \leq D_t^{\varepsilon} + 1, \tag{14}$$
$$G_{t+1}^{\varepsilon} \geq G_t^{\varepsilon}\,. \tag{15}$$

Proof As the algorithm moves at most 1 step per iteration, see (12), for $t \leq T_1$, it holds

$$||v_t| - |w_t^{\varepsilon}|| \geq ||v_{t+1}| - |w_t^{\varepsilon}|| - 1\,.$$

Noting that,

$$\begin{aligned} D_t^{\varepsilon} &= ||v_t| - |w_t^{\varepsilon}|| + (|z_{\mathtt{root}}^{\varepsilon}| - |w_t^{\varepsilon}|)^+ \\ &\geq ||v_{t+1}| - |w_t^{\varepsilon}|| + (|z_{\mathtt{root}}^{\varepsilon}| - |w_t^{\varepsilon}|)^+ - 1 \\ &\geq \left||v_{t+1}| - \left|w_{t+1}^{\varepsilon}\right|\right| + \left(|z_{\mathtt{root}}^{\varepsilon}| - |w_{t+1}^{\varepsilon}|\right)^+ - 1 \\ &= D_{t+1}^{\varepsilon} - 1\,, \end{aligned}$$

where the third line comes from the definition of w_{t+1}^{ε}, see (7), we obtain $D_{t+1}^{\varepsilon} \leq D_t^{\varepsilon} + 1$. By (13) we have, for $t \leq T_1$,

$$|S_t^{2\varepsilon}| \leq |S_{t+1}^{2\varepsilon}| \leq |S_t^{2\varepsilon}| + 1\,,$$

hence $G_{t+1}^{\varepsilon} \geq G_t^{\varepsilon}$. ■

Step 3: Actions of the algorithm on ξ_t^{ε}

Lemma 19 *On execution of algorithm `Explore` and algorithm `Choose` with parameter $\varepsilon > 0$ for all $t \leq T_1$, on ξ_t^{ε}, we have the following,*

$$D_{t+1}^{\varepsilon} \leq \max(D_t^{\varepsilon} - 1, 0)\,, \tag{16}$$
$$G_{t+1}^{\varepsilon} \geq G_t^{\varepsilon} + \mathbb{1}_{\{D_t^{\varepsilon}=0\}}\,. \tag{17}$$

Proof We first prove (17). Note that if the arm $v_t(k)$ is added in $S_{t+1}^{2\varepsilon}$ then either $|\hat{\mu}_{k,t} - \tau| \leq 2\varepsilon$ or $v_t(k) = v_t(r) = v_t(l) + 1$ and $\hat{\mu}_{l,t} + \varepsilon \leq \tau \leq \hat{\mu}_{r,t}$. Thus, on ξ_t^{ε}, we obtain in the first case $\Delta_{v_t(k)} \leq 3\varepsilon$ and in the second case

$$v_t(k) = v_t(l) = v_t(r) - 1 \text{ and } \mu_{v_t(l)} + \varepsilon \leq \tau \leq \mu_{v_t(r)} - \varepsilon\,.$$

In both case we have $v_t(k) \in W^{3\varepsilon}$, hence $G^\varepsilon_{t+1} \geq G^\varepsilon_t + 1$. It remains to prove that, when $D_t = 0$, an arm is effectively added in $S^{2\varepsilon}_{t+1}$. If $D^\varepsilon_t = 0$ then we know $v_t \in Z^\varepsilon$. If $v_t \in Z^\varepsilon_1$ then under ξ^ε_t there exists $k \in \{l, m, r\}$ such that $|\hat{\mu}_{k,t} - \tau| \leq 2\varepsilon$. Otherwise we know that

$$v_t(l) = v_t(r) - 1 \text{ and } \mu_{v_t(l)} + \varepsilon \leq \tau \leq \mu_{v_t(r)} - \varepsilon \,,$$

which implies on ξ^ε_t that

$$\hat{\mu}_{l,t} \leq \tau \leq \hat{\mu}_{r,t} \,.$$

In both case an arm is added to $S^{2\varepsilon}_{t+1}$.

Now we prove (16). Note that on the favorable event ξ^ε_t, we have $\forall k \in \{l, m, r\}$,

$$\mu_{v_t(k)} \geq \tau + \varepsilon \Rightarrow \hat{\mu}_{k,t} \geq \tau \,, \tag{18}$$

$$\mu_{v_t(k)} \leq \tau - \varepsilon \Rightarrow \hat{\mu}_{k,t} \leq \tau \,. \tag{19}$$

We consider the following three cases:

- If $\tau \notin \left[\mu_{v_t(l)} + \varepsilon,\, \mu_{v_t(r)} - \varepsilon\right]$. From (18) and (19), under ξ^ε_t, we get $\tau \notin [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$, and therefore $v_{t+1} = P(v_t)$. Since in this case we are getting closer to the set of ε-good nodes by going up in the tree we know that $w^\varepsilon_t = w^\varepsilon_{t+1}$. Thus thanks to Lemma 17, under ξ^ε_t,

$$D^\varepsilon_{t+1} = |v_{t+1}| - \left|w^\varepsilon_{t+1}\right| + \left(|z^\varepsilon_{\texttt{root}}| - |w^\varepsilon_{t+1}|\right)^+ = |v_t| - 1 - |w^\varepsilon_t| + (|z^\varepsilon_{\texttt{root}}| - |w^\varepsilon_t|)^+ = D^\varepsilon_t - 1 \,.$$

- If $\tau \in \left[\mu_{v_t(l)} + \varepsilon,\, \mu_{v_t(r)} - \varepsilon\right]$ and $v_t \notin Z^\varepsilon$. Note that in this case v_t can not be a leaf and we just need to go down in the subtree of v_t to find an ε-good node, id est $w_t = v_t$. Since $v_t \notin Z^\varepsilon$, without loss of generality, we can assume for example $\mu_{v_t(m)} > \tau + \varepsilon$. From (18) and (19), under ξ^ε_t, we then have $\tau \in [\hat{\mu}_{l,t},\, \hat{\mu}_{r,t}]$ and $\hat{\mu}_{m,t} \geq \tau$. Hence algorithm `Explore` goes to the correct subtree, $v_{t+1} = L(v_t)$. In particular we also have for this node

$$\tau \in \left[\mu_{v_{t+1}(l)} - \varepsilon,\, \mu_{v_t(m)} + \varepsilon\right] \,,$$

  therefore it holds again $w_{t+1} = v_{t+1}$. Thus combining the previous remarks we obtain thanks to Lemma 17, under ξ^ε_t,

$$D^\varepsilon_{t+1} = (|w_{t+1}| - |z^\varepsilon_{\texttt{root}}|)^+ = (|w_t| - |z^\varepsilon_{\texttt{root}}|)^+ - 1 = D^\varepsilon_t - 1 \,.$$

- If $\tau \in \left[\mu_{v_t(l)} + \varepsilon,\, \mu_{v_t(r)} - \varepsilon\right]$ and $v_t \in Z^\varepsilon$. We distinguish two cases: Z^ε_2 is empty or not. In both cases we will show that, under ξ^ε_t, $v_{t+1} \in Z^\varepsilon$ and thus

$$D^\varepsilon_{t+1} = D^\varepsilon_t = 0 \,.$$

  Hence it remains to consider these two cases:
  - If $Z^\varepsilon_2 \neq \emptyset$. Via the definition of Z^ε_2, see (4), and the fact $Z^\varepsilon_1 = \emptyset$, v_t is a leaf with $\mu_{v_t(r)} \leq \tau - \varepsilon$ and $\mu_{v_t(l)} \geq \tau + \varepsilon$. Hence from (18) and (19) we have $\hat{\mu}_{l,t} \leq \tau \leq \hat{\mu}_{r,t}$. Therefore by the action of algorithm `Explore` we will stay in the same node $v_{t+1} = v_t$.

– Else $Z_2^\varepsilon = \emptyset$. If $\mu_{v_t(m)} \in [\tau - \varepsilon, \tau + \varepsilon]$, we have $R(v_t), L(v_t), P(v_t) \in Z^\varepsilon$ hence trivially $v_{t+1} \in Z^\varepsilon$. Else we have $\mu_{v_t(m)} \notin [\tau - \varepsilon, \tau + \varepsilon]$. Without loss of generality we assume $\mu_{v_t(m)} > \tau + \varepsilon$. This implies that $\mu_{v_t(r)} > \tau + \varepsilon$ and since $v_t \in Z^\varepsilon = Z_1^\varepsilon$ it holds $\mu_{v_t(l)} \in [\tau - \varepsilon, \tau + \varepsilon]$. Thus, under ξ_t^c we then get as previously $\tau \in [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$ and $\hat{\mu}_{m,t} \geq \tau$. Therefore by the action of algorithm `Explore` we will go to the left child $v_{t+1} = L(v_t) \in Z^\varepsilon$.

■

Step 4: Lower bound on $G^\varepsilon_{T_1+1}$ We denote by $\bar{\xi}_t^\varepsilon$ the complement of ξ_t^ε.

Lemma 20 *For any execution of algorithm `Explore` and subsequent execution of `Choose` with parameter $\varepsilon \geq \varepsilon_0$,*

$$G^\varepsilon_{T_1+1} \geq \frac{3}{4} T_1 - 2 \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t^\varepsilon} .$$

Proof Combining (16) and (14) from Lemma 18 and Lemma 19 respectively we have

$$\begin{aligned} D^\varepsilon_{t+1} &\leq D^\varepsilon_t + \mathbb{1}_{\bar{\xi}_t^\varepsilon} - \mathbb{1}_{\xi_t^\varepsilon} \mathbb{1}_{\{D_t^\varepsilon > 0\}} \\ &= D^\varepsilon_t + 2\mathbb{1}_{\bar{\xi}_t^\varepsilon} - 1 + \mathbb{1}_{\xi_t^\varepsilon} \mathbb{1}_{\{D_t^\varepsilon = 0\}} . \end{aligned}$$

Using this inequality with (17) we obtain

$$\begin{aligned} G^\varepsilon_{T_1+1} &= \sum_{t=1}^{T_1} G^\varepsilon_{t+1} - G^\varepsilon_t \\ &\geq \sum_{t=1}^{T_1} \mathbb{1}_{\xi_t^\varepsilon} \mathbb{1}_{\{D_t^\varepsilon = 0\}} \\ &\geq \sum_{t=1}^{T_1} \left(D^\varepsilon_{t+1} - D^\varepsilon_t - 2\mathbb{1}_{\xi_t^\varepsilon} + 1 \right) \\ &\geq T_1 - D_1^\varepsilon - 2 \sum_{t=1}^{T_1} \mathbb{1}_{\xi_{t,\varepsilon}} \\ &\geq \frac{3}{4} T_1 - 2 \sum_{t=1}^{T_1} \mathbb{1}_{\xi_{t,\varepsilon}} , \end{aligned}$$

where we used in the last inequality the fact that $D_1 \leq \log_2(K)$ and that $\log_2(K) \leq T_1/4$ by definition of T_1 . ■

Step 5: First high probability bound on the regret

Lemma 21 *For all $\varepsilon \geq \varepsilon_0$, following the execution of algorithm `MTB`,*

$$\mathbb{P}(R_T > 3\varepsilon) \leq e^{-3\log(K)/4} . \tag{20}$$

Before proving Lemma 21 we need to show that the number of times a favorable events $\xi_t^{\varepsilon_0}$ occurs is not to small with high probability. Precisely in the following lemma we upper bound the probability of the event

$$\xi^{\varepsilon_0} = \left\{ \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t^{\varepsilon_0}} \leq \frac{T_1}{8} \right\} .$$

Lemma 22 *For any execution of algorithm* `Explore` *and subsequent execution of* `Choose` *with parameter* ε_0,

$$\mathbb{P}(\bar{\xi}^{\varepsilon_0}) \leq e^{-3\log(K)/4} .$$

Proof Let $\mathcal{F}_t$ be the information available at and including time t. Thanks to the Hoeffding inequality and the choice of T_2, we have for all $k \in \{l, m, r\}$,

$$\mathbb{P}\left(\left|\hat{\mu}_{k,t} - \mu_{v_t(k)}\right| \geq \varepsilon_0 | \mathcal{F}_{t-1}\right) \leq 2\exp\left(-\frac{T_2\varepsilon_0^2}{2\sigma^2}\right) \leq \frac{1}{24} ,$$

hence by a union bound $\mathbb{P}(\bar{\xi}_t^{\varepsilon_0} | \mathcal{F}_{t-1}) \leq 1/8$. Then the Azuma-Hoeffding inequality applied to the martingale

$$\sum_{t=1}^{T_1} \left[\mathbb{1}_{\bar{\xi}_t^{\varepsilon_0}} - \mathbb{P}(\bar{\xi}_t^{\varepsilon_0} | \mathcal{F}_{t-1}) \right] ,$$

with respect to the filtration $(\mathcal{F}_t)_{t \leq T_1}$ allows us to conclude

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \left[\mathbb{1}_{\bar{\xi}_t^{\varepsilon_0}} - \mathbb{P}(\bar{\xi}_t^{\varepsilon_0} | \mathcal{F}_{t-1}) \right] \geq \frac{T_1}{4} \right) \leq e^{-2T_1/16} \leq e^{-3\log(K)/4} , \tag{21}$$

where we used that $T_1 = \lceil 6\log(K) \rceil$. ■

We are now ready to prove Lemma 21.

Proof [Proof of Lemma 21] We first prove it for $\varepsilon = \varepsilon_0$. Thanks to Lemma 20 on the event ξ^{ε_0} we have

$$G_{T_1+1}^{\varepsilon_0} \geq \frac{3}{4}T_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t^{\varepsilon_0}} \geq \frac{T_1}{2} .$$

But thanks to the choice of $\hat{\varepsilon} \geq 2\varepsilon_0$ we know that

$$S_{T_1+1}^{2\varepsilon_0} \subset S_{T_1+1}^{\hat{\varepsilon}} .$$

Thus there is more than the half of the arms of $S_{T_1+1}^{\hat{\varepsilon}}$ in $W_{3\varepsilon_0}$, since this list is at most of size T_1. In particular this implies that $\hat{k} = \text{Median}(S_{T_1+1}^{\hat{\varepsilon}}) \in W_{3\varepsilon_0}$. Indeed $W_{3\varepsilon_0}$ is a segment in $[K]$, see (6). Therefore, on the event ξ^{ε_0} we have

$$R_T \leq 3\varepsilon_0.$$

Lemma 22 allows us to conclude, for $\varepsilon \geq \varepsilon_0$,

$$\mathbb{P}(R_T > 3\varepsilon) \leq \mathbb{P}(R_T > 3\varepsilon_0) \leq e^{-3\log(K)/4} .$$

■

Step 6: Second high probability bound on the regret

Lemma 23 *For all $\varepsilon \geq \varepsilon_0$, following the execution of algorithm MTB,*

$$\mathbb{P}(R_T > 3\varepsilon) \leq 72 \log(K) \exp\left(-\frac{T\varepsilon^2}{36\sigma^2 \log(K)}\right) . \tag{22}$$

Proof We consider the event where all the favorable events ξ_t^ε occur,

$$\xi_a^\varepsilon := \bigcap_{t=1}^{T_1} \xi_t^\varepsilon .$$

On this event ξ_a^ε thanks to Lemma 20 we have

$$\begin{aligned} G_{T_1+1}^\varepsilon &\geq \frac{3}{4} T_1 - 2 \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t^\varepsilon} \\ &= \frac{3}{4} T_1 , \end{aligned}$$

hence $S_{T_1+1}^{2\varepsilon} \neq \emptyset$ is not empty. Furthermore following the same arguments of the beginning of the proof of Lemma 19 all arms in the list $S_{T_1+1}^{2\varepsilon} \neq \emptyset$ are also in $W_{3\varepsilon}$. Then noting that by construction

$$\hat{\varepsilon} = \inf_{\varepsilon' \geq 2\varepsilon_0 :\ S_{T_1+1}^{\varepsilon'} \neq \emptyset} \varepsilon' ,$$

we get $\hat{\varepsilon} \leq 2\varepsilon$ therefore $S_{T_1+1}^{\hat{\varepsilon}} \subset S_{T_1+1}^{2\varepsilon}$. Thanks to the remarks above we know that $\hat{k} \in W_{3\varepsilon}$ thus on ξ_a^ε,

$$R_T \leq 3\varepsilon .$$

The Hoeffding inequality in combination with a union bound allows us to conclude,

$$\mathbb{P}(\bar{\xi}_a^\varepsilon) \leq \sum_{t=1}^{T_1} \mathbb{E}\left[\mathbb{P}(\bar{\xi}_t^\varepsilon | \mathcal{F}_{t-1})\right] \leq 72 \log(K) \exp\left(-\frac{T_2 \varepsilon^2}{2\sigma^2}\right) \tag{23}$$

$$\leq 72 \log(K) \exp\left(-\frac{T\varepsilon^2}{36\sigma^2 \log(K)}\right) . \tag{24}$$

■

Conclusion The proof of Proposition 12 is straightforward combining Lemma 21 and Lemma 23. Thus we obtain for all $\varepsilon \geq 3\varepsilon_0$,

$$\mathbb{P}(R_T \geq \varepsilon) \leq \min\left(\exp\left(-\frac{3\log(K)}{4}\right), 72 \log(K) \exp\left(-\frac{T\varepsilon^2}{324\sigma^2 \log(K)}\right)\right) .$$

We can integrate the high probability upper bound obtained in Proposition 12 to prove Corollary 13.

Proof [Proof of Corollary 13.] Thanks to Proposition 12, for $\varepsilon_1 = \log(72\log(K))\sqrt{324\sigma^2\log(K)/T}$, we have

$$\begin{aligned}\mathbb{E}[R_T] &\leq \varepsilon_0 + (\varepsilon_1 - \varepsilon_0)e^{-3\log(K)/4} + \int_{\varepsilon=\varepsilon_1}^{+\infty} 72\log(K)\exp\left(-\frac{T\varepsilon^2}{324\sigma^2\log(K)}\right) \\ &\leq \sqrt{\frac{36\sigma^2\log(48)\log(K)}{T}} + \left(\underbrace{\frac{\log(72\log(K))}{K^{3/4}}}_{\leq 3} + \frac{\sqrt{\pi}}{2}\right)\sqrt{\frac{324\sigma^2\log(K)}{T}} \\ &\leq 80\sqrt{\frac{\sigma^2\log(K)}{T}}\,.\end{aligned}$$

■

Setting $\sigma = 1$ Theorem 2 follows directly from Proposition 11 and Corollary 13.

Appendix E. Proof of Theorem 3

To prove Theorem 3 we first demonstrate, in Proposition 24, a lower bound on the expected regret of any strategy on the *UTBP*. We will then show, with Proposition 25, that the UTB achieves said lower bound. The proof of Theorem 3 will then follow directly. For all proofs during this section we make the assumption that arms are distributed as σ^2-sub-Gaussian with $\sigma = 1$. Also, we explain in the proof of the lower bound how it is possible to straightforwardly adapt the proof to the case where the distributions are supported in $[0, 1]$.

Proposition 24 *For any $T \geq 1$ and any strategy π, there exists an unimodal bandit problem $\underline{\nu} \in \mathcal{B}_u$, such that*

$$\bar{R}_T^{\pi,\underline{\nu}} \geq \frac{1}{8}\sqrt{\frac{K}{T}}\,.$$

Proof We will proceed as in the proof of Proposition 8. Fix some positive real number $0 < \varepsilon < 1$. Without loss of generality we can assume that $\tau = \varepsilon/2$. And consider the family of Gaussian bandit problems $\underline{\nu}^k$ indexed by $k \in \{0, \dots, K\}$, such that for all $k \in \{0, \dots, K\}$, $l \in [K]$,

$$\nu_l^k = \begin{cases}\mathcal{N}(\varepsilon, \sigma^2) & \text{if } k = l \\ \mathcal{N}(0, \sigma^2) & \text{else}\end{cases}\,.$$

Note that if we wish to consider distributions in $[0, 1]$ we can consider instead $\tau = 1/2 + \varepsilon/2$

$$\nu_l^k = \begin{cases}\mathcal{B}(1/2 + \varepsilon) & \text{if } k = l \\ \mathcal{B}(1/2) & \text{else}\end{cases}\,,$$

up to minor alterations of the constants, and to considering $\tau = 1/2$.

Note that all these bandit problems belong to the set of unimodal bandit problems, $\underline{\nu}^k \in \mathcal{B}_u$. Following the same steps as in the proof of Proposition 8 one can lower bound the maximum of the expected regrets over all the bandit problems introduced above,

$$\max_{k\in[K]} \bar{R}_T^{\underline{\nu}^k,\pi} \geq \frac{\varepsilon}{2}\left(1 - \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q}=Q^k\}}\right),$$

where we denote by $\mathbb{E}^k$ the expectation and by Q^k the true classification in the problem $\underline{\nu}^k$. Thanks to the contraction and the convexity of the relative entropy we have

$$\mathrm{kl}\left(\underbrace{\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_0\, \mathbb{1}_{\{\hat{Q}=Q^k\}}}_{\leq 1/K}, \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q}=Q^k\}}\right) \leq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_0\big[N_k(T)\big]\frac{\varepsilon^2}{2\sigma^2}$$
$$\leq \frac{T\varepsilon^2}{2K\sigma^2}.$$

Then using the Pinsker inequality $\mathrm{kl}(x,y) \geq 2(x-y)^2$, we obtain

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_k\, \mathbb{1}_{\{\hat{Q}=Q^k\}} \leq \frac{1}{K} + \sqrt{\frac{T\varepsilon^2}{4\sigma^2 K}}.$$

Hence combining the last three inequalities we get

$$\max_{k\in[K]} \bar{R}_T^{\underline{\nu}^k,\pi} \geq \frac{\varepsilon}{2}\left(\frac{1}{2} - \sqrt{\frac{T\varepsilon^2}{4\sigma^2 K}}\right).$$

Choosing $\varepsilon = \sqrt{4\sigma^2 K/T}$ allows us to conclude. ■

Proposition 25 *There exists a universal constant $c_{\mathrm{uni}} > 0$ such that for any unimodal bandit problem $\underline{\nu} \in \mathcal{B}_u$, UTB satisfies*

$$\bar{R}_T^{\mathtt{UTB},\underline{\nu}} \leq c_{\mathrm{uni}}\sqrt{\frac{K}{n}}.$$

Proof

Step 1: Definitions Write

$$\hat{\Delta} = \mu^* - \mu_{\hat{m}},$$

and

$$\hat{\varepsilon} = |\hat{\mu}_{\hat{l}} - \mu_{\hat{l}}| \vee |\hat{\mu}_{\hat{r}} - \mu_{\hat{r}}| \vee |\hat{\mu}_{\hat{m}} - \mu_{\hat{m}}| \vee |\hat{\mu}_{\hat{r}+1} - \mu_{\hat{r}+1}| \vee |\hat{\mu}_{\hat{l}-1} - \mu_{\hat{l}-1}|.$$

and we write $R^{(l)}$ for the regret of MTB on $\{1,\dots,\hat{m}\}$ when played by algorithm UTB, and $R^{(r)}$ for the regret of DEC-MTB on $\{\hat{m},\dots,K\}$ when played by algorithm UTB. Let us also write $R_T = R_T^{\mathtt{UTB},\nu}$ for the regret associated to the outputted set $\hat{S}$.

$$\mathcal{E}^{(l)} = \{|\hat{\mu}_{\hat{l}} - \tau| \leq \hat{\mu}_{\hat{m}} - \tau\} \cup \{\hat{\mu}_{\hat{l}-1} \leq \tau \leq \hat{\mu}_{\hat{l}}\},$$

and define similarly $\mathcal{E}^{(r)}$ replacing l by r. Define

$$\mathcal{E} = \{\mathcal{E}^{(l)} \cap \mathcal{E}^{(r)}\}.$$

Step 2: Bound on the regret on the events Assume without loss of generality that $R^{(l)} \geq R^{(r)}$. By definition of the algorithm this implies under this condition that

$$R_T = R^{(l)}\mathbb{1}_{\{\mathcal{E}\}} + (\mu^* - \tau)_+\mathbb{1}_{\{\mathcal{E}^C\}},$$

which implies directly

$$R_T \leq R^{(l)}\mathbb{1}_{\{\mathcal{E}\}} + (\mu_{\hat{m}} - \tau)_+\mathbf{1}\{\mathcal{E}^C\} + \hat{\Delta}. \tag{25}$$

Note that

$$\mathcal{E} \subset \{|\mu_{\hat{l}} - \tau| \leq \mu_{\hat{m}} - \tau + 2\hat{\varepsilon}\} \cup \{\mu_{\hat{l}-1} - \hat{\varepsilon} \leq \tau \leq \mu_{\hat{l}} + \hat{\varepsilon}\}.$$

And so since $R^{(l)} \leq |\mu_{\hat{l}} - \tau|$, we have

$$R^{(l)}\mathbb{1}_{\{\mathcal{E}\}} \leq R^{(l)} \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon}. \tag{26}$$

Note also that on $\mathcal{E}^C$ and under our condition $R^{(l)} \geq R^{(r)}$, we have that

$$\mathcal{E}^C \cap \{R^{(l)} \geq R^{(r)}\} \subset \{|\mu_{\hat{l}} - \tau| \geq \mu_{\hat{m}} - \tau - 2\hat{\varepsilon}\}.$$

And on $\mathcal{E}^C \cap \{R^{(l)} \geq R^{(r)}\}$, we have that $R^{(l)} \geq (\mu_{\hat{l}} - \tau)_+ - 2\hat{\varepsilon}$, which leads to under our assumption $R^{(l)} \geq R^{(r)}$

$$(\mu_{\hat{m}} - \tau)_+\mathbb{1}_{\{\mathcal{E}^C\}} \leq R^{(l)} \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon}. \tag{27}$$

So we have combining (26) and (27) all cases in (25) that if $R^{(l)} \geq R^{(r)}$

$$R_T \leq (R^{(l)}) \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon} + \hat{\Delta}.$$

Considering similarly the case $R^{(r)} \geq R^{(l)}$ gives

$$R_T \leq (R^{(l)} \vee R^{(r)}) \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon} + \hat{\Delta}.$$

Step 3: Integration of the regret Consider $\varepsilon_0 = 4c_{SR}\sqrt{\frac{K}{n}}$. Consider the event where $(\mu_{\hat{m}} - \tau)_+ = \tilde{\varepsilon} \geq \varepsilon_0$. On this event, and since the sequence of arms's means is unimodal, MTB satisfies the assumptions of Corollary 31 for $\tilde{\varepsilon}$ and a set of arms $\{1, \ldots, \hat{m}\}$, and integrating over the tail probability between ε_0 and $\tilde{\varepsilon}$ - conditional to we know that there exists an absolute constant $C > 0$ such that

$$\mathbb{E}[R^{(l)} \wedge \tilde{\varepsilon} | (\mu_{\hat{m}} - \tau)_+ = \tilde{\varepsilon}] \leq C\sqrt{\frac{\log K + 1}{n}}.$$

Similarly

$$\mathbb{E}[R^{(r)} \wedge \tilde{\varepsilon} | (\mu_{\hat{m}} - \tau)_+ = \tilde{\varepsilon}] \leq C\sqrt{\frac{\log K + 1}{n}}.$$

And so

$$\mathbb{E}\Big[(R^{(l)} \vee R^{(r)}) \wedge (\mu_{\hat{m}} - \tau)_+\Big] \leq C\sqrt{\frac{\log K + 1}{n}}.$$

combining this with the sub-Gaussian properties of the means which give that

$$\mathbb{E}\hat{\varepsilon} \leq c\sqrt{\frac{1}{T}},$$

where $c > 0$ is some absolute constant, and with the minimax optimality of SR which gives

$$\mathbb{E}\hat{\Delta} \leq 4c_{SR}\sqrt{\frac{K}{T}},$$

this provides the result.

■

Appendix F. Proof of Theorem 4

For the proof of Proposition 27 we make the assumption that the distribution of all arms is bounded on the $[0, 1]$ interval. In the case of the lower bound we consider σ^2-sub-Gaussian distributions. Also, we explain in the proof of the lower bound how it is possible to straightforwardly adapt the proof to the case where the distributions are supported in $[0, 1]$.

Proposition 26 *For any $T \geq 1$, $K \geq 3$ and any strategy π, there exists a structured bandit problem $\underline{\nu} \in \mathcal{B}_c$, such that*

$$\bar{R}_T^{\pi,\underline{\nu}} \geq \frac{3}{4}\sqrt{\frac{\sigma^2 \max\big(2, \mathrm{loglog}(K)\big)}{8T}}\,.$$

Proof We will proceed as in the previous proofs but with a different alternative set. Without loss of generality we can assume that $\tau = 0$. Fix some positive real number $0 < \varepsilon < 1$. And consider the family of Gaussian bandit problems $\underline{\nu}^l$ indexed by $l \in \{0, \ldots, L := \lfloor \log_2(K) \rfloor\}$ defined by $\underline{\nu}^l = \mathcal{N}(\mu^l, 1)$ with

$$\mu_k^l = \begin{cases} \frac{k}{k_l}\varepsilon & \text{if } k \leq k_l := 2^l \\ \varepsilon & \text{else}\,. \end{cases}.$$

Note that if we want to consider distributions supported in $[0, 1]$ we can consider $\underline{\nu}^l = \mathcal{B}(1/2 + \mu^l)$ and $\tau = 1/2$ instead of the Gaussian distributions, up to minor adaptations of the constants, and to considering $\tau = 1/2$.

Note that all these bandit problems belong to the set of convex bandit problems, $\underline{\nu}^k \in \mathcal{B}_c$. We will lower bound the maximum of the expected regrets over all the bandit problems introduced above,

$$\max_{l\in[L]} \bar{R}_T^{\underline{\nu}^l,\pi} = \max_{l\in[L]} \mathbb{E}_l\left[\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\hat{Q}_k \neq Q_k^l\}}\right],$$

where we denote by $\mathbb{E}^l$ the expectation and by Q^l the true classification in the problem ν^l. In particular we have $Q^l = [-1, \dots, -1, 1, \dots, 1]$ where the first one is at position k_l.

Let $\hat{k} := \max\{j : \ \hat{Q}_j = -1\}$ be the estimated k_l and $\widetilde{Q}$ an other possible answer exploiting the structure of the problems

$$\widetilde{Q}_k = \begin{cases} -1 & \text{if } j \leq \hat{k} \\ 1 & \text{else} \end{cases} .$$

Then we have for all $l \in [L]$ it holds

$$R_T^l := \max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\hat{Q}_k = Q_k^l\}} \geq \max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widetilde{Q}_k = Q_k^l\}} ,$$

indeed if $\hat{k} \geq k_l$ then $R_T^l = \varepsilon$ and since for all k the gaps are smaller than $\varepsilon \geq \Delta_k^l$ the inequality is trivially true. Else $\hat{k} < k_l$, then for all $k \leq \hat{k}$ we have $Q_k = -1$ thus $\mathbb{1}_{\hat{Q}_k \neq Q_k^l} \geq \mathbb{1}_{\widetilde{Q}_k \neq Q_k^l}$ and since $\hat{Q}_k = \widetilde{Q}_k$ for all $k \geq \hat{k}$ the inequality is verified. Now let $\hat{l} = \arg\min_{l\in[L]:\ k_l > \hat{k}} l$ be an estimate for the index of the problem. Then we have

$$\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widetilde{Q}_k = Q_k^l\}} \geq \max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^l\}} .$$

Indeed using the same arguments as above, if $\hat{k} \geq k_l$ then $\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widetilde{Q}_k = Q_k^l\}} = \varepsilon$. Else $\hat{k} < k_l$, then for all k, $\mathbb{1}_{\{\widetilde{Q}_k = Q_k^l\}} \geq \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^l\}}$. Finally, we prove that

$$\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^l\}} \geq \frac{\varepsilon}{2} \mathbb{1}_{\{\hat{l} \neq l\}} ,$$

indeed, if $\hat{l} > l$ then $\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^l\}} = \varepsilon$ and if $\hat{l} < l$ we have using $k_{\hat{l}+1} = 2k_{\hat{l}}$,

$$\begin{aligned} \max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^l\}} &\geq \max_{k\in[K]} \Delta_k^{\hat{l}+1} \mathbb{1}_{\{Q_k^{\hat{l}} = Q_k^{\hat{l}+1}\}} \\ &\geq \varepsilon - \frac{k_{\hat{l}}}{k_{\hat{l}+1}} \varepsilon = \frac{\varepsilon}{2} . \end{aligned}$$

Combining all the previous inequalities we obtain

$$\max_{l\in[L]} \bar{R}_T^{\nu^l,\pi} \geq \frac{\varepsilon}{2} \max_{l\in[L]} \mathbb{E}_l \, \mathbb{1}_{\{\hat{l} \neq l\}} \geq \frac{\varepsilon}{2} \left(1 - \frac{1}{L} \sum_{l\in[L]} \mathbb{E}_l \, \mathbb{1}_{\{\hat{l} = l\}} \right) .$$

We can conclude as previously. Thanks to the contraction and the convexity of the relative entropy we have

$$\begin{aligned} \mathrm{kl}\left(\frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_l \, \mathbb{1}_{\{\hat{l}=l\}}, \underbrace{\frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_0 \, \mathbb{1}_{\{\hat{l}=l\}}}_{\leq 1/L} \right) &\leq \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{E}_l\big[N_k(T)\big] \frac{\varepsilon^2}{2\sigma^2} \\ &\leq \frac{T\varepsilon^2}{2\sigma^2} . \end{aligned}$$

Then using a refined Pinsker inequality $\mathrm{kl}(x,y) \geq (x-y)^2 \max\big(2, \log(1/y)\big)$, we obtain

$$\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_l \, \mathbb{1}_{\{\hat{l}=Q^l\}} \leq \frac{1}{L} + \sqrt{\frac{T\varepsilon^2}{2\sigma^2 \max\big(2, \log(L)\big)}} \, .$$

Hence combining the last three inequalities we get

$$\max_{l\in[L]} \bar{R}_T^{\nu^l,\pi} \geq \varepsilon \left(\frac{1}{2} - \sqrt{\frac{T\varepsilon^2}{2\sigma^2 \max\big(2, \log(L)\big)}} \right) .$$

Choosing $\varepsilon = \sqrt{\sigma^2 \max\big(2, \log(L)\big)/(8T)}$ allows us to conclude. ■

Proposition 27 *There exists a universal constant $c_{\mathrm{conv}} > 0$ such that for any convex bandit problem $\underline{\nu} \in \mathcal{B}_c$, CTB satisfies*

$$\bar{R}_T^{CTB,\underline{\nu}} \leq c_{\mathrm{conv}} \sqrt{\frac{\log\log K \vee 1}{T}} \, .$$

Before going on to prove Proposition 27 we first show the following.

Lemma 28 *Consider $1 \leq p \leq q \leq K$, $\tilde{\varepsilon} > 0, \tilde{\tau} \in \mathbb{R}$. Consider any $1 \leq p < q \leq K$, such that,*

$$\mu_{\lfloor \frac{p+q}{2} \rfloor} \geq \tilde{\tau} + \frac{1}{8}\tilde{\varepsilon} \, . \tag{28}$$

Then

$$\left(\min(|\mu_k - \tilde{\tau}|, \frac{1}{8}\tilde{\varepsilon}) \operatorname{sign}(\mu_k + \tilde{\tau}) \right)_k ,$$

is monotonically increasing on $[p : \lfloor \frac{p+q}{2} \rfloor]$ and monotonically decreasing on $[\lfloor \frac{p+q}{2} \rfloor : q]$.

Proof We just prove that the sequence is monotonically increasing on $[p : \lfloor \frac{p+q}{2} \rfloor]$, the other case is proven similarly.

Since $(\mu_k)_{k\leq K}$ is concave, we know that there exists $k^* \in \{1, \dots, K\}$ such that $(\mu_k)_{k\leq k^*}$ is increasing and $(\mu_k)_{k\geq k^*}$ is decreasing.

- If $k^* \in [p, \lfloor \frac{p+q}{2} \rfloor]$, and since (28) holds, we have that $\forall k \in [k^*, \lfloor \frac{p+q}{2} \rfloor]$, $\mu_k - \tilde{\tau} \geq \tilde{\varepsilon}/8$. This implies the result.
- If $k^* \notin [p : \lfloor \frac{p+q}{2} \rfloor]$, we have either (i) that μ_k is increasing on the interval which implies the result or (ii) that μ_k is decreasing on the interval. In case (ii), we know by (28) that $\forall k \in [p, \lfloor \frac{p+q}{2} \rfloor]$, $\mu_k - \tilde{\tau} \geq \tilde{\varepsilon}/8$. This implies the result.

■

Lemma 29 *Let $\tilde{\varepsilon} > 0, \tilde{\tau} \in \mathbb{R}$. For any $1 \leq p \leq q \leq K$, such that,*

$$\mu_p \wedge \mu_q \geq \tilde{\tau} - \tilde{\varepsilon} \, ,$$
$$\mu_{\lfloor \frac{p+q}{2} \rfloor} \leq \tilde{\tau} - \frac{5}{8}\tilde{\varepsilon} \, ,$$

we have that, $\forall k \in \{p, \dots, q\}$ that $\mu_k \leq \tilde{\tau} - \varepsilon$.

Proof
We assume $\exists k \in \{p, \dots, q\}$ such that $\mu_k > \tau - \frac{1}{8}\tilde{\varepsilon}$ and aim to prove by contradiction. Without loss of generality assume $k < \frac{p+q}{2}$, in combination with the assumptions of Lemma 29 we have $(\mu_k - \mu_{\lfloor \frac{p+q}{2} \rfloor}) > \frac{1}{2}\tilde{\varepsilon}$. However, via the convex property $(\mu_k - \mu_{\lfloor \frac{p+q}{2} \rfloor}) \leq (\mu_{\rfloor \frac{p+q}{2} \lfloor} - \mu_q)$, a contradiction as it implies with the forelast equation that $\mu_q < \tau - \frac{1}{8}\tilde{\varepsilon}$. ■

We now define the event,

$$\begin{aligned}
\xi_i &:= \left(\xi_i^{(L)} \cap \xi_i^{(R)}\right) \cup \xi_i^{(A)} \\
&:= \left(\left\{ \mu_{l_i} \geq \tau - \varepsilon_i \, , \forall k < l_i : \mu_k \leq \tau - \frac{1}{2}\varepsilon_i \right\} \right. \\
&\left. \cap \left\{ \mu_{r_i} \geq \tau - \varepsilon_i \, , \forall k > r_i : \mu_k \leq \tau - \frac{1}{2}\varepsilon_i \right\} \right) \\
&\cup \left\{ \forall k \leq K, \mu_k \leq \tau - \frac{1}{8}\varepsilon_i \right\}.
\end{aligned}$$

Consider the event

$$\mathcal{E}_i = \{\mu_{m_i} \geq \tau_i + \frac{1}{8}\varepsilon_i\}. \tag{29}$$

Proposition 30 *Let $i \leq M$ and set*

$$\delta_i' = \min\left(\exp\left(-\frac{3\log\log(K)}{4} \right), 72\log\log(K)\exp\left(-\frac{T_2^{(i)}\varepsilon_i^2}{216 \times 64 \log\log(K)} \right) \right).$$

Let l'_{i+1} be the largest arm smaller than l_{i+1} in $\mathcal{S}^{\log}_{l_i, r_i}$. It holds that

$$\mathbb{P}\left(|\mu_{l_{i+1}} - \tau_i| \leq \varepsilon_i/8 \;\; OR \;\; \mu_{l'_{i+1}} + \varepsilon_i/8 < \tau_i < \mu_{l_{i+1}} - \varepsilon_i/8 \middle| \mathcal{E}_i \right) \geq 1 - \delta_i'.$$

Also for r'_{i+1} be the smallest arm smaller than r_i in $-\mathcal{S}^{\log}_{-r_i, l_i}$.

$$\mathbb{P}\left(|\mu_{r_{i+1}} - \tau_i| \leq \varepsilon_i/8 \;\; OR \;\; \mu_{r'_{i+1}} + \varepsilon_i/8 < \tau_i < \mu_{r_{i+1}} - \varepsilon_i/8 \middle| \mathcal{E}_i \right) \geq 1 - \delta_i'.$$

Proof A straightforward corollary of Proposition 12 is as follows.

Corollary 31 *Consider $\varepsilon \geq \sqrt{\frac{2\log(48)6\log(K)}{T}}$ and a problem $\nu \in \mathcal{B}(1, K)$ such that $(\min(|\mu_k - \tau|, \varepsilon)\text{sign}(\mu_k - \tau) + \tau)_k$ is increasing with k. Then the MTB Algorithm will allow us to identify and arm $\hat{k}$ such that,*

$$|\mu_{\hat{k}} - \tau| \leq \varepsilon \text{ OR } \mu_{\hat{k}-1} + \varepsilon \leq \tau \leq \mu_{\hat{k}-1} - \varepsilon$$

with probability greater than,

$$1 - \min\left(\exp\left(-\frac{3\log(K)}{4}\right), 72\log(K)\exp\left(-\frac{T\varepsilon^2}{216\log(K)}\right)\right).$$

The result of the proposition follows by applying this corollary and noting that

- in any case, $|\mathcal{S}^{\log}_{l_i,r_i}| \leq \log K$ so that we apply MTB on a problem that has less than $\log\log K$ arms,
- that on $\mathcal{E}_i$, we have that $(\min(|\mu_k - \tau_i|, \varepsilon_i/8)\text{sign}(\mu_k - \tau_i))_{k\in[l_i,m_i]}$ is increasing (respectively, $(\min(|\mu_k - \tau_i|, \varepsilon_i/8)\text{sign}(\mu_k - \tau_i))_{k\in[m_i,r_i]}$ is decreasing) - see Lemma 28.
- Moreover $\varepsilon_i \geq \varepsilon_M \geq \sqrt{\frac{2\log(48)6\log(K)}{T}}$. And so since $\mathcal{S}^{\log}_{l_i,r_i} \subset [l_i, m_i]$ (resp. $-\mathcal{S}^{\log}_{-r_i,-l_i} \subset [m_i, r_i]$), the conditions of Corollary 31 are satisfied, for the set $\mathcal{S}^{\log}_{l_i,r_i}$ of arms.

Therefore we can apply Corollary 31 to show that when running MTB $(\mathcal{S}^{\log}_{l_i,r_i}, \tau_i, T_2^{(i)})$ we are able to identify an arm $\hat{k}$ such that setting $l_{i+1} = \hat{k}$ satisfies our result with probability greater than $1 - \delta_i'$.

■

Proposition 32 *We have that for $i \leq M$*

$$\mathbb{P}\left(\xi_{i+1}^{(L)} \middle| \xi_i \cap \mathcal{E}_i\right) \geq 1 - \delta_i',$$

and

$$\mathbb{P}\left(\xi_{i+1}^{(R)} \middle| \xi_i \cap \mathcal{E}_i\right) \geq 1 - \delta_i'.$$

Proof

We prove this proposition only for $\xi_{i+1}^{(L)}$ as the proof for $\xi_{i+1}^{(R)}$ is similar. Consider the high probability event of Proposition 30, where we just have two possibilities for the mean of l_{i+1} which we summarize below.

Case 1 Consider the case where MTB outputs l_{i+1} such that,

$$\mu_{l'_{i+1}} + \varepsilon_i/8 < \tau_i < \mu_{l_{i+1}} - \varepsilon_i/8\,, \tag{30}$$

where l'_{i+1} is defined in Proposition 30. Since $(\mu_k)_{k<K}$ is concave and since by definition of the concave grid $\mathcal{S}^{\log}_{l_i,r_i}$ we have that for $l'_{i+1} \neq l_i$,

$$\mu_{l'_{i+1}} - \mu_{l_{i+1}} \geq \frac{\varepsilon_i}{4}\,.$$

However this would imply

$$\mu_{l_i} < \tau_i - \frac{\varepsilon_i}{8} - \frac{\varepsilon_i}{4} < \tau - \varepsilon_i\,,$$

contradicting ξ_i, hence $l_i = l'_{i+1}$ and therefore via choice of l'_{i+1}, $l_i + 1 = l_{i+1}$. Therefore as $\mu_{k<K}$ is concave,

$$\forall k < l_{i+1}, \mu_k \leq \mu_{l_{i+1}}\,.$$

The property $\mu_{l_{i+1}} \geq \tau - \varepsilon_{i+1}$ follows directly from (4), we have $\xi_{i+1}^{(L)}$

Case 2 Consider the case where MTB outputs l_{i+1} such that,

$$|\mu_{l_{i+1}} - \tau_i| \leq \varepsilon_i/8.$$

From Lemma 28 we have that the sequence $(\mu_k)_{k<K}$ is increasing on $[\tau_i - \frac{1}{8}\varepsilon_i, \tau_i + \frac{1}{8}\varepsilon_i]$ Therefore $\forall k < l_{i+1}, \mu_k \leq \mu_{l_{i+1}}$. Hence ξ^L_{i+1} holds.

And so we have as desired that

$$\xi^{(L)}_{i+1} \cap \xi_i \cap \mathcal{E}_i \subset \{|\mu_{l_{i+1}} - \tau_i| \leq \varepsilon_i/8 \;\; OR \;\; \mu_{l'_{i+1}} + \varepsilon_i/8 < \tau_i < \mu_{l_{i+1}} - \varepsilon_i/8\} \cap \xi_i \cap \mathcal{E}_i.$$

This concludes the proof. ■

Proposition 33 *We have that for $i \leq M$*

$$\mathbb{P}\left(\xi^{(A)}_{i+1} \middle| \xi_i \cap \mathcal{E}^c_i\right) = 1.$$

Proof On $\xi_i \cap \mathcal{E}^c_i$, we know that $m_i = \lfloor \frac{l_i + r_i}{2} \rfloor$ and

$$\mu_{m_i} \leq \tau_i + \frac{1}{8}\varepsilon_i = \tau - \frac{5}{8}\varepsilon_i,$$

and

$$\mu_{l_i} \vee \mu_{r_i} \geq \tau - \varepsilon_i,$$

and so by Lemma 29 we conclude that for any $k \leq K$, $\mu_k < \tau - \frac{1}{8}\varepsilon_i$. And so $\xi^{(A)}_{i+1}$ holds. ■

Corollary 34 *We have that*

$$\mathbb{P}(\xi_{i+1}|\xi_i) \geq 1 - 2\delta'_i$$

Proof This holds by combining Propositions 32 and Proposition 33. ■

Hence by Corollary 34 and for any $I \leq M$ we have,

$$\mathbb{P}(\cap_{i \leq I} \xi_i) \geq \prod_{i \leq I}(1 - 2\delta'_i) \geq 1 - 2\sum_{i=1}^{I} \delta'_i.$$

For $I, i \leq M$ consider the event

$$\eta^I_i := \left\{ \begin{array}{l} |\hat{\mu}_{m,i} - \mu_{m_i}| \vee |\hat{\mu}_{l,i} - \mu_{l_i}| \vee |\hat{\mu}_{r,i} - \mu_{r_i}| \vee \\ |\hat{\mu}_{l-1,i} - \mu_{l_i - 1}| \vee |\hat{\mu}_{r+1,i} - \mu_{r_i+1}| \leq \dfrac{1}{16}\varepsilon_i \vee \varepsilon_I \end{array} \right\}, \tag{31}$$

which via Azuma's martingale inequality occurs with probability greater than,

$$1 - 10\exp\left(-\frac{1}{2}T^{(i)}_2 \varepsilon^2_i\right) \geq 1 - 10\delta_i. \tag{32}$$

Proposition 35 *Fix $I \leq M$ and assume that there exists k such that $\mu_k > \tau - \frac{1}{8}\varepsilon_I$. On ξ_I, we have that $\{k : \mu_k \geq \tau\} \subset \{l_I, \dots, r_I\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$.*

Proof First note that under the condition $\mu_k > \tau - \frac{1}{8}\varepsilon_I$ we have that $\xi_I^{(L)} \cap \xi_I^{(R)}$ holds. Therefore the second inclusion holds, see Corollary 34 and the definition of ξ_I. Now assume $\{k : \mu_k = \tau\} \neq \emptyset$. Let k^* be as in the proof of Lemma 28. By definition of ξ_I and since $(\mu_k)_k$ is concave, it is clear that $l_I \leq k^* \leq r_I$. The first inclusion then follows again by definition of ξ_I. In the case where $\{k : \mu_k = \tau\} = \emptyset$ the first inclusion is obvious. ■

Proposition 36 *Fix $I \leq M$ and assume that for all k, $\mu_k \leq \tau - \frac{1}{8}\varepsilon_I$. On $\xi_I \cap (\cap_{i \leq M} \eta_i^I)$, we have that $\hat{S} = \emptyset$.*

Proof Under the conditions of the proposition we have that $\mu_{m_i} \leq \tau - \frac{1}{8}\varepsilon_I$, for all i and this implies the result by definition of the η_i^I and $\mathcal{I}_m$. ■

Proposition 37 *Fix $I \leq M$. On $\xi_I \cap \left(\cap_{i \leq M} \eta_i^I \right)$, we have that*

$$\mathcal{I}_m \subset \{l_I, \dots, r_I\},$$

and also

$$l_I \in \mathcal{I}_l \quad r_I \in \mathcal{I}_r.$$

Proof On $\cap_{i \leq I} \eta_i^I$, we have that $\mathcal{I}_m \subset \{k : \mu_k \geq \tau\} \cup \{l_I, \dots, r_I\}$, and so from Propositions and 35 and 36, we have on $\cap_{i \leq I} \eta_i^I \cap \xi_I$, that $\mathcal{I}_m \subset \{l_I, \dots, r_I\}$.

The proof that $l_I \in \mathcal{I}_l$ on $\xi_I \cap \eta_I^I$ - as well as the fact that $r_I \in \mathcal{I}_r$ - follows immediately by combining the definition of $\mathcal{I}_l$ - resp. $\mathcal{I}_r$ - with Proposition 35 and 36, and the definition of η_I^I. ■

Proposition 38 *Fix $I \leq M$, and assume that $m_I \notin \mathcal{I}_m$. On $\xi_I \cap \left(\cap_{i \leq I} \eta_i^I \right)$, we have that $\{k : \mu_k \geq \tau + 4\varepsilon_i\} \subset \emptyset \subset \{\hat{l}, \dots, \hat{r}\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$.*

Proof On $\xi_I \cap \left(\cap_{i \leq I} \eta_i^I \right)$ we have from Proposition 37 that $\mathcal{I}_m \subset \{l_I, \dots, r_I]$ and that $l_I \in \mathcal{I}_l, r_I \in \mathcal{I}_r$. This implies that on $\xi_I \cap \left(\cap_{i \leq I} \eta_i^I \right)$, $\{\hat{l}, \dots, \hat{r}\} \subset \{l_I, \dots, r_I\}$. Together with Propositions 35 and 36 this implies that on $\xi_I \cap \left(\cap_{i \leq I} \eta_i^I \right)$ we have $\{\hat{l}, \dots, \hat{r}\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$.

Moreover, on η_I^I, we have by the assumption of Proposition 38 that $\mu_{m_I} \leq \tau + \frac{17}{8}\varepsilon_I$. Together with Proposition 35 and 36 and Lemma 29, this implies that on $\xi_I \cap \eta_i^I$, $\forall k \leq K, \mu_k \leq \tau + 4\varepsilon_I$. This concludes the proof with the fact that $\{\hat{l}, \dots, \hat{r}\} \subset \{l_I, \dots, r_I\}$. ■

Proposition 39 *Fix $I \leq M$, and assume that $m_I \in \mathcal{I}_m$. On $\left(\cap_{i \leq I} \xi_i \right) \cap \left(\cap_{i \leq M} \eta_i^I \right)$, we have that $\{k : \mu_k \geq \tau + \varepsilon_I\} \subset \{\hat{l}, \dots, \hat{r}\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$.*

Proof As in the proof of Proposition 38, we have on $\xi_I \cap \left(\cap_{i \leq I} \eta_i^I \right)$ that it holds that $\{\hat{l}, \dots, \hat{r}\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$. Under the event η_I^I as $\hat{l} \in \mathcal{I}_l, \hat{r} \in \mathcal{I}_r$ we have that,

$$\mu_{\hat{l}-1} < \tau + \varepsilon_I \ \& \ \mu_{\hat{r}+1} < \tau + \varepsilon_I.$$

Moreover, on η_I^I, we have by the assumption of Proposition 39 that $\mu_{m_I} \geq \tau + \frac{15}{8}\varepsilon_I$. Therefore, as $\mu_{m_I} \in \{\hat{l} - 1, \dots, \hat{r} + 1\}$ via the concavity of $(\mu_k)_{k<K}$ we have that $\{k : \mu_k \geq \tau + \varepsilon_I\} \subset \{\hat{l}, \dots, \hat{r}\}$. This concludes the proof. ■

Proof [Proof of Proposition 27] Let $I \leq M$. Combining Propositions 38 and 39, we have on $\left(\cap_{i \leq I} \xi_i \right) \cap \left(\cap_{i \leq M} \eta_i^I \right)$ that

$$\{k : \mu_k \geq \tau + 4\varepsilon_I\} \subset \{\hat{l}, \dots, \hat{r}\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}.$$

Note that

$$\mathbb{P}\left[\left(\cap_{i \leq I} \xi_i \right) \cap \left(\cap_{i \leq M} \eta_i^I \right) \right] \geq 1 - 10 \sum_{i \leq I} \delta_i - \sum_{i \leq I} \delta_i' - (M - I)\delta_I.$$

We have by definition of $\delta_i', T_2^{(i)}$ that

$$\delta_i' \leq \min\left(\frac{1}{\log(K)^{3/4}}, 72 \log\log(K) \delta_i^2 \right),$$

and also we have that $\delta_i = 2^{i-M}$ so that whenever $M - i \geq \log\log\log(K)$, we have that $\log\log(K)\delta_i^2 \leq \delta_i$. And so

$$\sum_{i \leq I} \delta_i' \leq 144 \delta_i,$$

since when $M - i \geq \log\log\log(K)$, we have $72\delta_i \geq \frac{1}{\log(K)^{3/4}}$. And so

$$\mathbb{P}\left[\left(\cap_{i \leq I} \xi_i \right) \cap \left(\cap_{i \leq M} \eta_i^I \right) \right] \geq 1 - 164\delta_I - (M - I)\delta_I = 1 - (M - I + 164)2^{I-M}.$$

Thus for any $i \in \{0, \dots, M\}$ we have

$$\mathbb{P}\left[R_T \geq 4\left(\frac{7}{8}\right)^{M-i} \right] \leq (i + 164)2^{-i} \leq 200\left(\frac{2}{3}\right)^i.$$

This concludes the proof by summing over I for finding the expected regret, and noting that there exists a universal constant $C > 0$ such that $\left(\frac{7}{8}\right)^M = \varepsilon_M \leq C\sqrt{\frac{\log\log K}{T}}$, by definition of M. ■

Appendix G. Extension of results to fixed confidence setting

Fixed confidence setting. In this section we extend our results to the fixed confidence setting for the *MTBP* and *TBP*. In this case, we define $\delta, \varepsilon > 0$, to be respectively the target confidence, and target precision of our algorithm. We say that a strategy π is (ε, δ)-PAC if it stops sampling at some stopping time $\hat{T}^{\pi}_{\varepsilon,\delta}$ of its choice, and satisfies that with probability larger than $1-\delta$, $R_T^{\underline{\nu},\pi} \leq \varepsilon$. In this setting the aim is to find a (ε, δ)-PAC strategy that minimises the expected stopping time $\mathbb{E}_{\underline{\nu}}[\hat{T}^{\pi}_{\varepsilon,\delta}]$. The following Corollaries are an immediate consequence of our previous results, thus we omit proofs.

G.1. Lower Bounds

The following corollary is a direct extension to Proposition 8 which provides a lower bound in the unstructured case.

Corollary 40 *Let $\varepsilon, \delta > 0$. It holds that for any strategy π that stops at a stopping time $\hat{T}^{\pi}_{\varepsilon,\delta}$ and that is (ε, δ)-PAC, there exists a unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, such that*

$$\mathbb{E}_{\underline{\nu}}[\hat{T}^{\pi}_{\varepsilon,\delta}] \geq \frac{2\sigma^2 K \max(\log(K), 2)(1 - K^{-1} - \delta)^2}{\varepsilon^2}\,.$$

Proof Consider the notations of the proof of Proposition 8. Assume that there exists an (ε, δ)-PAC strategy π such that for all $Q \in \{-1, 1\}^K$, we have

$$\mathbb{E}_{Q}[\hat{T}^{\pi}_{\varepsilon,\delta}] < \frac{2\sigma^2 K \max(\log(K), 2)(1 - 1/K - \delta)^2}{\varepsilon^2}\,.$$

From the proof of Proposition 8 it holds

$$\frac{1}{2^K}\sum_{Q} \mathbb{P}_Q(\hat{Q} = Q) \leq 1/K + \sqrt{\sup_{Q' \in \{-1,1\}^K} \mathbb{E}_Q[\hat{T}^{\pi}_{\varepsilon,\delta}]\varepsilon^2/(2K\sigma^2 \max(\log(K), 2))}\,.$$

And so there is a contradiction:

$$\inf_{Q} \mathbb{P}_Q(\hat{Q} = Q) < 1 - \delta\,.$$

■

Combining this result with the lower bound from Theorem 2 of Chen et al. (2014), we obtain that for any (ε, δ)-PAC strategy, there exists a bandit problem where all arms are 1/4-sub-Gaussian and such that the expected stopping time is of higher order than $\frac{K \log(K/\delta)}{\varepsilon^2}$, since they prove that the expected stopping time for any (ε, δ)-PAC strategy is higher than $\frac{K \log(1/\delta)}{\varepsilon^2}$, on some bandit problem.

The following corollary is a direct extension to Proposition 11 which provides a lower bound in the monotone case.

Corollary 41 *Let $\varepsilon, \delta > 0$ and $K \geq 2$. It holds that for any strategy π that stops at a stopping time $\hat{T}_{\varepsilon,\delta}$ and that is (ε,δ)-PAC, there exists a unstructured bandit problem $\underline{\nu} \in \mathcal{B}_m$, such that*

$$\mathbb{E}_{\underline{\nu}}[\hat{T}_{\varepsilon,\delta}] \geq \frac{2\sigma^2 \max(2, \log(K))(1 - K^{-1} - \delta)^2}{\varepsilon^2}.$$

A very similar result was already obtained in Karp and Kleinberg (2007) , but for Bernoulli random variables in the lower bound, and without providing an explicit dependence on δ. In the paper Ben-Or and Hassidim (2008), they refine this bound in the case of fixed probability of error which implies that for any strategy that (ε,δ)-PAC, there exists a structured bandit problem where all arms are $1/4$-sub-Gaussian and such that the expected stopping time is of higher order than $(1-\delta)\log(K)/\varepsilon^2$ *up to terms that are negligible with respect to* $\log(K)/\varepsilon^2$ - which is essentially the same as what we have.

We say that a strategy is optimal if its expected simple regret (or its expected stopping time for the fixed confidence setting) matches one of this lower bounds *up to a universal constant.*

G.2. Upper Bounds

The following Corollary is a direct extension to Proposition 10, which provides an upper bound on regret of the `Uniform` algorithm.

Corollary 42 *Let $\varepsilon, \delta > 0$. For any unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, Algorithm `Uniform` launched with parameter $T := \lfloor \frac{2\sigma^2 K \log(2K/\delta)}{\varepsilon^2} \rfloor + K$ is (ε,δ)-PAC.*

Interestingly the stopping time can be taken here as deterministic, and this matches up to a multiplicative constant the lower bound in Corollary 40 combined with the one in Chen et al. (2014).

The following Corollary is a direct extension to Corollary13 which provides an upper bound on the regret of the `MTB` algorithm,

Corollary 43 *Let $\varepsilon, \delta > 0$. For any problem $\underline{\nu} \in \mathcal{B}_s$, algorithm `MTB` launched with parameter $T := \lfloor \frac{21\sigma^2 \log(K)}{\varepsilon^2} + 12\log(K) \rfloor$ if $\delta \geq K^{-3/4}$ and $T := \lfloor \frac{432\sigma^2 \log(K)\log(9/\delta)}{\varepsilon^2} + 12\log(K) \rfloor$ otherwise, is (ε,δ)-PAC.*

Interestingly, the stopping time can be taken here as constant. For δ large enough i.e. $\delta \geq K^{-3/4}$, yet smaller than any universal constant strictly smaller than 1, this is order optimal up to a multiplicative constant - see Corollary 41. For δ smaller, this is order optimal up to a multiplicative constant that depends on δ - and it is an open question to obtain optimality in this case.

Similar results can be obtained in *UTBP* and *CTBP*.

Appendix H. Supplementary discussion concerning the *TBP* and *MTBP*

H.1. Comparison of *TBP* and *MTBP* and focus on the main difference coming from the monotone structure

In the *TBP*, the proof of the bound of algorithm `Uniform` is very classical. It is, as usual in bandits, event based. We consider the event where all arms concentrate around their mean

with error bounded by $O(\sqrt{K\log(K/\delta)/T})$ - where the $\log(K/\delta)$ term comes from a union bound over all K arms - and prove that on this event the regret is bounded. The lower bound is slightly less classical when it comes to the bandit literature, and is close in spirit to the use of a sequential version of Fano's inequality - stating effectively that the union bound in the analysis of the event on the means is tight.

In the *MTBP*, however, both the algorithm MTB and its proof are far less classical. As discussed in Section 1 a naive, yet suboptimal, approach to the *MTBP* is a binary search. At each step we sample an arm $O(T/\log(K))$ times and then decide to go left or right. This kind of strategy relies on making a correct decision at each step, and requires an event based analysis. The event is here that all $O(\log(K))$ sampled arms have their empirical means that concentrate around the true means at rate $\sqrt{\log(K)\log(\log(K)/\delta)/T}$ - the $\log(\log(K)/\delta)$ term coming from the union bound. This results in a regret of order $\sqrt{\log(K)\log(\log(K))/T}$, which is strictly sub-optimal. With this in mind we consider a different algorithm that performs a 'corrective' version of the binary search, i.e. a version where the algorithm can self-correct if it realises that it made a mistake This subtle, yet fundamental difference highlights the very big gap between *TBP* and *MTBP*.

H.2. Supplementary details of the related works: *TBP*

Comparing *TBP* and *MTBP* thoroughly to related work is tricky since many related works are written in the fixed confidence setting. We extend the discussion here with respect to what is done in the paper.

In the *problem independent regime* of the *TBP*, current state of the art results can be deduced from the paper Locatelli et al. (2016). A corollary to the lower bound in Locatelli et al. (2016) in the problem independent case is that for any algorithm, there exists a bandit problem where all arms have their distribution on $[0,1]$ and such that with probability larger than $1/2$, at least one arm is missclassified and at more than a strictly positive constant times $\sqrt{K/T}$ from the threshold - this is also a corollary from the lower bound in Bubeck et al. (2009) for the different problem of best arm identification. Reciprocally, the state of the art upper bound in the problem independent case is a corollary to the upper bound in Locatelli et al. (2016). In the problem independent setting, with probability larger than $1-\delta$, all arms are within a strictly positive constant times $\sqrt{K\log(K\log T/\delta)/T}$ from τ. As one can see, current state of the art upper and lower bounds are are far from matching in the *problem independent case*.

H.3. Supplementary details of the related works: *MTBP*

The papers Feige et al. (1994), Ben-Or and Hassidim (2008) and Emamjomeh-Zadeh et al. (2016) introduce a noisy binary search *with corrections*. However in the above papers the probability of making an error during the binary search is treated as fixed. But this assumption does not hold in the setting of the *MTBP*. In Nowak (2009) a more generalised version of the binary search is considered with weaker assumptions on structure, however there is no contribution to classical binary search beyond that of Karp and Kleinberg (2007).

Karp and Kleinberg (2007) consider the special case where all arms k follows a Bernoulli distribution with parameter p_k and $p_1 < ... < p_K$, and the aim is to find a i such that p_i is

close to $1/2$. In the *fixed confidence setting*, they prove that the naive binary search approach is not optimal and propose an involved exponential weight algorithm, as well as a random walk binary search, for solving the problem. They prove that for $\varepsilon, \delta > 0$ fixed, then the algorithm returns all arms above threshold with probability larger than $1-\delta$ and tolerance ε in an expected number of pulls less than a multiplicative constant *that depends on δ in a non-specified way* times $\log_2(K)/\varepsilon^2$. They prove that this is optimal up to a constant depending on δ. In the paper Ben-Or and Hassidim (2008) they refine the dependence in δ in a slightly different setting - where one has a fixed error probability. They prove that *up to terms that are negligible with respect to* $\log(K)/\varepsilon^2$, a lower bound in the expected stopping time is of order $(1-\delta)\log(K)/\varepsilon^2$.

H.4. Contribution with respect to the literature

Our contributions can be summarised are as follows:

- *Problem independent optimal rate for TBP* We provide the first -to the best of our knowledge - upper and lower bounds in the *problem independent regime* for the *TBP*- both in the fixed confidence and fixed budget setting - as well as an associated parameter-free algorithm, `Uniform`.
- *Extension of MTBP to σ^2-sub-Gaussian distribution* The lower bound and optimal algorithm proposed in Karp and Kleinberg (2007) is specific to the assumption that all arms follow a Bernoulli distribution - and related literature makes even more constraining assumptions Feige et al. (1994); Ben-Or and Hassidim (2008); Emamjomeh-Zadeh et al. (2016). An extension of their algorithms- even in the fixed confidence setting - beyond this assumption is non-trivial. We propose an algorithm whose only assumption is that the arms follow a σ^2-sub-Gaussian distribution.
- *MTBP in the fixed budget setting* We treat in a problem independent optimal way the *fixed budget setting*. The algorithms proposed in Karp and Kleinberg (2007) - as well as in Feige et al. (1994); Ben-Or and Hassidim (2008); Emamjomeh-Zadeh et al. (2016) in a more restricted setting regarding the error distributions - operate in the fixed confidence setting. Adapting their results to a fixed budget setting is challenging, in particular since we consider the *expected maximal gap* as a measure of performance - see Section 2.
- *Simultaneous bound on all probability* The `MTB` regret bound holds simultaneously across all probabilities. That is for all $\delta > 0$ and after T rounds of our algorithm, we have a guarantee that with probability larger than $1-\delta$, the simple regret will be bounded depending on δ. This is in strong contrast to what is done in the fixed confidence literature Karp and Kleinberg (2007); Ben-Or and Hassidim (2008); Emamjomeh-Zadeh et al. (2016); Chen et al. (2014), where δ is given as a parameter to the algorithm, and where the behaviour of the algorithm is only studied on an event of probability $1-\delta$, and a clear improvement with respect to Karp and Kleinberg (2007) where the dependence in δ is not explicitly stated in the bound on regret. Our result is more general, as it allows us to get a bound on the *expected simple regret* for the fixed budget setting, but also to easily transform our algorithm to the fixed confidence setting.

We also refer to Table 3 for a comprehensive summary of state of the art rates, as well as of our rates.

H.5. Problem dependent regime

While not the focus of this paper we comment on the performance of our algorithms in the problem dependent regime for the *TBP* and *MTBP*. The problem dependent regime is defined as follows: for some sequence $\Delta \in \mathbb{R}_+^K$ we consider a sub class of problems $\mathcal{B}^\Delta \subset \mathcal{B}$ where

$$\mathcal{B}^\Delta = \{\nu \in \mathcal{B} : \forall k \in [K],\ |\mu_k - \tau| = \Delta_k\}\ .$$

Similarly we can define

$$\mathcal{B}_m^\Delta = \{\nu \in \mathcal{B}_m : \forall k \in [K],\ |\mu_k - \tau| = \Delta_k\}\ .$$

The mechanics of the game are then identical to those described in Section 2 with the exception that we consider a modified version the simple regret

$$\tilde{R}_T^{\nu,\pi} = \mathbb{P}_{\underline{\nu}}\Big(\exists k \in [K] : \hat{Q}_k^\pi \neq Q_k\Big)\,,$$

that is, the probability the learner makes at least one miss classification - which is more relevant than the simple regret considered in this paper in the regime where the Δ_k are not very small, depending on T, K.

In the case of the *TBP* consider the class of problems $\mathcal{B}^\Delta$ for some $\Delta \in \mathbb{R}_+^K$. An upper bound on the simple regret of the order $\exp\left(-c\frac{1}{K}\sum \Delta_i^2 \frac{T}{K} + c' \log(\log(T)K)\right)$ is provided from Locatelli et al. (2016), for the APT algorithm that does not take any parameters - where $c, c' > 0$ are universal constants. A matching lower bound is also provided in Locatelli et al. (2016), up to universal constants in the exponential. In the same setting we can upper bound the simple regret of the `Uniform` algorithm by $\sum_k \exp\left(-c\Delta_k^2 \frac{T}{K}\right)$, where $c > 0$ is a universal constant. Clearly the uniform algorithm under performs heavily in cases with high variance across the gaps, this should not come as a surprise.

In the case of the *MTBP* consider the class of problems $\mathcal{B}_m^\Delta$ for some $\Delta \in \mathbb{R}_+^K$. We can construct and immediate lower bound on the simple regret of the order $\exp\left(-cT \min_{k\in[K]} \Delta_k^2\right)$ - where $c > 0$ is some universal constant - while the `MTB` algorithm achieves an upper bound of the order $\exp\left(-c\frac{T}{\log(K)} \min_{k\in[K]} \Delta_k^2\right)$ - where $c > 0$ is some (different) universal constant. Thus, while it is not optimal, the algorithm `MTB` is nevertheless quite efficient in the problem dependent setting.

Appendix I. Supplementary discussion

I.1. Parameters of the algorithms

The `Uniform` algorithm only takes T as a parameter, see Subsection I.2 for a discussion on how to make it anytime. The `MTB` algorithm takes only σ, K, T as parameters. Again, see Subsection I.2 for an anytime version. Getting rid of σ is however more tricky and is an open problem. We believe that in some pathological situations, the knowledge of σ is necessary.

	State of the art		Our results	
	LB	UB	LB	UB
TBP FB [1] (Locatelli et al., 2016)	$\sqrt{\frac{K}{T}}$	$\sqrt{\frac{K\log(K\log T)}{T}}$	$\sqrt{\frac{K\log(K)}{T}}$ 4	$\sqrt{\frac{K\log(K)}{T}}$ 5
TBP FC (Chen et al., 2014)	$\frac{K\log(\delta^{-1})}{\varepsilon^2}$	$\frac{K\log(K^2\varepsilon^{-2}\delta^{-1})}{\varepsilon^2}$	$\frac{K\log(K)(1-K^{-1}-\delta)}{\varepsilon^2}$ 6	$\frac{K\log(K\delta^{-1})}{\varepsilon^2}$
MTBP FB	None	None	$\sqrt{\frac{\log(K)}{T}}$	$\sqrt{\frac{\log(K)}{T}}$
MTBP FC [2] (Karp and Kleinberg, 2007)	$\frac{\underline{c}_\delta\log(K)}{\varepsilon^2}$ 3	$\frac{\bar{c}_\delta\log(K)}{\varepsilon^2}$	$\frac{(1-K^{-1}-\delta)\log(K)}{\varepsilon^2}$ 7	$\frac{\log(K)\log(\delta^{-1})}{\varepsilon^2}$ 8

Table 3: Upper and lower bounds on the expected simple regret in the fixed budget (FB) setting and on the expected stopping time for (ε,δ)-PAC strategies in the fixed confidence (FC) setting. All results are given up to universal multiplicative constant - in the case where the sub-Gaussian parameter σ is set to 1. *Left:* previous state of the art bounds. *Right:* bounds from our paper.

Note however that it is a very mild assumption. Indeed σ comes from Definition 7.In many case, natural choices for σ are available - for instance if reward are bounded. Regarding UTB and CTB, simple extensions can be made so that they also consider the sub-Gaussian case.

I.2. Making the algorithms anytime

Although the Uniform algorithm, for simplicity, takes a known budget T it can trivially be extended to an anytime algorithm. With T unknown one can easily obtain a uniform distribution of pulls by repeatedly pulling all arms once in a batch until the "unknown" budget is expended.

In the case of the MTB Algorithm such a trivial extension is not possible. At each time step the number of times the arms in the current node are pulled is dependant upon budget T. Now note that it is possible to apply a doubling trick to our problem. I.e. first call the algorithm MTB with budget $T = \lfloor 6\log(K)\rfloor + 1$, and then until the algorithm is stopped,

4. See also Bubeck et al. (2009) for the LB.
5. Here $\underline{c}_\delta, \bar{c}_\delta > 0$ is a function of δ that is left unspecified in Karp and Kleinberg (2007).
6. See also Ben-Or and Hassidim (2008) for the LB $\frac{(1-\delta)\log(K)}{\varepsilon^2}$ up to terms that are negligible with respect to $\log(K)/\varepsilon^2$.
7. In Locatelli et al. (2016) The problem complexity H is upper bounded by K/ε^2. Replacing H with such provides the given upper bound
8. The lower bound is well known, see Bubeck et al. (2009).
9. And combining this with the lower bound in Chen et al. (2014), we get the problem independent lower bound of order $\frac{K\log(K\delta^{-1})}{\varepsilon^2}$ that matches our upper bound.
10. See also Ben-Or and Hassidim (2008) for a LB that is essentially equivalent to this.
11. In the case where $\delta \geq K^{-3/4}$ and is smaller than any universal constant strictly smaller than 1, our UB is more refined and of order $\frac{\log(K)}{\varepsilon^2}$, which is order optimal.

always double the budget and call algorithm MTB from scratch. Then when the algorithm is stopped, recommend the arm recommended by the last full iteration. Note that this arm will have been selected with at least a fourth of the budget, and so Proposition 12 and Corollary 13 hold with the doubling trick and therefore without taking T as parameter, and replacing T by $T/4$ in the bound. Similar tricks hold also for UTB and CTB.

I.3. Computational complexity

The computational complexity of both our algorithms is very low. Algorithm Uniform is just uniform sampling, and then a computation of K empirical means and their comparison to the threshold. I.e. this is in total n operations (where by operations we mean addition or comparisons), and needs to store only K variables, i.e. the empirical means.

Algorithm MTB consists of

- first running Algorithm Explore, which consists just in computing about $\log(K)$ empirical means, and taking decisions based on them. The algorithm just needs to perform n operations (where by operations we mean addition or comparisons), and needs to store only about $\log K$ variables, i.e. the empirical means and position of sampled arms.
- then running Algorithm Choose which consists in scanning one time the list of sampled arms, i.e. doing about $\log(K)$ operations, and returning the median. The number of operations is therefore of order $\log(K)$ and the algorithm needs to store only about $\log(K)$ variables, i.e. the empirical means and position of relevant sampled arms.

Similarly, the computational complexity of UTB and CTB is also low.