An optimal algorithm for the Thresholding Bandit Problem

Andrea Locatelli   ANDREA.LOCATELLI@UNI-POTSDAM.DE
Maurilio Gutzeit   MGUTZEIT@UNI-POTSDAM.DE
Alexandra Carpentier   CARPENTIER@UNI-POTSDAM.DE
Department of Mathematics, University of Potsdam, Germany

Abstract

We study a specific *combinatorial pure exploration stochastic bandit problem* where the learner aims at finding the set of arms whose means are above a given threshold, up to a given precision, and *for a fixed time horizon*. We propose a parameter-free algorithm based on an original heuristic, and prove that it is optimal for this problem by deriving matching upper and lower bounds. To the best of our knowledge, this is the first non-trivial pure exploration setting with *fixed budget* for which optimal strategies are constructed.

1. Introduction

In this paper we study a specific *combinatorial, pure exploration, stochastic bandit setting*. More precisely, consider a stochastic bandit setting where each arm has mean μ_k. The learner can sample sequentially $T > 0$ samples from the arms and aims at finding as efficiently as possible the set of arms whose means are larger than a threshold $\tau \in \mathbb{R}$. In this paper, we refer to this setting as the *Thresholding Bandit Problem (TBP)*, which is a specific instance of the combinatorial pure exploration bandit setting introduced in (Chen et al., 2014). A simpler "one armed" version of this problem is known as the SIGN-ξ problem, see (Chen & Li, 2015).

This problem is related to the popular combinatorial pure exploration bandit problem known as the TopM problem where the aim of the learner is to return the set of M arms with highest mean (Bubeck et al., 2013b; Gabillon et al., 2012; Kaufmann et al., 2015; Zhou et al., 2014; Cao et al., 2015) - which is a combinatorial version of the best arm identification problem (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004; Bubeck et al., 2009; Audibert & Bubeck,

2010; Gabillon et al., 2012; Jamieson et al., 2014; Karnin et al., 2013; Kaufmann et al., 2015; Chen & Li, 2015). To formulate this link with a simple metaphor, the TopM problem is a "contest" and the TBP problem is an "exam": in the former, the learner wants to select the M arms with highest mean, in the latter the learner wants to select the arms whose means are higher than a certain threshold. We believe that this distinction is important and that in many applications the TBP problem is more relevant than the TopM, as in many domains one has a natural "efficiency", or "correctness" threshold above which one wants to use an option. For instance in industrial applications, one wants to keep a machine if its production's value is above its functioning costs, in crowd-sourcing one wants to hire a worker as long as its productivity is higher than its wage, etc. In addition to these applications derived from the TopM problem, the TBP problem has applications in dueling bandits and is a natural way to cast the problem of active and discrete level set detection, which is in turn related to the important applications of active classification, and active anomaly detection - we detail this point more in Subsection 3.1.

As mentioned previously, the TBP problem is a specific instance of the combinatorial pure exploration bandit framework introduced in (Chen et al., 2014). Without going into the details of the combinatorial pure exploration setting for which the paper (Chen et al., 2014) derives interesting general results, we will summarize what these results imply for the particular TBP and TopM problems, which are specific cases of the combinatorial pure exploration setting. As it is often the case for pure exploration problems, the paper (Chen et al., 2014) distinguishes between two settings:

- The *fixed budget setting* where the learner aims, given a fixed budget T, at returning the set of arms that are above the threshold (in the case of TBP) or the set of M best arms (in the case of TopM), with highest possible probability. In this setting, upper and lower bounds are on the *probability of making an error when returning the set of arms*.

- The *fixed confidence setting* where the learner aims, given a probability δ of acceptable error, at returning the set of arms that are above the threshold (in the case of TBP) or the set of M best arms (in the case of TopM) with as few pulls of the arms as possible. In this setting, upper and lower bounds are on the *number of pulls T that are necessary to return the correct set of arm with probability at least $1 - \delta$*.

The similarities and dissemblance of these two settings have been discussed in the literature in the case of the TopM problem (in particular in the case $M = 1$), see (Gabillon et al., 2012; Karnin et al., 2013; Chen et al., 2014). While as explained in (Audibert & Bubeck, 2010; Gabillon et al., 2012), the two settings share similarities in the specific case when additional information about the problem is available to the learner (such as the complexity H defined in Table 1), they are very different in general and results do not transfer from one setting to the other, see (Bubeck et al., 2009; Audibert & Bubeck, 2010; Karnin et al., 2013; Kaufmann et al., 2015). In particular we highlight the following fact: while the *fixed confidence setting* is relatively well understood in the sense that there are constructions for optimal strategies (Kalyanakrishnan et al., 2012; Jamieson et al., 2014; Karnin et al., 2013; Kaufmann et al., 2015; Chen & Li, 2015), there is an important knowledge gap in the *fixed budget setting*. In this case, without the knowledge of additional information on the problem such as e.g. the complexity H defined in Table 1, there is a gap between the known upper and lower bounds, see (Audibert & Bubeck, 2010; Gabillon et al., 2012; Karnin et al., 2013; Kaufmann et al., 2015). This knowledge gap is more acute for the general combinatorial exploration bandit problem defined in the paper (Chen et al., 2014) (see their Theorem 3) - and therefore for the TBP problem (where in fact no fixed budget lower bound exists to the best of our knowledge). We summarize in Table 1 the state of the art results for the TBP problem and for the TopM problem with $M = 1$.

The summary of Table 1 highlights that in the fixed budget setting, both for the TopM and the TBP problem, the correct *complexity* H^* that should appear in the bound, i.e. what is the problem dependent quantity H^* such that the upper and lower bounds on the probability of error is of order $\exp(-n/H^*)$, is still an open question. In the TopM problem, Table 1 implies that $H \leq H^* \leq \log(2K)H_2$. In the TBP problem, Table 1 implies $0 \leq H^* \leq \log(2K)H_2$, since to the best of our knowledge a lower bound for this problem exists only in the case of the fixed confidence setting. Note that although this gap may appear small in particular in the case of the TopM problem as it involves "only" a $\log(K)$ multiplicative factor, it is far from being negligible since the $\log(K)$ gap factor acts on a term of order exponential minus T exponentially.

Problem	Lower Bound	Upper Bound
TBP (FC)	$H \log\left(\frac{1}{\delta}\right)$	$H \log\left(\frac{1}{\delta}\right)$
TBP (FB)	No results	$K \exp\left(-\frac{T}{\log(K)H_2}\right)$
TopM (FC)	$H \log\left(\frac{1}{\delta}\right)$	$H \log\left(\frac{1}{\delta}\right)$
TopM (FB)	$\exp\left(-\frac{T}{H}\right)$	$K \exp\left(-\frac{T}{\log(K)H_2}\right)$

Table 1. State of the art results for the TBP problem and the TopM problem with $M = 1$ with fixed confidence Δ for δ small enough (FC) and fixed budget (FB) - for FC, bound on the expected total number of samples needed for making an error of at most δ on the set of arms and for FB, bound on the probability of making a mistake on the returned set of arms. The quantities H, H_2 depend on the means μ_k of the arm distributions and are defined in (Chen et al., 2014) (and are not the same for TopM and TBP). In the case of the TBP problem, set $\Delta_k = |\tau - \mu_k|$ and set $\Delta_{(k)}$ for the Δ_k ordered in increasing order, we have $H = \sum_i \Delta_i^{-2}$ and $H_2 = \min_i i\Delta_{(i)}^{-2}$. For the TopM problem with $M = 1$, the same definitions holds with $\Delta_k = \max_i \mu_i - \mu_k$.

In this paper we close, up to constants, the gap in the fixed budget setting for the TBP problem - we prove that $H^* = H$. In addition, we also prove that our strategy minimizes at the same time the cumulative regret, and identifies optimally the best arm, provided that the highest mean of the arms is known to the learner. Our findings are summarized in Table 1. In order to do that, we introduce a new algorithm for the TBP problem which is entirely *parameter free*, and based on an original heuristic. In Section 2, we describe formally the TBP problem, the algorithm, and the results. In Section 3, we describe how our algorithm can be applied to the active detection of discrete level sets, and therefore to the problem of active classification and active anomaly detection. We also describe what are the implications of our results for the TopM problem. Finally Section 4 presents some simulations for evaluating our algorithm with respect to the state of the art competitors. The proofs of all theorems are in Appendix A, as well as additional simulation results.

Problem	Results
TBP (FB) : UB	$\exp\left(-\frac{T}{H} + \log\left(\log(T)K\right)\right)$
TBP (FB) : LB	$\exp\left(-\frac{T}{H} - \log\left(\log(T)K\right)\right)$
TopM (FB) : UB (with μ^* known)	$\exp\left(-\frac{T}{H} + \log\left(\log(T)K\right)\right)$

Table 2. Our results for the TopM and the TBP problem in the fixed budget setting - i.e. upper and lower bounds on the probability of making a mistake on the set of arms returned by the learner.

2. The Thresholding Bandit Problem

2.1. Problem formulation

Learning setting Let K be the number of arms that the learner can choose from. Each of these arms is characterized by a distribution ν_k that we assume to be R-sub-Gaussian.

Definition (R-sub-Gaussian distribution). *Let $R > 0$. A distribution ν is R-sub-Gaussian if for all $t \in \mathbb{R}$ we have*

$$\mathbb{E}_{X \sim \nu}[\exp(tX - t\mathbb{E}[X])] \leq \exp(R^2 t^2/2).$$

This encompasses various distributions such as bounded distributions or Gaussian distributions of variance R^2 for $R \in \mathbb{R}$. Such distributions have a finite mean, let $\mu_k = \mathbb{E}_{X \sim \nu_k}[X]$ be the mean of arm k.

We consider the following dynamic game setting which is common in the bandit literature. For any time $t \geq 1$, the learner chooses an arm I_t from $\mathbb{A} = \{1, ..., K\}$. It receives a noisy reward drawn from the distribution ν_{I_t} associated to the chosen arm. An adaptive learner bases its decision at time t on the samples observed in the past.

Set notations Let $u \in \mathbb{R}$ and $\mathbb{A}$ be the finite set of arms. We define S_u as the set of arms whose means are over u, that is $S_u := \{k \in \mathbb{A}, \mu_k \geq u\}$. We also define S_u^C as the complimentary set of S_u in $\mathbb{A}$, i.e. $S_u^C = \{k \in \mathbb{A}, \mu_k < u\}$.

Objective Let $T > 0$ (not necessarily known to the learner beforehand) be the horizon of the game, let $\tau \in \mathbb{R}$ be the *threshold* and $\epsilon \geq 0$ be the *precision*. We define the (τ, ϵ) thresholding problem as such : after T rounds of the game described above, the goal of the learner is to correctly identify the *arms whose means are over or under the threshold τ up to a certain precision ϵ*, i.e. to correctly discriminate arms that belong to $S_{\tau+\epsilon}$ from those in $S_{\tau-\epsilon}^C$. In the rest of the paper, the sentence "the arm is over the threshold τ" is to be understood as "the arm's mean is over the threshold".

After T rounds of the previously defined game, the learner has to output a set $\widehat{S}_\tau := \widehat{S}_\tau(T) \subset \mathbb{A}$ of arms and it suffers the following loss:

$$\mathcal{L}(T) = \mathbb{I}(S_{\tau+\epsilon} \cap \widehat{S}_\tau^C \neq \emptyset \quad \vee \quad S_{\tau-\epsilon}^C \cap \widehat{S}_\tau \neq \emptyset).$$

A good learner minimizes this loss by correctly discriminating arms that are outside of a 2ϵ band around the threshold: arms whose means are smaller than $(\tau - \epsilon)$ should not belong to the output set $\widehat{S}_\tau$, and symmetrically those whose means are bigger than $(\tau + \epsilon)$ should not belong to $\widehat{S}_\tau^C$. If it manages to do so, the algorithm suffers no loss and otherwise it incurs a loss of 1. For arms that lie inside this 2ϵ strip, mistakes on the other hand bear no cost. If we set ϵ to 0 we recover the exact TBP thresholding problem described in the introduction, and the algorithm suffers no loss if it discriminates exactly arms that are over the threshold from those under.

Let $\mathbb{E}$ be the expectation according to the samples collected by an algorithm, its expected loss is:

$$\mathbb{E}[\mathcal{L}(T)] = \mathbb{P}(S_{\tau+\epsilon} \cap \widehat{S}_\tau^C \neq \emptyset \quad \vee \quad S_{\tau-\epsilon}^C \cap \widehat{S}_\tau \neq \emptyset),$$

i.e. it is the probability of making a mistake, that is rejecting an arm over $(\tau + \epsilon)$ or accepting an arm under $(\tau - \epsilon)$. The lower this probability of error, the better the algorithm, as an oracle strategy would simply rightly classify each arm and suffer an expected loss of 0.

Our problem is a pure exploration bandit problem, and is in fact, shifting the means by $-\tau$, a specific case of the pure exploration bandit problem considered in (Chen et al., 2014) - namely the specific case where the set of sets of arms that they call $\mathcal{M}$ and which is their decision class is the set of all possible set of arms. We will comment more on this later in Subsection 2.4.

Problem complexity We define $\Delta_i^{\tau,\epsilon}$ the gap of arm i with respect to τ and ϵ as:
$$\Delta_i := \Delta_i^{\tau,\epsilon} = |\mu_i - \tau| + \epsilon. \tag{1}$$

We also define the complexity H_ϵ of the problem as
$$H := H_{\tau,\epsilon} = \sum_{i=1}^{K} (\Delta_i^{\tau,\epsilon})^{-2}. \tag{2}$$

We call H complexity as it is a characterization of the hardness of the problem. A similar quantity was introduce for general combinatorial bandit problems (Chen et al., 2014) and is similar in essence to the complexity introduced for the best arm identification problem, see (Audibert & Bubeck, 2010).

2.2. A lower bound

In this section, we exhibit a lower bound for the thresholding problem. More precisely, for any sequence of gaps $(d_k)_k$, we define a finite set of problems where the distributions of the arms of these problems correspond to these gaps and are Gaussian of variance 1. We lower bound the largest probability of error among these problems, for the best possible algorithm.

Theorem 1. *Let $K, T \geq 0$. Let for any $i \leq K$, $d_i \geq 0$. Let $\tau \in \mathbb{R}, \epsilon > 0$.*

For $0 \leq i \leq K$, we write $\mathcal{B}^i$ for the problem where the distribution of arm $j \in \{1, \ldots, K\}$ is $\mathcal{N}(\tau + d_i + \epsilon, 1)$ if $i \neq j$ and $\mathcal{N}(\tau - d_i - \epsilon, 1)$ otherwise. For all these problems, $H := H_{\tau,\epsilon} = \sum_i (d_i + 2\epsilon)^{-2}$ is the same by definition.

It holds that for any bandit algorithm
$$\max_{i \in \{0,\ldots,K\}} \mathbb{E}_{\mathcal{B}^i}(\mathcal{L}(T)) \geq \exp\big(-3T/H -$$
$$4\log(12(\log(T)+1)K)),$$

where $\mathbb{E}_{\mathcal{B}^i}$ is the expectation according to the samples of problem $\mathcal{B}^i$.

This lower bound implies that even if the learner is given the distance of the mean of each arm to the threshold and the shape of the distribution of each arm (here Gaussian of variance 1), any algorithm still makes an error of at least $\exp(-3T/H - 4\log(12(\log(T) + 1)K))$ on one of the problems. This is a lower bound in a very strong sense because we really restrict the set of possibilities to a setting where we know all gaps and prove that nevertheless this lower bounds holds. Also it is non-asymptotic and holds for any T, and implies therefore a non-asymptotic minimax lower bound. The closer the means of the distributions to the threshold, the larger the complexity H, and the larger the lower bound. The proof is to be found in Appendix A.

This theorem's lower bound contains two terms in the exponential, a term that is linear in T and a term that is of order $\log((\log(T) + 1)K) \approx \log(\log(T)) + \log(K)$. For large enough values of T, one has the following simpler corollary.

Corollary. *Let $\bar{H} > 0$ and $R > 0$, $\tau \in \mathbb{R}$ and $\epsilon \geq 0$. Consider $\mathbb{B}_{\bar{H},R}$ the set of K-armed bandit problems where the distributions of the arms are R-sub-Gaussian and which have all a complexity smaller than $\bar{H}$.*

Assume that $T \geq 4\bar{H}R^2 \log(12(\log(T) + 1)K)$. It holds that for any bandit algorithm

$$\sup_{\mathcal{B} \in \mathbb{B}_{\bar{H},R}} \mathbb{E}_{\mathcal{B}}(\mathcal{L}(T)) \geq \exp\left(-4T/(R^2\bar{H})\right),$$

where $\mathbb{E}_{\mathcal{B}}$ is the expectation according to the samples of problem $\mathcal{B} \in \mathbb{B}_{\bar{H},R}$.

2.3. Algorithm APT and associated upper bound

In this section we introduce APT (Anytime Parameter-free Thresholding algorithm), an anytime parameter-free learning algorithm. Its heuristic is based on a simple observation, namely that a near optimal static strategy that allocates T_k samples to arm k is such that $T_k\Delta_k^2$ is constant across k (and increasing with T) - see Theorem 1, and in particular the second half of Step 3 of its proof in Appendix A - and that therefore a natural idea is to simply pull at time t the arm that minimizes an estimator of this quantity. Note that in this paper, we consider for the sake of simplicity that each arm is tested against the same threshold, however this can be relaxed to $(\tau_k)_k$ at no additional cost.

Algorithm The algorithm receives as input the definition of the problem (τ, ϵ). First, it pulls each arm of the game once. At time $t > K$, APT updates $T_i(t)$, the number of pulls up to time t of arm i, and the empirical mean $\hat{\mu}_i(t)$ of arm k after $T_i(t)$ pulls. Formally, for each $k \in \mathbb{A}$ it computes $T_i(t) = \sum_{s=1}^{t} \mathbb{I}(I_s = i)$ and the updated means

$$\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} X_{i,s}, \tag{3}$$

Algorithm 1 APT algorithm
> **Input:** τ, ϵ
> Pull each arm once
> **for** $t = K + 1$ **to** T **do**
>   Pull arm $I_t = \arg\min_{k \leq K} B_k(t)$ from Equation (5)
>   Observe reward $X \sim \nu_{I_t}$
> **end for**
> **Output:** $\hat{S}_\tau = \{k : \hat{\mu}_k(T) \geq \tau\}$

where $X_{i,s}$ denotes the sample received when pulling i for the s-th time. The algorithm then computes:

$$\widehat{\Delta}_i(s) := \widehat{\Delta}_i^{\tau,\epsilon}(s) = |\hat{\mu}_i(t) - \tau| + \epsilon, \tag{4}$$

the current empirical estimate of the gap associated with arm i. The algorithm then computes:

$$B_i(t + 1) = \sqrt{T_i(t)}\widehat{\Delta}_i(t). \tag{5}$$

and pulls the arm $I_{t+1} = \arg\min_{i \leq K} B_i(t+1)$ that minimizes this quantity. At the end of the horizon T, the algorithm outputs the set of arms $\widehat{S}_\tau = \{k : \widehat{\mu}_k(T) \geq \tau\}$.

The expected loss of this algorithm can be bounded as follows.

Theorem 2. *Let $K \geq 0, T \geq 2K$, and consider a problem $\mathcal{B}$. Assume that all arms ν_k of the problem are R-sub-Gaussian with means μ_k. Let $\tau \in \mathbb{R}, \epsilon \geq 0$*

Algorithm APT's expected loss is upper bounded on this problem as

$$\mathbb{E}(\mathcal{L}(T)) \leq \exp\left(-\frac{1}{64R^2}\frac{T}{H} + 2\log((\log(T) + 1)K)\right),$$

where we remind that $H = \sum_i (|\mu_i - \tau| + \epsilon)^{-2}$ and where $\mathbb{E}$ is the expectation according to the samples of the problem.

The bound of Theorem 2 holds for any R-sub-Gaussian bandit problem. *Note that one does not need to know R in order to implement the algorithm*, e.g. if the distributions are bounded, one does not need to know the bound. This is a desirable feature for an algorithm, yet e.g. all algorithms based on upper confidence bounds need a bound on R. This bound is non-asymptotic (one just needs $T \geq 2K$ so that one can initialize the algorithm) and therefore Theorem 2 provides a minimax upper bound result over the class of problems that have sub-Gaussian constant R and complexity H.

The term in the exponential of the lower bound of Theorem 2 matches the lower bound of Theorem 1 up to a multiplicative factor and the $\log((\log(T) + 1)K)$ term. Now as in the case of the lower bound, for large enough values of T, one has the following simpler corollary.

Corollary. *Let $\bar{H} > 0$ and $R > 0$, $\tau \in \mathbb{R}$ and $\epsilon \geq 0$. Consider $\mathbb{B}_{\bar{H},R}$ the set of K-armed bandit problems where the distributions of the arms are R-sub-Gaussian and whose complexity is smaller than $\bar{H}$.*

Assume that $T \geq 256\bar{H}R^2 \log((\log(T)+1)K)$. For Algorithm APT it holds that

$$\sup_{\mathcal{B} \in \mathbb{B}_{\bar{H},R}} \mathbb{E}_{\mathcal{B}}(\mathcal{L}(T)) \leq \exp\big(-T/(128R^2\bar{H})\big),$$

where $\mathbb{E}_{\mathcal{B}}$ is the expectation according to the samples of problem $\mathcal{B} \in \mathbb{B}_{\bar{H},R}$

This corollary and Corollary 2.2 imply that for T large enough - i.e. of larger order than $HR^2 \log((\log(T)+1)K)$ - Algorithm APT is order optimal over the class of problems whose complexity is bounded by $\bar{H}$ and whose arms are R-sub-Gaussian.

2.4. Discussion

A parameter free algorithm An important point that we want to highlight for our strategy APT is that it does not need any parameter, such as the complexity H, the horizon T or the sub-Gaussian constant R. This contrasts with any upper confidence based approach as in e.g. (Audibert & Bubeck, 2010; Gabillon et al., 2012) (e.g. the UCB-E algorithm in (Audibert & Bubeck, 2010)), which need as parameter an upper bound on R and the exact knowledge of H, while the bound of Theorem 2 will hold for any R and any H, and our algorithm adapts to these quantities. Also we would like to highlight that for the related problem of best arm identification, existing fixed budget strategies need to know the budget T in advance (Audibert & Bubeck, 2010; Karnin et al., 2013; Chen et al., 2014) - while our algorithm can be stopped at any time and the bound of Theorem 2 will hold.

Extensions to distributions that are not sub-Gaussian as opposed to adaptation to sub- models It is easy to see in the light of (Bubeck et al., 2013a) that one could extend our algorithm to non sub-Gaussian distributions by using an estimator other than the empirical means, as e.g. the estimators in (Catoni et al., 2012) or in (Alon et al., 1996). These estimators have sub-Gaussian concentration asymptotically under the only assumption that the distributions have a finite $(1 + v)$ moment with $v > 0$ (and the sub-Gaussian concentration will depend on v). Using our algorithm with a such estimator will therefore provide a result that is similar to the one of Theorem 2 - and that without requiring the knowledge of v, which means that our algorithm APT modified for using these robust estimators instead of the empirical mean will work *for any bandit problem where the arm distributions have a finite $(1 + v)$ moment with $v > 0$.* On the other hand, if we consider more specific, e.g. exponential, models, it is possible to obtain a refined lower

bound in terms of Kullback- Leibler divergences rather than gaps following (Kaufmann et al., 2015). However, an upper bound of the same order clearly comes at the cost of a more complicated strategy and holds in less generality than our bound.

Optimality of our strategy As explained previously, the upper bound on the expected risk of algorithm APT is comparable to the lower bound on the expected risk up to a $\log\big((\log(T)+1)K\big)$ term (see Theorems 2 and Theorems 1) - and this term vanishes when the horizon T is large enough, namely when $T \geq O(HR^2 \log\big((\log(T)+1)K\big))$, which is the case for most problems. So for T large enough, our strategy is order optimal over the class of problems that have complexity smaller than H and sub-Gaussian constant smaller than R.

Comparison with existing results Our setting is a specific combinatorial pure exploration setting with fixed budget where the objective is to find the set of arms that are above a given threshold. Settings related to ours have been analyzed in the literature and the state of the art result on our problem can be found (to the best of our knowledge) in the paper (Chen et al., 2014). In this paper, the authors consider a general pure exploration combinatorial problem. Given a set $\mathcal{M}$ of subsets of $\{1, \ldots, K\}$, they aim at finding a subset of arms $M^* \in \mathcal{M}$ such that $M^* = \arg\max_{M \in \mathcal{M}} \sum_{k \in M^*} \mu_k$. In the specific case where $\mathcal{M}$ is the set of all subsets of $\{1, \ldots, K\}$, their problem in the *fixed budget setting* is exactly the same as ours when $\epsilon = 0$ and the means are shifted by $-\tau$. Their algorithm CSAR's upper bound on the loss is (see their Theorem 3):

$$\mathbb{E}(\mathcal{L}(T)) \leq K^2 \exp\Big(-\frac{T-K}{72R^2 \log(K)H_{\text{CSAR},2}}\Big),$$

where $H_{\text{CSAR},2} = \max_i i\Delta_{(i)}^{-2}$. As $H_{\text{CSAR},2} \log(K) \geq H$ by definition, there is a gap for their strategy in the fixed budget setting with respect to the lower bound of Theorem 1, which is smaller and of order $\exp(-T/(HR^2))$. Our strategy on the contrary does not have this gap, and improves over the CSAR strategy. We believe that this lack of optimality for CSAR is not an artefact of the proof of the paper (Chen et al., 2014), and that CSAR is sub-optimal, as it is a successive reject algorithm with fixed and non-adaptive reject phase length. A similar gap between upper and lower bounds for successive reject based algorithms in the *fixed budget setting* was also observed for the best arm identification problem when no additional information such as the complexity are known to the learner, see (Audibert & Bubeck, 2010; Karnin et al., 2013; Kaufmann et al., 2015; Chen et al., 2014). It is therefore an interesting fact that there is a *parameter free* optimal algorithm for our *fixed budget* problem.

The paper (Chen et al., 2014) also provides results in the *fixed confidence setting*, where the objective is to provide an ϵ optimal set using the smallest possible sample size. In these results such a gap in optimality does not appear and the algorithm CLUCB they propose is almost optimal, see also (Kalyanakrishnan et al., 2012; Jamieson et al., 2014; Karnin et al., 2013; Kaufmann et al., 2015; Chen & Li, 2015) for related results in the fixed confidence setting. This highlights that the fixed budget setting and the fixed confidence setting are fundamentally different (at least in the absence of additional information such as the complexity H), and that providing optimal strategies in the fixed budget setting is a more difficult problem than providing an adaptive strategy in the fixed confidence problem - adaptive algorithms that are nearly optimal in the absence of additional information have only been exhibited in the latter case. To the best of our knowledge, all strategies except ours have such an optimality gap for fixed budget pure exploration combinatorial bandit problems, while there exists fixed confidence strategies for general pure exploration combinatorial bandits that are very close to optimal, see (Chen et al., 2014).

Now in the case where the learner has additional information on the problem, as e.g. the complexity H, it has been proved in the TopM problem that a UCB-type strategy has probability of error upper bounded as $\exp(-T/H)$, see (Audibert & Bubeck, 2010; Gabillon et al., 2012). A similar UCB type of algorithm would also work in the TBP problem, implying the same upper bound results as APT. But we would like to highlight that the *exact* knowledge of H is needed by these algorithms for reaching this bound - which is unlikely in applications. Our strategy on the other hand reaches, up to constants, the optimal expected loss for the TBP problem, without needing any parameter.

3. Extensions of our results to related settings

In this section we detail some implications of the results of the previous section to some specific problems.

3.1. Active level set detection : Active classification and active anomaly detection

Here we explain how a simple modification of our setting transforms it into the setting of *active level set detection*, and therefore why it can be applied to active classification and active anomaly detection. We define the problem of discrete, active level set detection as the problem of deciding as efficiently as possible, in our bandit setting, whether for any k the probabilities that the samples of arms ν_k are above or below a given level L are higher or smaller than a threshold τ up to a precision ϵ, i.e. it is the problem of deciding for all k whether $\tilde{\mu}_k(L) := \mathbb{P}_{X \sim \nu_k}(X > L) \geq \tau$, or not up to a precision ϵ.

This problem can be immediately solved by our approach with a simple change of variable. Namely, for the sample $X_t \sim \nu_{I_t}$ collected by the algorithm at time t, consider the transformation $\tilde{X}_t = \mathbf{1}_{X_t > L}$. Then $\tilde{X}_t$ is a Bernoulli random variable of parameter $\tilde{\mu}_{I_t}(L)$ (which is a $1/2$-sub-Gaussian distribution) - and applying our algorithm to the transformed samples $\tilde{X}_t$ solves the active level set detection problem. This has two interesting applications, namely in active binary classification and in active anomaly detection.

Active binary classification   In active binary classification, the learner aims at deciding, for k points (the arms of the bandit), whether each point belongs to the class 1 or the class 0.

At each round t, the learner can request help from a homogeneous mass of experts (which can be a set of previously trained classifiers, where one wants to minimize the computational cost, or crowd-sourcing, when one wants to minimize the costs of the task), and obtain a *noisy* label for the chosen data point I_t. We assume that for any point k, the expert's responses are independent and stochastic random variables in $\{0, 1\}$ of mean μ_k (i.e. the arm distributions are Bernoulli random variables of parameter μ_k). We assume that the experts are right on average and that the label l_k of k is equal to $l_k := \mathbf{1}\{\tilde{\mu}_k > 1/2\}$. The active classification task therefore amounts to deciding whether $\mu_k > \tau := 1/2$ or not, possibly up to a given precision ϵ. Our strategy therefore directly applies to this problem by choosing $\tau = 1/2$.

Active anomaly detection   In the case anomaly detection, a common way to characterize anomalies is to describe them as naturally *not concentrated* (Steinwart et al., 2005). A natural way to characterize anomalies is thus to define a *cutoff level* L, and classify the samples e.g. above this level L as anomalous. Such an approach has already received attention for anomaly detection e.g in (Streeter & Smith), albeit in a cumulative regret setting.

Here we consider an active anomaly detection setting where we face K sources of data (the arms), and we aim at sampling them actively to detect which sources emit anomalous samples with a probability higher than a given threshold τ - this threshold is chosen e.g. as the maximal tolerable amount of anomalous behavior of a source. This illustrates the fact that as described in (Steinwart et al., 2005), the problem of anomaly detection is indeed a problem of level set detection - and so the problem of active anomaly detection is a problem of active level set detection on which we can use our approach as explained above.

3.2. Best arm identification and cumulative reward maximization with known highest mean value

Two classical bandit problems are the best-arm identification problem and the cumulative reward maximization problem. In the former, the goal of the learner is to identify the arm with the highest mean (Bubeck et al., 2009). In the latter, the goal is to maximize the sum of the samples collected by the algorithm up to time T (Auer et al., 1995). Intuitively, both problems should call for different strategies - in the best arm identification problem one wants to explore all arms heavily while in the cumulative reward maximization problem one wants to sample as much as possible the arm with the highest mean. Such intuition is backed up by Theorem 1 of (Bubeck et al., 2009), which states that in the absence of additional information and with a fixed budget, the lower the regret suffered in the cumulative setting, expressed in terms of rewards, the higher the regret suffered in the identification problem, expressed in terms of probability of error. We prove in this section the somewhat non intuitive fact that if one knows the value of best arm's mean, its possible to perform both tasks *simultaneously* by running our algorithm where we choose $\epsilon = 0$ and $\tau = \mu^* := \max_k \mu_k$. Our algorithm then reduces to the GCL^* algorithm that can be found in (Salomon & Audibert, 2011).

Best arm identification In the best arm identification problem, the game setting is the same as the one we considered but the goal of the learner is different: it aims at returning an arm J_T that with the highest possible mean. The following proposition holds for our strategy APT that runs for T times, and then returns the arm J_T that was the most pulled.

Theorem 3. *Let $K > 0$, $R > 0$ and $T \geq 2K$ and consider a problem where the distribution of the arms ν_k is R-sub-Gaussian and has mean μ_k. Let $\mu^* := \max_k \mu_k$ and $H_{\mu^*} = \sum_{i:\mu_i \neq \mu^*} (\mu^* - \mu_i)^{-2}$.*

Then APT run with parameters $\tau = \mu^*$ and $\epsilon = 0$, recommending the arm $J_T = \arg\max_{k \in \mathbb{A}} T_k(T)$, is such that

$$\mathbb{P}(\mu_{J_T} \neq \mu^*) \leq \exp\left(-\frac{T}{36R^2 H_{\mu^*}} + 2\log(\log(T)+1)K\right).$$

If the complexity H is also known to the learner, algorithm UCB-E from (Audibert & Bubeck, 2010) would attain a similar performance.

Remark This implies that if μ^* is known to the learner, there exists an algorithm such that its probability of error is of order $\exp(-cT/H)$. The recent paper (Carpentier & Locatelli, 2016) actually implies that the knowledge of μ^* is actually key here, since without this information, the simple regret is at least of order $\exp(-cT/(\log(K)H))$ in a minimax sense.

Cumulative reward maximization In the cumulative reward maximization problem, the game setting is the same as the the one we considered but the aim of the learner is different : if we write X_t for the sample collected at time t by the algorithm, it aims at maximizing $\sum_{t \leq T} X_t$. The following proposition holds for our strategy APT that runs for T times.

Theorem 4. *Let $K > 0$, $R > 0$ and $T \geq 2K$ and consider a problem where the distribution of the arms ν_k is R-sub-Gaussian.*

Then APT run with parameters $\tau = \mu^*$ and $\epsilon = 0$ is such that

$$T\mu^* - \mathbb{E}\sum_{t \leq T} X_t \leq \inf_{\delta \geq 1}\Big[\sum_{k \neq k^*} \frac{4R^2 \log(T)\delta}{\mu^* - \mu_i} + (\mu^* - \mu_i)(1 + \frac{K}{T^{2\delta-2}})\Big].$$

This bound implies both the problem dependent upper bound of order $\sum_i \Delta_i^{-1} \log(T)$ and the problem independent upper bound of order $\sqrt{TK\log(T)}$, and this matches the performance of algorithms like UCB for any tuning parameter. A similar result can also be found in (Salomon & Audibert, 2011).

Discussion Propositions 3 and 4, whose proofs are provided in Appendix A, imply that our algorithm APT is a good strategy for solving *at the same time* both problems when μ^* is known. As mentioned previously, this is counter intuitive since one would expect a good strategy for the best arm identification problem to explore significantly more than a good strategy for the cumulative reward maximization problem. To convince oneself, it is sufficient to look at the two-armed case, for which in the fixed budget it is optimal to sample both arms equally, while this strategy has linear regret in the cumulative setting. This intuition is formalized in (Bubeck et al., 2009) where the authors prove that no algorithm can achieve this without additional information. Our results therefore imply that the knowledge of μ^* by the learner is a sufficient information so that Theorem 1 of (Bubeck et al., 2009) does not hold anymore and there exists algorithms that solve both problems at the same time, as APT does.

TopM problem An extension of the best arm identification problem is known as TopM arms identification problem, where one is concerned with identifying the set of the M arms with the highest means (Bubeck et al., 2013b; Gabillon et al., 2012; Kaufmann et al., 2015; Zhou et al., 2014; Chen et al., 2014; Cao et al., 2015). If the learner has some additional information, such as the mean values of the arms with Mth and $(M + 1)$th highest means, then it is straightforward that one can apply our algorithm APT, setting τ in the middle between the Mth and $(M + 1)$th highest means. The set $\widehat{S}_\tau$ would then be returned as the es-
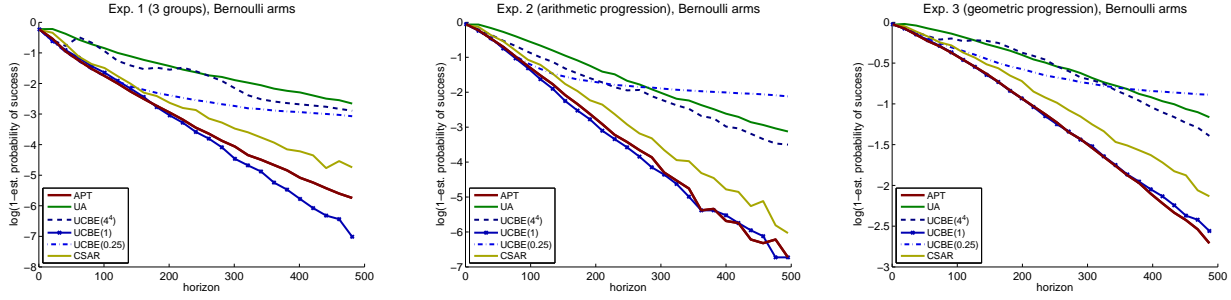
Figure 1. Results of Experiments 1-3 with Bernoulli distributions. The average error of the specified methods is displayed on a logarithmic scale with respect to the horizon.

timated set of M optimal arms. The upper bound and proof for this problem is a direct consequence of Theorem 2, and granted one has such extra-information, outperforms existing results for the fixed budget setting, see (Bubeck et al., 2013b; Kaufmann et al., 2015; Chen et al., 2014; Cao et al., 2015). If the complexity H were also known to the learner, the strategy in (Gabillon et al., 2012) would attain a similar performance.

4. Experiments

We illustrate the performance of algorithm APT in a number of experiments. For comparison, we use the following methods which include the state of the art CSAR algorithm of (Chen et al., 2014) and two minor adaptations of known methods that are also suitable for our problem.

Uniform Allocation (UA): For each $t \in \{1, 2, \ldots, T\}$, we choose $I_t \sim \mathcal{U}_{\mathbb{A}}$. This method is known to be optimal if all arms are equally difficult to classify, that is in our setting, if the quantities $\Delta_i^{\tau,\epsilon}$, $i \in \mathbb{A}$, are very close.

UCB-type algorithm: The algorithm UCBE given and analyzed in (Audibert & Bubeck, 2010) is designed for finding the best arm - and its heuristic is to pull the arm that maximizes a UCB bound - see also (Gabillon et al., 2012) for an adaptation of this algorithm to the general TopM problem. The natural adaptation of the method for our problem corresponds to pulling the arm that minimizes $\widehat{\Delta}_k(t) - \sqrt{\frac{a}{T_k(t)}}$. From the theoretical analysis in the paper (Audibert & Bubeck, 2010; Gabillon et al., 2012), it is not hard to see that setting $a \approx (T - K)/H$ minimizes their upper bound, and that this algorithm attains the same expected loss as ours - but it requires the knowledge of H. In the experiments we choose values $a_i = 4^i \frac{T-K}{H}$, $i \in \{-1, 0, 4\}$, and denote the respective results as UCBE(4^i). The value a_0 can be seen as the optimal choice, while the two other choices give rise to strategies that are sub-optimal because they respectively explore too little or too much.

CSAR: As mentioned before, this method is given in (Chen et al., 2014). In our specific setting, via the shift $\widetilde{\mu}_i = \mu_i - \tau$, the lines 7-17 of the algorithm reduce

to classifying the arm i that maximizes $|\widetilde{\mu}_i|$ based on its current mean. The set A_t corresponds to $\widehat{S}_\tau$ at time t. In fact in our specific setting the CSAR algorithm is a successive reject-type strategy (see (Audibert & Bubeck, 2010) where the arm whose empirical mean is furthest from the threshold is rejected at the end of each phase.

Figure 1 displays the estimated probability of success on a logarithmic scale with respect to the horizon of the six algorithms based on $N = 5000$ simulated games with $\tau = \frac{1}{2}$, $\epsilon = 0.1$, $K = 10$, and $T = 500$.

Experiment 1 (3 groups setting): K Bernoulli arms with means $\mu_{1:3} \equiv 0.1$, $\mu_{4:7} = (0.35, 0.45, 0.55, 0.65)$ and $\mu_{8:10} \equiv 0.9$, which amounts to 2 difficult relevant arms (that is, outside the 2ϵ- band), 2 difficult irrelevant arms and six easy relevant arms.

Experiment 2 (arithmetic progression): K Bernoulli arms with means $\mu_{1:4} = 0.2 + (0:3) \cdot 0.05$, $\mu_5 = 0.45$, $\mu_6 = 0.55$ and $\mu_{7:10} = 0.65 + (0:3) \cdot 0.05$, which amounts to 2 difficult irrelevant arms and eight arms arithmetically progressing away from τ.

Experiment 3 (geometric progression): K Bernoulli arms with means $\mu_{1:4} = 0.4 - 0.2^{1:4}$, $\mu_5 = 0.45$, $\mu_6 = 0.55$ and $\mu_{7:10} = 0.6 + d^{5-(1:4)}$, which amounts to 2 difficult irrelevant arms and eight arms geometrically progressing away from τ.

The experimental results confirm that our algorithm may only be outperformed by methods that have an advantage in the sense that they have access to the underlying problem complexity and, in the case of UCBE(1), an additional optimal parameter choice. In particular, other choices for that parameter lead to significantly less accurate results comparable to the naive strategy UA. These effects are also visible in the further results given in Appendix B.

Conclusion In this paper we proposed a parameter free algorithm based on a new heuristic for the TBP problem in the fixed confidence setting - and we prove that it is optimal which is a kind of result which is highly non trivial for combinatorial pure exploration problems with fixed budget.

References

Alon, Noga, Matias, Yossi, and Szegedy, Mario. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 20–29. ACM, 1996.

Audibert, Jean-Yves and Bubeck, Sébastien. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory*, 2010.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert. Gambling in a Rigged Casino: The Adversarial Multi-Armed Bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pp. 322–331, 1995.

Bubeck, Sebastian, Cesa-Bianchi, Nicolo, and Lugosi, Gábor. Bandits with heavy tail. *Information Theory, IEEE Transactions on*, 59(11):7711–7717, 2013a.

Bubeck, Sébastien, Munos, Rémi, and Stoltz, Gilles. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pp. 23–37. Springer, 2009.

Bubeck, Séebastian, Wang, Tengyao, and Viswanathan, Nitin. Multiple identifications in multi-armed bandits. In *Proceedings of The 30th International Conference on Machine Learning (ICML-13)*, pp. 258–265, 2013b.

Cao, Wei, Li, Jian, Tao, Yufei, and Li, Zhize. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Advances in Neural Information Processing Systems*, pp. 1036–1044, 2015.

Carpentier, Alexandra and Locatelli, Andrea. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory*, 2016.

Catoni, Olivier et al. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pp. 1148–1185. Institut Henri Poincaré, 2012.

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

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

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

Gabillon, Victor, Ghavamzadeh, Mohammad, and Lazaric, Alessandro. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Jamieson, Kevin, Malloy, Matthew, Nowak, Robert, and Bubeck, Sébastien. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

Kalyanakrishnan, Shivaram, Tewari, Ambuj, Auer, Peter, and Stone, Peter. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 655–662, 2012.

Karnin, Zohar, Koren, Tomer, and Somekh, Oren. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1238–1246, 2013.

Kaufmann, Emilie, Cappé, Olivier, and Garivier, Aurélien. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2015.

Mannor, S and Tsitsiklis, J N. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Salomon, Antoine and Audibert, Jean-Yves. Deviations of stochastic bandit regret. In *Algorithmic Learning Theory*, pp. 159–173. Springer, 2011.

Steinwart, Ingo, Hush, Don R, and Scovel, Clint. A classification framework for anomaly detection. In *Journal of Machine Learning Research*, pp. 211–232, 2005.

Streeter, Matthew J and Smith, Stephen F. Selecting among heuristics by solving thresholded k-armed bandit problems. *ICAPS 2006*, pp. 123.

Zhou, Yuan, Chen, Xi, and Li, Jian. Optimal pac multiple arm identification with applications to crowdsourcing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 217–225, 2014.

A. Proofs

A.1. Proof of Theorem 1

Proof. In this proof, we will prove that on at least one instance of the problem, any algorithm makes a mistake of order at least $\exp(-cT/H)$.

Step 0: Setting and notations. Let us consider K real numbers $\Delta_i \geq 0$, and let us set $\tau = 0, \epsilon = 0$. Let us write $\nu_i := \mathcal{N}(\Delta_i, 1)$ for the Gaussian distribution of mean Δ_i and variance 1, and $\nu_i' := \mathcal{N}(-\Delta_i, 1)$ for the Gaussian distribution of mean $-\Delta_i$ and variance 1. Note that this construction is easily generalised to cases where $\tau \neq 0$ or $\epsilon \neq 0$ by translation or careful choice of the Δ_i.

We define the product distributions $\mathcal{B}^i$ where $i \in \{0, ..., K\}$ as $\nu_1^i \otimes ... \otimes \nu_K^i$ where for $k \leq K$, $\nu_k^i := \nu_i \mathbf{1}_{k \neq i} + \nu_i' \mathbf{1}_{k=i}$ is ν_i if $k \neq i$ and ν_i' otherwise. We also extend this notation to $\mathcal{B}^0$, where none of the arms is flipped with respect to the threshold ($\forall k, \nu_k^0 := \nu_i$). It is straightforward that the gap Δ_i of arm i with respect to the threshold $\tau = 0$ does not depend on $\mathcal{B}^i$ and is equal to Δ_i. It follows that all these problems have the same complexity H as defined previously (with $\epsilon = 0$ and $\tau = 0$).

We write for $i \leq K$, $\mathbb{P}_{\mathcal{B}}^i$ for the probability distribution according to all the samples that a strategy could possibly collect up to horizon T, i.e. according to the samples $(X_{k,s})_{k \leq K, s \leq T} \sim (\mathcal{B}^i)^{\otimes T}$. Let $(T_k)_{k \leq K}$ denote the numbers of samples collected by the algorithm on arm k.

Let $k \in \{0, ..., K\}$. Note that

$$\text{KL}_k := \text{KL}(\nu_k', \nu_k) = 2\Delta_k^2,$$

where KL is the Kullback Leibler divergence. Let $T \geq t \geq 0$. We define the quantity:

$$\widehat{\text{KL}}_{k,t} = \frac{1}{t} \sum_{s=1}^{t} \log(\frac{d\nu_k'}{d\nu_k}(X_{k,s})) = -\frac{1}{t} \sum_{s=1}^{t} 2X_{k,s}\Delta_k.$$

Step 1: Concentration of the empirical KL. Let us define the event:

$$\xi = \left\{ \forall k \leq K, \forall t \leq T, |\widehat{\text{KL}}_{k,t} - \text{KL}_k| \leq 4\Delta_k \sqrt{\frac{\log(4(\log(T) + 1)K)}{t}} \right\}.$$

Since $\widehat{\text{KL}}_{k,t} = -\frac{1}{t} \sum_{s=1}^{t} 2X_{k,s}\Delta_k$ and $\text{KL}_k = 2\Delta_k^2$, by Gaussian concentration (a peeling and the maximal martingale inequality), it holds that for any i that $\mathbb{P}_{\mathcal{B}^i}(\xi) \geq 3/4$.

Step 2: A change of measure. We will now use the change of measure introduced previously for a well chosen event $\mathcal{A}$. Namely, we consider $\mathcal{A}_i = \{i \in \widehat{S}_\tau\}$, the event where the algorithm classified arm i as being above the threshold. We have by doing a change of measure between $\mathcal{B}^i$ and $\mathcal{B}^0$ (since they only differ in arm i and only the T_i first samples of arm i by the algorithm):

$$\begin{aligned}
\mathbb{P}_{\mathcal{B}^i}(\mathcal{A}_i) &= \mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}_i} \exp\left(-T_i \widehat{\text{KL}}_{i,T_i} \right) \right] \\
&\geq \mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}_i \cap \xi} \exp\left(-T_i \widehat{\text{KL}}_{i,T_i} \right) \right] \\
&\geq \mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}_i \cap \xi} \exp\left(-2\Delta_i^2 T_i - 4\Delta_i \sqrt{T_i} \sqrt{\log((4\log(T)+1)K)} \right) \right],
\end{aligned}$$

by definition of ξ and KL_i.

Step 3: A union of events. We now consider the event $\mathcal{A} = \bigcap_{i=1}^{K} \mathcal{A}_i$, i.e. the event where all arms are classified as

being above the threshold $\tau = 0$. We have:

$$\max_{i \in \{1,...,K\}} \mathbb{P}_{\mathcal{B}^i}(\mathcal{A}_i) \geq \frac{1}{K} \sum_{i=1}^{K} \mathbb{P}_{\mathcal{B}^i}(\mathcal{A}_i) \tag{6}$$

$$\geq \frac{1}{K} \sum_{i=1}^{K} \mathbb{P}_{\mathcal{B}^i}(\mathcal{A}_i \cap \xi)$$

$$\geq \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}_i \cap \xi} \exp\left(-2T_i \Delta_i^2 - 4\Delta_i \sqrt{T_i}\sqrt{\log(4(\log(T)+1)K))}\right)\right]$$

$$\geq \mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A} \cap \xi} \frac{1}{K} \sum_{i=1}^{K} \exp\left(-3T_i\Delta_i^2 - 4\log(4(\log(T)+1)K))\right)\right]$$

$$\geq \exp\left(-4\log(4(\log(T)+1)K)\right)\mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}\cap\xi}S\right], \tag{7}$$

where the fourth line comes from using $2ab \leq a^2 + b^2$ with $a = \Delta_i\sqrt{T_i}$ and where:

$$S = \frac{1}{K} \sum_{i=1}^{K} \exp\left(-3T_i\Delta_i^2\right).$$

Since $\sum_i T_i = T$ and all T_i are positive, there exists an arm i such that $T_i \leq \frac{T}{H\Delta_i^2}$. This yields:

$$S \geq \frac{1}{K} \exp\left(-\frac{3T}{H}\right) = \exp\left(-\frac{3T}{H} - \log(K)\right).$$

This implies by definition of the risk:

$$\max_{i \in \{0,...,K\}} \mathbb{E}_{\mathcal{B}^i}(\mathcal{L}(T)) \geq \max\left(\max_{i \in \{1,...,K\}} \mathbb{P}_{\mathcal{B}^i}(\mathcal{A}_i), 1 - \mathbb{P}_{\mathcal{B}^0}(\mathcal{A})\right)$$

$$\geq \frac{1}{2} \exp\left(-\frac{3T}{H} - 4\log(4(\log(T)+1)K)\right) - \log(K)\mathbb{E}_{\mathcal{B}^0}\left[\mathbf{1}_{\mathcal{A}\cap\xi}\right] + \frac{1}{2}(1 - \mathbb{P}_{\mathcal{B}^0}(\mathcal{A}))$$

$$= \frac{1}{2} \exp\left(-\frac{3T}{H} - 4\log(4(\log(T)+1)K - \log(K)))\right)\mathbb{P}_{\mathcal{B}^0}\left[\mathcal{A}\cap\xi\right] + \frac{1}{2}(1 - \mathbb{P}_{\mathcal{B}^0}(\mathcal{A}))$$

$$\geq \frac{1}{8} \exp\left(-\frac{3T}{H} - 4\log(4(\log(T)+1)K) - \log(K)\right)$$

$$\geq \exp\left(-\frac{3T}{H} - 4\log(12(\log(T)+1)K)\right),$$

The fourth line comes from $\mathbb{P}(\xi) \geq 3/4$, and we consider two cases $\mathbb{P}_{\mathbb{B}^0}(\mathcal{A}) \geq 1/2$ and $\mathbb{P}_{\mathbb{B}^0}(\mathcal{A}) \leq 1/2$. The first leads directly to the condition as the intersection is at least of probability $1/4$; in the latter case, we have the same bound via

$$\max_{i \in \{0,...,K\}} \mathbb{E}_{\mathcal{B}^i}(\mathcal{L}(T)) \geq \mathbb{E}_{\mathcal{B}^0}(\mathcal{L}(T)) = \mathbb{P}_{\mathcal{B}^0}(\mathcal{A}^C) \geq 1/2.$$

This concludes the proof.

$\square$

A.2. Proof of Theorem 2

Proof. In this proof, we will show that on a well chosen event ξ, we classify correctly the arms which are over $\tau + \epsilon$, and reject the arms that are under $\tau - \epsilon$.

Step 1: A favorable event. Let $\delta = (4\sqrt{2})^{-1}$. Towards this goal, we define the event ξ as follows:

$$\xi = \left\{\forall i \in \mathbb{A}, \forall s \in \{1,...,T\} : |\frac{1}{s}\sum_{t=1}^{s} X_{i,t} - \mu_i| \leq \sqrt{\frac{T\delta^2}{Hs}}\right\}.$$

We know from Sub-Gaussian martingale inequality that for each $i \in \mathbb{A}$ and each $u \in \{0, ..., \lfloor \log(T) \rfloor\}$:

$$\mathbb{P}\left(\exists v \in [2^u, 2^{u+1}], \{|\frac{1}{v}\sum_{t=1}^{v} X_{i,t} - \mu_i| \geq \sqrt{\frac{T\delta^2}{Hv}}\}\right) \leq \exp(-\frac{T\delta^2}{2R^2H}).$$

ξ is the union of these events for all $i \leq K$ and $s \leq \lfloor \log(T) \rfloor$. As there are less than $(\log(T) + 1)K$ such combinations, we can lower-bound its probability of occurrence with a union bound by:

$$\mathbb{P}(\xi) \geq 1 - 2(\log(T) + 1)K \exp(-\frac{T\delta^2}{2R^2H}).$$

Step 2: Characterization of some helpful arm. At time T, we consider an arm k that has been pulled after the initialization phase and such that $T_k(T) - 1 \geq \frac{(T-K)}{H\Delta_k^2}$. We know that such an arm exists otherwise we get:

$$T - K = \sum_{i=1}^{K}(T_i(T) - 1) < \sum_{i=1}^{K}\frac{T-K}{H\Delta_i^2} = T - K,$$

which is a contradiction. Note that since $T \geq 2K$, we have that $T_k(T) - 1 \geq \frac{T}{2H\Delta_k^2}$

We now consider $t \leq T$ the last time that this arm k was pulled. Using $T_k(t) \geq 2$ (by the initialisation of the algorithm), we know that:

$$T_k(t) \geq T_k(T) - 1 \geq \frac{T}{2H\Delta_k^2}. \tag{8}$$

Step 3: Lower bound on the number of pulls of the other arms. On ξ, at time t as we defined previously, we have for every arm i:

$$|\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{T\delta^2}{HT_i(t)}}. \tag{9}$$

From the reverse triangle inequality and Equation (4), we have:

$$
\begin{aligned}
|\hat{\mu}_i(t) - \mu_i| &= |(\hat{\mu}_i(t) - \tau) - (\mu_i - \tau)| \\
&\geq ||\hat{\mu}_i(t) - \tau| - |\mu_i - \tau|| \\
&\geq |(|\hat{\mu}_i(t) - \tau| + \epsilon) - (|\mu_i - \tau| + \epsilon)| \\
&\geq |\widehat{\Delta}_i(t) - \Delta_i|.
\end{aligned}
$$

Combining this with (9) yields the following:

$$\Delta_k - \sqrt{\frac{T\delta^2}{HT_k(t)}} \leq \widehat{\Delta}_k(t) \leq \Delta_k + \sqrt{\frac{T\delta^2}{HT_k(t)}}. \tag{10}$$

By construction, we know that at time t we pulled arm k, which yields for every $i \in \mathbb{A}$:

$$B_k(t) \leq B_i(t). \tag{11}$$

We can lower bound the left-hand side of (11) using (8):

$$\left(\Delta_k - \sqrt{\frac{T\delta^2}{HT_k(t)}}\right)\sqrt{T_k(t)} \leq B_k(t)$$

$$\left(\Delta_k - \sqrt{2}\delta\Delta_k\right)\sqrt{\frac{T}{2H\Delta_k^2}} \leq B_k(t)$$

$$\left(\frac{1}{\sqrt{2}} - \delta\right)\sqrt{\frac{T}{H}} \leq B_k(t), \tag{12}$$

and upper bound the right hand side using (10) by:

$$B_i(t) = \widehat{\Delta}_i \sqrt{T_i(t)}$$

$$\leq \left(\Delta_i + \sqrt{\frac{T\delta^2}{HT_i(t)}}\right)\sqrt{T_i(t)}$$

$$\leq \Delta_i \sqrt{T_i(t)} + \delta\sqrt{\frac{T}{H}}. \tag{13}$$

As both $\widehat{\Delta}_i$ and Δ_i are positive by definition, combining (12) and (13) yields the following lower bound on $T_i(T) \geq T_i(t)$:

$$\left(1 - 2\sqrt{2}\delta\right)^2 \frac{T}{2H\Delta_i^2} \leq T_i(T). \tag{14}$$

Step 4: Conclusion. On ξ, as Δ_i is a positive quantity, combining (9) and (14) yields:

$$\mu_i - \Delta_i \frac{\sqrt{2}\delta}{1 - 2\sqrt{2}\delta} \leq \hat{\mu}_i(T) \leq \mu_i + \Delta_i \frac{\sqrt{2}\delta}{1 - 2\sqrt{2}\delta}, \tag{15}$$

where $\frac{\sqrt{2}\delta}{1-2\sqrt{2}\delta}$ simplifies to $1/2$ for $\delta = (4\sqrt{2})^{-1}$.
For arms such that $\mu_i \geq \tau + \epsilon$, then $\Delta_i = \mu_i - \tau + \epsilon$ and we can rewrite (15):

$$\mu_i - \tau - \frac{1}{2}\Delta_i \leq \hat{\mu}_i(T) - \tau$$

$$(\mu_i - \tau)(1 - \frac{1}{2}) - \frac{\epsilon}{2} \leq \hat{\mu}_i(T) - \tau$$

$$0 \leq \hat{\mu}_i(T) - \tau,$$

where the last line uses $\mu_i \geq \tau + \epsilon$. One can easily check through similar derivations that $\hat{\mu}_i(T) - \tau < 0$ holds for $\mu_i < \tau - \epsilon$. On ξ, arms over $\tau + \epsilon$ are all accepted, and arms under $\tau - \epsilon$ are all rejected, which means the loss suffered by the algorithm is 0. As $1 - \mathbb{P}(\xi) \leq 2(\log(T) + 1)K \exp(-\frac{1}{64R^2}\frac{T}{H})$, this concludes the proof. $\square$

A.3. Proof of Theorem 3

Proof. We will prove that on a well defined event ξ, sub-optimal arms are pulled at most $\frac{T}{2\Delta_k^2 H} - 1$ times, which translates to the best arm being chosen at the end of the horizon as it was pulled more than half of the time.

Step 1: A favorable event. Let $\delta = 1/18$. We define the following events $\forall i \in \mathbb{A}$:

$$\xi_i = \{\forall s \leq T : |\mu^* - \widehat{\mu}_i(s)| \leq \sqrt{\frac{T\delta}{HT_i(s)}}\},$$

We now define ξ as the intersection of these events:

$$\xi = \bigcap_{k \in \mathbb{A}} \xi_k.$$

Using the same Sub-Gaussian martingale inequality as in the proof of Theorem 2, we can lower bound its probability of occurrence with a union bound by:

$$P(\xi) \geq 1 - 2(\log(T) + 1)K \exp(-\frac{T}{36R^2 H})$$

Step 2: The wrong arm at the wrong time. Let us now suppose that a sub-optimal arm k was pulled at least $\frac{T-K}{2\Delta_k^2 H}$ times after the initialization which translates to $T_k(T) - 1 \geq \frac{T-K}{2\Delta_k^2 H}$. Let us now consider the last time $t \leq T$ that this arm was pulled. As it was pulled at time t, the following inequality holds:

$$B_k(t) \leq B_{k^*}(t). \tag{16}$$

On ξ, we can now lower bound the left hand side by:

$$(\Delta_k - \sqrt{\frac{T\delta}{HT_k(t)}})\sqrt{T_k(t)} \le B_k(t)$$

$$\Delta_k \sqrt{T_k(t)} - \sqrt{\frac{T\delta}{H}} \le B_k(t), \tag{17}$$

We also upper bound the right hand side of (16) by:

$$B_{k^*}(t) \le \sqrt{\frac{T\delta}{H}}. \tag{18}$$

Combining both bounds (17) and (18) with (16), as well as rearranging the terms yields:

$$\Delta_k \sqrt{T_k(t)} \le 2\sqrt{\frac{T\delta}{H}}$$

$$T_k(t)\Delta_k^2 \le \frac{4T\delta}{H}. \tag{19}$$

Using $T_k(t) \ge T_k(T) - 1 \ge \frac{T-K}{2\Delta_k^2 H}$ as well as $T \ge 2K$, we have

$$T_k(t) \ge \frac{T}{4\Delta_k^2 H}. \tag{20}$$

Plugging this in (19) brings the following condition:

$$\frac{T}{4\Delta_k^2 H}\Delta_k^2 \le \frac{4T\delta}{H}. \tag{21}$$

which directly reduces to $\delta \ge 1/16$, which is a contradiction as we have set $\delta = 1/18$.

As we have proved that for any sub-optimal arm $i \ne k^*$ it satisfies $T_i(T) < \frac{T}{2\Delta_i^2 H}$, summing for all arms yields:

$$T - T_{k^*}(T) = \sum_{i \ne k^*} T_i(T)$$

$$< \frac{T}{2H} \sum_{i \ne k^*} \frac{1}{\Delta_i^2} = \frac{T}{2}. \tag{22}$$

We conclude by observing that $T_{k^*}(T) > T/2$, and as such will be chosen by the algorithm at the end as being the best arm. $\square$

A.4. Proof of Theorem 4

Proof. In this proof we will show that with high probability the sub-optimal arms have been pulled at most at a logarithmic rate, and will then bound the expectation of the number of pulls of these arms.

Step 1: A favorable event. We define the following events $\forall s \le T$:

$$\xi_{k^*,s} = \{\mu^* - \hat{\mu}_{k^*}(s) \le R\sqrt{\frac{\log(T)\delta}{T_{k^*}(s)}}\},$$

as well as for all arms $i \ne k^*$:

$$\xi_{i,s} = \{\hat{\mu}_k(s) - \mu_k \le R\sqrt{\frac{\log(T)\delta}{T_{k^i}(s)}}\}.$$

By Hoeffding's inequality, the complimentary $\bar{\xi}_k$ of each of these events has probability at most $T^{-2\delta}$.

We now consider ξ the intersection of these events for all $k \in \mathbb{A}$. By a union bound, as there are T such events for each arm, we have:

$$\mathbb{P}(\xi) \geq 1 - \frac{K}{T^{2\delta-1}}. \tag{23}$$

We also have:

$$\mathbb{P}(\bar{\xi}) \leq \frac{K}{T^{2\delta-1}}. \tag{24}$$

We will now prove a bound on the number of pulls on ξ.

Step 2: Bound on pulls of sub-optimal arms. We now consider the last time t that arm $k \neq k^*$ was pulled, under the assumption that it was pulled at least once after the initialization. The decision rule of the algorithm yields:

$$B_k(t) \leq B_{k^*}(t). \tag{25}$$

On ξ, we can now lower-bound the left-side and upper-bound the right hand side, which yields:

$$(\Delta_k - R\sqrt{\frac{\log(T)\delta}{T_k(t)}})\sqrt{T_k(t)} \leq R\sqrt{\frac{\log(T)\delta}{T_{k^*}(t)}}\sqrt{T_{k^*}(t)}, \tag{26}$$

which can be rearranged as such:

$$\Delta_k\sqrt{T_k(t)} \leq 2R\sqrt{\log(T)\delta}, \tag{27}$$

and the following bound on $T_k(T)$:

$$T_k(T) \leq \frac{4R^2\log(T)\delta}{\Delta_k^2} + 1. \tag{28}$$

Note that we here make the assumption that the arm was pulled at least once by the algorithm after the initialization. If it has only been pulled during the initialization, the bound still trivially holds as we have at least one pull.

Step 3: Conclusion. We can thus upper-bound the expectation of $T_k(t)$, as when ξ does not hold we get at most T pulls:

$$\mathbb{E}[T_k(T)] \leq \frac{4R^2\log(T)\delta}{\Delta_k^2} + 1 + \frac{K}{T^{2\delta-2}}, \tag{29}$$

and we get the following bound on the pseudo-regret when ξ holds:

$$\bar{R}_T \leq \sum_{k \neq k^*} \frac{4R^2\log(T)\delta}{\Delta_k} + \Delta_k(1 + \frac{K}{T^{2\delta-2}}). \tag{30}$$

Plugging $\delta = 1$ yields:

$$\bar{R}_T \leq \sum_{k \neq k^*} \frac{4R^2\log(T)}{\Delta_k} + \Delta_k(1 + K), \tag{31}$$

and we recover the classical bound of the UCB1 algorithm. $\qquad\square$

B. Further Experimental Results

We now also provide simulation results for our three settings in the case of Gaussian arms with means μ_i and variances $\sigma_i^2 = 0.25$. Again, only the correctly tuned UCBE- algorithm outperforms APT.
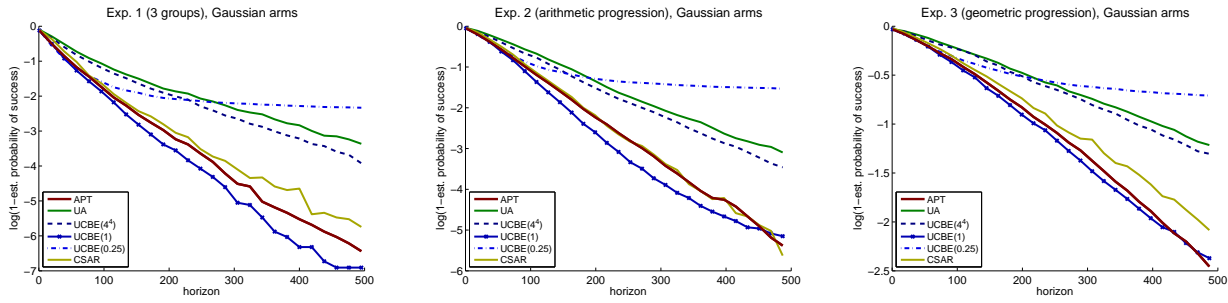

Figure 2. Results of Experiments 1-3 with Gaussian distributions. The average error of the specified methods is displayed on a logarithmic scale with respect to the horizon.