Structured Pure Exploration Bandit Problems and Extensions

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von M.Sci. James Cheshire

geb. am 5. April 1995 in Stourbridge, UK

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. Alexandra Carpentier
Prof. Dr. Christophe Giraud

eingereicht am: 24. Mai 2022

Verteidigung am: 26. Juli 2022

Declaration of Authorship

I, James Cheshire, declare that I produced this thesis without prohibited assistance and that all sources of information that were used in producing this thesis, including my own publications, have been clearly marked and referenced.

In particular I have not wilfully:

- Fabricated data or ignored or removed undesired results.
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data.
- Plagiarised data or publications or presented them in a disorted way.

I know that violations of copyright may lead to injunction and damage claims from the author or prosecution by the law enforcement authorities.

This work has not previously been submitted as a doctoral thesis in the same or a similar form in Germany or in any other country. It has not previously been published as a whole.

Declaration of any Criminal Conviction

I hereby declare that I have not been found guilty of scientific and/or academic misconduct.

Signed:

Date: 26/7/22

Otto von Guericke Universität Magdeburg

Doctoral Thesis

Structured Pure Exploration Bandit Problems and Extensions

Author:
James Cheshire

Supervisor:
Dr. Alexandra Carpentier
Dr. Sebastian Sager

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor rerum naturalium

in the

Institut für Mathematische Stochastik
Fakultät für Mathematik

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Abstract

Institut für Mathematische Stochastik
Fakultät für Mathematik

Structured Pure Exploration Bandit Problems and Extensions

by James Cheshire

The subject of this thesis is a study of several multi armed bandit problems (MAB), with a focus on structured pure exploration bandit problems. We begin with an extensive overview of the MAB literature, with specific weight given to classical techniques and results, for pure exploration bandit problems. Over the course of our analysis we will explore several novel extensions to the MAB. Our first contribution will be to classify the minimax rate for the Thresholding Bandit Problem (TBP). We will then go on to consider the TBP under several shape constraints and again classify the minimax rate in each of these cases. Our second contribution is to study the shape constrained TBP in a problem dependent setting. For the TBP, under both a monotone and concave constraint, we provide problem dependent upper and lower bounds, matching up to log terms. Our third contribution is to consider a potentially infinite armed formulation of the MAB, where a proportion of the arms are optimal. In this setting we provide problem dependent upper and lower bounds, matching up to log terms, for both cumulative regret and best arm identification.

Das Gegenstand dieser Dissertation ist die Untersuchung von mehreren Multi Armed Bandit-Probleme (MAB) mit einem Fokus auf Structured Pure Exploration Bandit-Probleme. Wir fangen mit einer umfangreichen Übersicht der MAB Literatur an, wobei wir in die klassischen Techniken und Ergebnisse für Pure Exploration Bandit-Probleme tiefer eingehen. Im Laufe unserer Analyse werden wir mehrere originelle Erweiterungen zur MAB untersuchen. Unser erstes Beitrag wird das Klassifizieren der Minimax-Raten für das Thresholding Bandit Problem (TBP) sein. Anschließend werden wir das TBP unter einigen Formbeschränkungen betrachten und die Minimax-Rate in jedem dieser Fälle klassifizieren. Unser zweites Beitrag wird die Untersuchung des formbeschränkten TBPs in einer problemabhängigen Umgebung sein. Für das TBP, unter sowohl einem monotonen, als auch konkaven Zwang, bieten wir problemabhängige Ober- und Untergrenzen an, die bis auf die Log-Terme übereinstimmt. Unser dritter Beitrag wird die Betrachtung einer potentiellen Infinite Armed Formulierung des MAB sein, wobei ein Anteil der Arme optimal sind. In dieser Umgebung schaffen wir problemabhängige Ober- und Untergrenzen, die für beide Cumulative Regret und Best Arm Identification bis auf die Log-Terme übereinstimmt. Das letzte Kapitel von dieser Dissertation widmet sich einer laufenden Arbeit, der Cluster-Identifikation mit Bandit Rückmeldung.

Contents

Chapter 1

Introduction

1.1 Stochastic multi armed bandit problems (MAB)

First introduced in the context of clinical trials, [81], in the stochastic multi armed bandit problem, (MAB), a learner is faced with many actions or “arms". At each time step they must choose a single action from which they observe a, typically noisy, reward. A classical objective of the learner is then to maximise their expected cumulative reward, see [81]. With this objective the learner wishes to choose arms with high expected pay off as much as possible. However, the expected rewards are unknown to the learner - leading to the exploration vs exploitation trade off.

The classical multi armed bandit problem is as follows. The learner is presented with an arm set $\mathcal{A}$ where each arm, $a \in \mathcal{A}$ follow a distribution ν_a with mean μ_a. The learner has a budget $T > 0$ and sequentially samples the arms for a total T times. At time $t < T$ let a_t be the arm chosen, the learner then receives a reward $Y_t \sim \nu_{a_t}$ conditionally independent of the past. We term the policy of the learner a sampling strategy, that at each time step maps the past observations, both previous rewards and arm choices, plus potentially some external randomness to the next choice of arm.

As mentioned the MAB was first considered in the context of maximising expected cumulative reward,

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] .$$

Clearly, the optimal policy would be to always pick the arm with highest expected reward. With this in mind, for an arm set $\mathcal{A}$ we define

$$\mu^* = \max_{a \in \mathcal{A}}(\mu_a) ,$$

as the maximum value attained by the means of the arms. An arm is said to be optimal, if its mean is equal to μ^*. We now define the cumulative regret of the learner as the difference between the learner's policy and the optimal policy,

$$R(T) := T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] . \tag{1.1}$$

Now the learner is faced with the “exploration vs exploitation" trade off, they wish to spend most of their budget on arms that are optimal - or at least near optimal, however the means of the arms are unknown. Therefore the learner must explore the arm set, sampling many arms to find ones with a high expected reward. There are also such "pure exploration bandit problems", here the learner does not care about their cumulative reward and only wishes to uncover some hidden property of the arms. Perhaps the simplest such setting is that of best arm identification (BAI). That is, at the end of T rounds the learner must output a prediction $\hat{a}$ of an optimal arm. One can then evaluate their performance on their expected simple regret,

$$r(T) := \mathbb{E}[\mu^* - \mu_{\hat{a}}] \ , \tag{1.2}$$

that is, the expected difference in the mean of the predicted arm to the true optimal. Or their probability of error,

$$e(T) := \mathbb{P}(\mu_{\hat{a}} \neq \mu^*) \ . \tag{1.3}$$

Over the course of this thesis we will cover both pure exploration MAB, as well as the more classical cumulative regret, in a variety of settings.

1.1.1 Preliminary notation and terminology

For simplicity, in the entirety of Section 2.1 we assume that all arms $a \in \mathcal{A}$ are distributed with support bounded on $[0, 1]$. Almost exclusively our results will extend to, and maintain optimality under, the assumption of sub Gaussian distributions with unbounded support. The single instance where this in not the case is clearly highlighted.

With the exception of Sections 1.8 and 1.7, for the entirety of the introduction to this thesis, we will be restricted to the K armed bandit setting, that is $\mathcal{A}$ is finite and $|\mathcal{A}| = K$. In this case we denote $\mathcal{B}$ the set of all possible bandit problems with K arms, where the distribution of each arm has bounded support on $[0, 1]$. We also define the set of all possible policies, $\mathcal{C}$. We suppress dependency of $\mathcal{B}$ and $\mathcal{C}$ on K in our notation. The regret of the learner is some function ϕ,

$$\phi : (\mathcal{B}, \mathcal{C}, T) \to \mathbb{R} \ .$$

We term the regret of a policy π on a problem $\nu \in \mathcal{B}$, for a regret function ϕ, with budget T, as $\phi_\pi^\nu(T)$. As an example, consider the cumulative regret, Equation (1.1), where for a policy $\pi \in \mathcal{C}$, problem $\nu \in \mathcal{B}$ and budget T, the cumulative regret of policy π, with budget T, on on problem ν, is given as $R_\pi^\nu(T)$. In some cases, when it is obvious to the reader, we may drop the dependency on π, ν in the notation.

For a bandit problem ν and policy π we will denote the distribution on the canonical bandit model, see [63](Chapter 4.6) as $\mathbb{P}_{\nu,\pi}$. Essentially this can be seen as the distribution on all samples gathered by a policy π on bandit problem ν. We similarly define $\mathbb{E}_{\nu,\pi}$ as the expectation on the canonical bandit model. For two distributions P, Q, on the same measurable space $(\Omega, \mathcal{F})$, for P absolutely continuous to Q, we write $P \ll Q$. For a integer $K \in \mathbb{N}$, we define,

$$[K] := \{1, ..., K\} \ .$$

For time $t > 0$ and arm $a \in \mathcal{A}$, let $N_a(t)$ be the number of times the learner has pulled arm a up to, but not including, time t, that is,

$$N_a(t) := \sum_{s<t} \mathbb{1}(a_s = a) \ .$$

One of the defining characteristics of the MAB is that the learner does not have access to the true means of the arms. We also denote for some time t and arm a, the empirical mean of arm a at time t as,

$$\widehat{\mu}_{a,t} := \frac{1}{N_a(t)} \sum_{s<t:a_s=a} Y_s \; .$$

Where obvious, we will occasionally drop the dependency on t in the notation and simply denote the empirical mean of arm a as $\widehat{\mu}_a$.

1.1.2 Minimax rates and problem independent vs dependent

One way to judge a policy π, on $\mathcal{B}$, is by its worst case performance, i.e.

$$\sup_{\nu \in \mathcal{B}} \phi_\pi^\nu(T) \; . \tag{1.4}$$

A benchmark to measure (1.4) against is the minimax rate, which we define as follows,

$$\phi_T^*(\mathcal{B}) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}} \phi_\pi^\nu(T) \; ,$$

that is the "best worst case performance" across all possible policies $\pi \in \mathcal{C}$ on the class of problems $\mathcal{B}$. By considering regret on the entire class of problems, $\mathcal{B}$, our results are termed to be *problem independent* in the sense that they hold in the worst case, across all problems. However, if we do not wish to default to worst case performance, we can instead consider a restricted class, that only contains problems of a certain difficulty. To perform well in the setting of cumulative regret, the learner must be able to distinguish arms as optimal or sub optimal; the difficulty of this problem depends upon the distance between the means of the arms. With this in mind, for some $a \in \mathcal{A}$ let us define the gap to an optimal arm, Δ_a, of the ath arm as,

$$\Delta_a := \mu^* - \mu_a \; .$$

We then say, that the sequence of gaps $(\Delta_a)_{a\in\mathcal{A}}$ classifies the complexity of the problem. Instead of evaluating the learner on their worst case performance, across all possible problems, we can classify their regret given a specific vector of gaps. That is, given a sequence of gaps $\bar{\Delta}$, we can consider the rate on the restricted set of bandit problems,

$$\mathcal{B}_{\bar{\Delta}} := \{\nu \in \mathcal{B} : \forall a : \mu_a \neq \mu^*, \mu^* - \mu_a \geq \bar{\Delta}_a\} \; .$$

In this case, our results will be dependent upon the gaps $(\Delta_a)_{a\in\mathcal{A}}$ and are termed *problem dependent* results. For a given $\bar{\Delta}$ and regret function ϕ, we can again define the minimax rate on this restricted set of problems as,

$$\phi_T^*(\mathcal{B}_{\bar{\Delta}}) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}_{\bar{\Delta}}} \phi_\pi^\nu(T) \; .$$

During this thesis we will present results from both the problem dependent and problem independent perspective.

1.1.3 Fixed budget vs fixed confidence

So far we have introduced the MAB in terms of a fixed budget, that is the learner provides a policy π which runs for a total of T time steps, were their goal is to minimise their regret, $\phi_\pi(T)$. This is in contrast to the *fixed confidence* setting. Here there is no fixed T, instead, along with a policy π, the learner must choose a stopping time τ. That is, their policy π will run for τ time steps. The stopping time τ may be dependent on both observed rewards and external randomness. Fix a problem class $\mathcal{B}$

and regret function ϕ, where the regret of some policy π, with stopping time τ, on some problem $\nu \in \mathcal{B}$, is given as $\phi_\pi^\nu(\tau)$. For some confidence level $\delta > 0$ and margin of error $\varepsilon > 0$, a policy π, with stopping rule τ, is said to be PAC(δ, ε) (probably approximately correct), under regret function ϕ on the class of problems $\mathcal{B}$, if,

$$\forall \nu \in \mathcal{B}, \mathbb{P}_{\nu,\pi}[\phi_\pi^\nu(\tau) \leq \varepsilon] \leq \delta \ . \tag{1.5}$$

The goal of the learner is to then obtain a PAC(δ, ε) policy such that the expected stopping time in the worst case,

$$\sup_{\nu \in \mathcal{B}} \mathbb{E}_\nu[\tau] \ ,$$

is minimised. In certain settings, e.g. when the regret is given as some probability of error, we will take $\varepsilon = 0$. In such cases, dependency on ε may be dropped in the notation and we refer to PAC(δ) policies.

While the majority of novel results presented in this thesis will be in the fixed budget setting, much of the related literature is in the fixed confidence setting.

1.2 The 2 armed bandit

In this section we will first present a toy problem, the two armed bandit. Our goal is to build the reader's intuition and also introduce some classical proof techniques. In the 2 armed bandit, we restrict the arm set to have cardinality 2, that is, $\mathcal{A} = \{1, 2\}$. We will study this problem under both probability of error for BAI and cumulative regret. For the two armed case, there is a single gap $\Delta = |\mu_1 - \mu_2|$. If $\mu_1 = \mu_2$, in the 2 armed case, then both BAI and minimisation of cumulative regret are trivial, therefore, for the remainder of this section, we assume $|\mu_1 - \mu_2| > 0$.

1.2.1 Probability of error in BAI for fixed budget 2 armed bandit

To minimise probability of error for BAI in the two armed bandit, we utilise perhaps the most simplistic algorithm in bandit theory, Uniform Allocation. That is, we pull both arms a total of $T/2$ times and then output the arm with highest empirical mean.

```
for $a \in \{1, 2\}$ do
    Pull arm a, $T/2$ times, let $\widehat{\mu}_a$ denote sample mean
end
Output: arm in $\arg\max_{a \in \{1,2\}}(\widehat{\mu}_a)$
```

Algorithm 1: Uniform allocation

Upper bound on regret of Uniform Allocation To upper bound the probability of error of Uniform Allocation, we recall a classical result, Hoeffding's inequality.

Theorem 1. *Let $X_1, ..., X_n$ be independent random variables all supported on the interval $[b, d]$ for some $b, d \in \mathbb{R}$ with $b < d$. Let $\sum_{i=1}^n X_i = S_n$, then for all $s > 0$,*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq s) \leq 2 \exp\left(-\frac{2s^2}{n(d-b)^2}\right) \ .$$

Concentration inequalities, such as the above, are very useful in bandits, as illustrated by the proof of the following theorem.

Theorem 2 (Specific version of Theorem 2 [2]). *Let $\Delta > 0$ and assume $T > 2$, $K = 2$, on all bandit problems $\nu \in \mathcal{B}_\Delta$,* `Uniform Allocation` *will satisfy,*

$$e^{\nu}_{\texttt{Uniform Allocation}}(T) \leq \exp(-T\Delta^2/4) \ .$$

Proof. Without loss of generality, assume that $\mu_1 > \mu_2$. Firstly note that,

$$e(T) \leq \mathbb{P}(\widehat{\mu}_1 \leq \widehat{\mu}_2) \ . \tag{1.6}$$

Now define the event,

$$\xi := \{\widehat{\mu}_1 > \widehat{\mu}_2\} \ .$$

Via Hoeffding's we have that,

$$\mathbb{P}(\widehat{\mu}_1 - \widehat{\mu}_2 \leq 0) \leq \exp(-T\Delta^2/4) \ ,$$

and thus,

$$\mathbb{P}(\xi) \geq 1 - \exp(-T\Delta^2/4) \ . \tag{1.7}$$

As under event ξ, the algorithm recommends the correct arm, the proof follows. □

Lower bound for BAI in the two armed case When constructing lower bounds, we will typically identify pairs or families of bandit problems, that are hard for the learner to distinguish between. An important quantity of interest will be the Kullback Leibler divergence. For two distributions Q, P, with $P \ll Q$, on the same measurable space $(\Omega, \mathcal{F})$ we denote the Kullback Leibler divergence as,

$$\mathrm{KL}(P, Q) = \int_\Omega \log\left(\frac{dP(x)}{dQ(x)}\right) dP(x) \ .$$

In the case where P is not absolutely continuous to Q then we let $\mathrm{KL}(P, Q) = \infty$, however, for the entirety of this thesis, we will only consider the KL divergence on distributions absolutely continuous to one another. To prove our lower bounds we will use information theoretic results, such as the Bretagnolle-Huber's Inequality (see Theorem 14.2 [63]).

Theorem 3 (Bretagnolle-Huber's Inequality). *For any two probability measures P, Q on a measurable space $(\Omega, \mathcal{F})$, and an arbitrary event $A \in \mathcal{F}$,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-\mathrm{KL}(P, Q)) \ .$$

Thus, for two bandit problems, ν, ν' and policy π, of particular interest to us will be the Kullback Leibler distance, $\mathrm{KL}(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi})$. We remind the reader that in this section we restrict to $K = 2$, however, as it will be of use later we generalise to the K armed bandit momentarily. Let

$$\nu = \{P_1, ..., P_K\} \ ,$$

and

$$\nu' = \{Q_1, ..., Q_K\} \ ,$$

via application of the chain rule, see [63](Chapter 15.1), one can show,

$$\mathrm{KL}(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) = \sum_{a=1}^{K} \mathbb{E}_{\nu,\pi}[N_a(T)] \mathrm{KL}(P_a, Q_a) \ . \tag{1.8}$$

In our restricted setting of the 2 armed bandit, the above simplifies to,

$$\mathrm{KL}(\mathbb{P}_{\nu,\pi}, \mathbb{P}_{\nu',\pi}) = \mathbb{E}_{\nu,\pi}[N_1(T)]\mathrm{KL}(P_1, Q_1) + \mathbb{E}_{\nu,\pi}[N_2(T)]\mathrm{KL}(P_2, Q_2) \ .$$

With Theorem 3 and equation (1.8) in mind, we are now ready to construct our lower bound.

Theorem 4. *Assume $T > 2$, $K = 2$ and $0 < \Delta < 1/4$, for any policy π there exists a problem $\nu \in \mathcal{B}_\Delta$, such that on problem ν the policy π suffers the following probability of error,*

$$e_\pi^\nu(T) \geq \frac{1}{4}\exp(-8T\Delta^2) \ .$$

Proof. For $x \in [0, 1]$, let $\mathcal{B}er(x)$ denote the Bernoulli distribution with mean x. Define the following two bandit problems, $\nu^{<1>}, \nu^{<2>}$, with,

$$\nu_1^{<1>} = \mathcal{B}er(1/2 + \Delta),\ \nu_2^{<1>} = \mathcal{B}er(1/2) \ ,$$

and

$$\nu_1^{<2>} = \mathcal{B}er(1/2),\ \nu_2^{<2>} = \mathcal{B}er(1/2 + \Delta) \ ,$$

Under problem $\nu^{<1>}$ arm 1 is optimal while under problem $\nu^{<2>}$ arm 2 is optimal. Now,

$$\mathrm{KL}(\mathcal{B}er(1/2 + \Delta), \mathcal{B}er(1/2)) \leq 4\Delta^2 \ ,$$

and

$$\mathrm{KL}(\mathcal{B}er(1/2), \mathcal{B}er(1/2 + \Delta)) \leq 8\Delta^2 \ ,$$

thus via (1.8),

$$\mathrm{KL}\big(\mathbb{P}_{\nu^{<1>},\pi}, \mathbb{P}_{\nu^{<2>},\pi}\big) \leq 8T\Delta^2 \ .$$

Now we define the event,

$$\xi := \{\hat{a} = 2\} \ ,$$

and then by application of the Bretagnolle-Huber Inequality, Theorem 3, on event ξ,

$$\mathbb{P}_{\nu^{<1>},\pi}(\xi) + \mathbb{P}_{\nu^{<2>}\pi}(\xi^c) \geq \frac{1}{2}\exp(-8T\Delta^2), \ ,$$

and thus,

$$\max_{j \in \{1,2\}} (e_\pi^{\upsilon^{<j>}}(T)) \geq \frac{1}{4}\exp(-8T\Delta^2) \ .$$

□

As we can see, for the 2 armed bandit, the upper bound of Theorem 2 matches our lower bound of Theorem 4, up to multiplicative constants, both in and outside the exponential. Later in this thesis, we will showcase the more general upper and lower bounds for the K-armed bandit, see Section 1.4.

1.2.2 Cummulative regret for 2 armed bandit

In the setting of cumulative regret, the learner is faced with the exploration vs exploitation trade-off. Perhaps the most naive way of approaching this, is to divide ones budget into two distinct phases. During the first phase, the learner uses a fixed amount of their budget to sample all of the arms evenly. The learner identifies the best performing arm from the first phase and then uses the remainder of their budget to sample said arm exclusively. This is the Explore Then Commit (ETC) algorithm, explicitly described as follows.

```
Input: exploration parameter $m < T/2$
for $a \in \{1, 2\}$ do
    Pull $m$ times arm $a$, let $\widehat{\mu}_a$ denote sample mean
end
Pull an arm $\arg\max_{a \in \{1,2\}}(\widehat{\mu}_a)$ for final $T - 2m$ rounds
```

Algorithm 2: ETC Algorithm for 2 armed bandit

Upper bound on the regret of ETC

Theorem 5. *Let $\Delta > 0$ and assume $T \geq \frac{1}{\Delta^2}$, $K = 2$, running the ETC algorithm with $m = (4\log(T\Delta^2)/\Delta) \wedge (T/2)$ will satisfy,*

$$R(T) \leq \frac{\log(T\Delta^2)}{\Delta} + \frac{4}{\Delta} ,$$

on all bandit problems $\nu \in \mathcal{B}_\Delta$.

Proof. In the case where $T/2 < 4\log(T\Delta^2)/\Delta$ the proof is immediate, thus, assume $T/2 \geq 4\log(T\Delta^2)/\Delta$. Without loss of generality, assume $\mu_1 > \mu_2$ and define the event,

$$\xi := \{\widehat{\mu}_1 > \widehat{\mu}_2\} .$$

Via application of Hoeffding's, as in the proof of Theorem 2, we have that

$$\mathbb{P}(\xi) \geq 1 - \exp(-m\Delta^2/4) .$$

On event ξ we have that, after $2m$ rounds, the algorithm will pull the optimal arm for the remainder of the budget. Therefore,

$$R(T) \leq T\Delta \exp(-m\Delta^2/4) + m\Delta .$$

Choosing $m = 4\log(T\Delta^2)/\Delta^2$ will minimise the above and complete the proof. □

Remark 1. The assumption $T \geq \frac{1}{\Delta^2}$, of Theorem 5, is reasonable, as in the case where $T < \frac{1}{\Delta^2}$, the trivial upper bound on the regret, $R(T) \leq T\Delta$, will match the lower bound, up to a constant, as shown in Theorem 6.

Lower bound on cumulative regret for 2 armed case

Theorem 6. *Let $0 < \Delta < 1/2$ and $K = 2$. Under the assumption, $T > \frac{1}{\Delta^2}$, for any policy π there exists a problem $\nu \in \mathcal{B}_\Delta$, such that on problem ν, the policy π suffers the following regret,*

$$R^{\nu}_{\pi}(T) \geq \left(\frac{c\log(T\Delta^2)}{\Delta}\right) \wedge (T\Delta) ,$$

where $c > 0$ is an absolute constant.

Under the assumption $T \leq \frac{1}{\Delta^2}$, for any policy π there exists a problem $\nu' \in \mathcal{B}_\Delta$, such that on problem ν', the policy π suffers the following regret,

$$R^{\nu'}_{\pi}(T) \geq c'T\Delta ,$$

where $c' > 0$ is an absolute constant.

Proof. Firstly we assume $T > \frac{1}{\Delta^2}$.

Step 1: $T > \frac{1}{\Delta^2}$ We define the following two bandit problems, $\nu^{<1>}, \nu^{<2>}$, with,

$$\nu_1^{<1>} = \mathcal{B}er(1/2 + 2\Delta),\ \nu_2^{<1>} = \mathcal{B}er(1/2 + \Delta)\ ,$$

and

$$\nu_1^{<2>} = \mathcal{B}er(1/2),\ \nu_2^{<2>} = \mathcal{B}er(1/2 + \Delta)\ .$$

Now via equation (1.8) and the fact that $\text{KL}(\mathcal{B}er(1/2 + 2\Delta), \mathcal{B}er(1/2)) \leq 16\Delta^2$ we have that,

$$\text{KL}\big(\mathbb{P}_{\nu^{<1>},\pi}, \mathbb{P}_{\nu^{<2>},\pi}\big) \leq 16\mathbb{E}[N_1(T)]\Delta^2\ .$$

Now firstly assume $\mathbb{E}_{\nu^{<1>}}[N_1(T)] \geq \frac{\log(T\Delta^2)}{32\Delta^2}$, under this assumption we have that,

$$R_{\nu^{<1>}}(T) \geq \frac{\log(T\Delta^2)}{32\Delta}\ ,$$

and the proof is complete. Otherwise we have $\mathbb{E}_{\nu^{<1>}}[N_1(T)] < \frac{\log(T\Delta^2)}{32\Delta^2}$ and thus,

$$\text{KL}\big(\mathbb{P}_{\nu^{<1>},\pi}, \mathbb{P}_{\nu^{<2>},\pi}\big) < \log(T\Delta^2)/2\ .$$

Define the event ξ

$$\xi := \{N_1(T) > T/2\}\ .$$

Note that on ξ, under problem $\nu^{<2>}$ the learner will suffer regret $T\Delta/2$ and will also suffer regret $T\Delta/2$ on event ξ^c, under problem $\nu^{<1>}$. Then by application of the Bretagnolle-Huber Inequality, Theorem 3, on event ξ,

$$\mathbb{P}_{\nu^{<2>},\pi}(\xi) + \mathbb{P}_{\nu^{<1>},\pi}(\xi^c) \geq \frac{1}{2}\exp(-\log(T\Delta^2)/2) = \frac{1}{2\Delta\sqrt{T}}\ ,$$

and thus, there exists $\nu \in \{\nu^{<1>}, \nu^{<2>}\}$ such that on problem ν, policy π suffers the following regret,

$$R(T) \geq \sqrt{T}/8\ .$$

As we assume $T > \frac{1}{\Delta^2}$, we have $\sqrt{T}/8 > \frac{\log(T\Delta^2)}{16\Delta}$ and the proof also follows.

Step 2: $T \leq \frac{1}{\Delta^2}$ Under this assumption we have that,

$$\text{KL}\big(\mathbb{P}_{\nu^{<1>},\pi}, \mathbb{P}_{\nu^{<2>},\pi}\big) \leq 16\ ,$$

and thus, defining the event ξ as in Step 1,

$$\mathbb{P}_{\nu^{<2>},\pi}(\xi) + \mathbb{P}_{\nu^{<1>},\pi}(\xi^c) \geq \frac{1}{2}\exp(-16)\ ,$$

and thus, there exists $\nu \in \{\nu^{<1>}, \nu^{<2>}\}$ such that on problem ν, policy π suffers the following regret,

$$R(T) \geq \frac{1}{4}\exp(-16)T\Delta\ .$$

□

When we compare to the lower bound of Theorem 6, the ETC appears near optimal. However, the issue with the ETC is that the learner must specify the length of the exploration phase m and to do this optimally, requires knowledge of Δ. In the case where the gaps are large, the arms are easy to distinguish from one another and one would want a relatively short exploration phase. When the gaps are small, the opposite holds. Therefore, for the two armed bandit, for $0 < \Delta < 1/4$, $T > \frac{1}{\Delta^2}$, we have classified the minimax rate,

$$R_T^*(\mathcal{B}_\Delta) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}_\Delta} R_\pi^\nu(T) \ ,$$

as of the order,

$$\log(T\Delta^2)/\Delta \ ,$$

only in the case where Δ *is known*. Knowing Δ is an unreasonable assumption in most practical applications. In the following section we will discuss an algorithm, the Upper Confidence Bound (UCB) algorithm, that overcomes this obstacle.

1.3 Cumulative regret for the K-armed bandit

In this section, we consider cumulative regret under the more general setting where $|\mathcal{A}| = K$. As seen in the 2 armed case, a naive approach to balance exploration and exploitation is to fix a proportion of the budget to be used for exploration. However, to maximise performance of such strategies, one requires knowledge of $(\Delta_a)_{a \in \mathcal{A}}$, in order to tune the length of the exploration phase. This is an unreasonable assumption in most practical applications.

1.3.1 UCB algorithm for the K-armed bandit

UCB stands for "upper confidence bound". Assume at time t we have n iid samples of some arm a, via Hoeffding's we can create a confidence bound around the empirical mean as follows,

$$\mu_a \in \left[\hat{\mu}_{a,t} - \sqrt{\frac{\log(2/\delta)}{2n}}, \hat{\mu}_{a,t} + \sqrt{\frac{\log(2/\delta)}{2n}} \right] \ ,$$

w.p. greater than $1 - \delta$. The principle of the UCB algorithm is "optimism in the face of uncertainty", that is, when choosing which arm to sample next, the learner picks the arm with the highest upper confidence bound. By doing this, we give more weight to arms that have been pulled less. Taking $\delta = 1/T^2$, we define the UCB index of arm a at time t as in [4],

$$\mathrm{UCB}(a,t) := \hat{\mu}_{a,t} + \sqrt{\frac{\log(2T^2)}{2N_a(t)}} \ .$$

The UCB algorithm is then as follows, see UCB.

```
Sample each arm once.
for $t = (K+1), ..., T$ do
    Sample arm $a_t = \arg\max_{\mathcal{A}} \mathrm{UCB}(a,t)$
end
```

Algorithm 3: UCB

One may point out, that pulling arms based on the UCB index opens up the possibility of pulling sub optimal arms. However, for a sub optimal arm, after pulling it several times its upper confidence bound will decrease and we will cease to pull it. This intuition is summarised in the following proposition.

Proposition 1 (Contained in Proof of Theorem 1 [4])**.** Following execution of the UCB algorithm, on some problem $\nu \in \mathcal{B}$, for a sub optimal arm a one has,

$$\mathbb{E}_{\nu,\texttt{UCB}}[N_a(T)] \leq 3 + \frac{4\log(\sqrt{2}T)}{\Delta_a^2} \,.$$

Proof. For $t \leq T$, $a \in \mathcal{A}$, let $X_{a,t}$ be the sample received pulling the ath arm for the tth time. Assuming all rewards are generated in advance for the learner to uncover, for all arms $a \in \mathcal{A}$, $t \leq T$, $X_{a,t}$ is well defined. Assume without loss of generality, that $\mu_1 = \mu^*$. For some arm $a \in \mathcal{A} : \mu_a \neq \mu^*$ define the event,

$$\xi_a := \left\{\forall t < T, \mu^* \leq \frac{1}{t}\sum_{s=1}^{t} X_{1,s} + \sqrt{\frac{\log(2T^2)}{2t}}\right\} \cup \left\{\forall t < T, \left|\frac{1}{t}\sum_{s=1}^{t} X_{a,s} - \mu_a\right| \leq \sqrt{\frac{\log(2T^2)}{2t}}\right\} .$$

Via Hoeffding's and a union bound we have that,

$$\mathbb{P}(\xi_a) \geq 1 - \frac{2}{T} \,.$$

Now say there exists a time s where,

$$N_a(s) > \frac{2\log(2T^2)}{\Delta_a^2} \,.$$

For all $t \geq s$, under event ξ_a we have the following,

$$\begin{aligned} \mathrm{UCB}(a,t) &< \hat{\mu}_{a,t} + \Delta_a/2 \\ &\leq \mu_a + \Delta_a/2 + \Delta_a/2 \leq \mu^* \,, \end{aligned}$$

and thus, under event ξ_a, for all $t \geq s$,

$$\mathrm{UCB}(a,t) < \mathrm{UCB}(a^*,t) \,,$$

and

$$N_a(T) \leq \frac{2\log(2T^2)}{\Delta_a^2} + 1 \,.$$

Therefore we have that,

$$\mathbb{E}[N_a(T)] \leq \mathbb{P}(\xi_a^c)T + \frac{2\log(2T^2)}{\Delta_a^2} + 1 \leq 3 + \frac{2\log(2T^2)}{\Delta_a^2} \,.$$

□

For a policy $\pi \in \mathcal{C}$ and problem $\nu \in \mathcal{B}$, we can decompose the cummulative regret of the learner as follows,

$$R_\pi^\nu(T) = \sum_{a \in \mathcal{A}} \mathbb{E}[N_a(T)]\Delta_a \,,$$

and thus by considering the pulls on all sub optimal arms, Proposition 1 leads to the following bound on the regret of the UCB.

Theorem 7 (Theorem 1 [4]). *Let $K = |\mathcal{A}|$. For $\bar{\Delta} \in [0,1]^K$, on all bandit problems $\nu \in \mathcal{B}_{\bar{\Delta}}$, running the UCB algorithm will satisfy,*

$$R^{\pi}_{\mathit{UCB}}(T) \leq \left(1 + \frac{\pi^2}{3}\right) \sum_{a=1}^{K} \bar{\Delta}_a + 8\log(T) \sum_{a:\mu_a \neq \mu^*} \frac{1}{\bar{\Delta}_a} .$$

1.3.2 Lower bound for cumulative regret in K-armed bandit

For the lower bound, it is relatively straightforward to generalise the proof of Theorem 6 for the K armed case.

Lemma 1 ([63] Lemma 16.3). *Let $i \in [K]$ and $\nu, \nu' \in \mathcal{B}$ be two K armed bandit problems such that, ν, ν' differ only in the distribution of the ith arm and furthermore, for problem ν arm i is strictly sub optimal and for problem ν' arm i is optimal. Let*

$$\lambda = \min\left(\mu_i' - \mu_i, \max_{a \in [K]}(\mu_a) - \mu_i\right) .$$

In this case, for any policy π,

$$\mathbb{E}_{\nu,\pi}[N_i(T)] \geq \frac{\log(\lambda/4) + \log(T) - \log(R^{\nu}_{\pi}(T) + R^{\nu'}_{\pi}(T))}{\mathrm{KL}(v_i, v_i')} .$$

The proof of Lemma 1 follows from a straight forward adaptation of our technique, in the proof of Theorem 6, to the K-armed case. Lemma 1 then leads to the following Theorem.

Theorem 8 (Specific version of Theorem 16.4 [63]). *Let $K = |\mathcal{A}|$. Consider $\bar{\Delta} \in [0, 1/2)^K$ and policy π. If the regret of policy π, on all problems $\nu \in \mathcal{B}_{\bar{\Delta}}$, satisfies,*

$$R^{\nu}_{\pi}(T) \leq c\sqrt{T} ,$$

for some constant $c > 0$, then their exists a problem $\tilde{\nu} \in \mathcal{B}_{\bar{\Delta}}$, such that on problem $\tilde{\nu}$, policy π suffers the following regret,

$$R^{\tilde{\nu}}_{\pi}(T) \geq c' \sum_{a=1}^{K} \frac{\log(T\bar{\Delta}_a^2)}{\bar{\Delta}_a} ,$$

where $c' > 0$ is an absolute constant depending only on c.

Proof. Consider the bandit problem $\nu \in \mathcal{B}_{\bar{\Delta}}$ where,

$$\nu_j = \begin{cases} \mathcal{B}er(1/2) & \text{if } j = 1 \\ \mathcal{B}er(1/2 - \bar{\Delta}_j) & \text{if } j \neq 1 . \end{cases}$$

Take a suboptimal arm $i > 1$ on problem ν. Consider the bandit problem $\nu' \in \mathcal{B}_{\bar{\Delta}}$, such that,

$$\nu_j' = \begin{cases} \nu_j & \text{if } j \neq i \\ \mathcal{B}er(1/2 + \bar{\Delta}_j) & \text{if } j = i . \end{cases}$$

We remind the reader that for some $\alpha \in [0, 1/2)$,

$$\mathrm{KL}(\mathcal{B}er(1/2 - \alpha), \mathcal{B}er(1/2 + \alpha)) \leq 8\alpha^2 ,$$

and thus by application of Lemma 1 we then have,

$$\mathbb{E}_{\nu,\pi}[N_i(T)] \geq \frac{\log(\bar{\Delta}_i/4) + \log(T) - \log(R^{\nu}_{\pi}(T) + R^{\nu'}_{\pi}(T))}{8\bar{\Delta}_i^2} .$$

Therefore under the assumption that,

$$\forall \tilde{\nu} \in \mathcal{B}_{\bar{\Delta}}, R^{\tilde{\nu}}_{\pi}(T) \leq \sqrt{T}/2 ,$$

we have that,

$$\mathbb{E}_{\nu,\pi}[N_i(T)] \geq \frac{\log(T)/2 + \log(\bar{\Delta}_i/4)}{8\bar{\Delta}_i^2} ,$$

and thus,

$$R^{\nu}_{\pi}(T) \geq \sum_{a=1}^{K} \frac{\log(T\bar{\Delta}_a^2/16)}{16\bar{\Delta}_a} ,$$

providing the result.

□

Thus via combinations of Theorem 7 and Theorem 8, for a given $\bar{\Delta} \in [0,1]^K$, $\alpha > 1$, with $T \geq \min_{a\in\mathcal{A}}(\bar{\Delta}_a)^{-2\alpha}$, we have identified the minimax rate, $R^*_T(\mathcal{B}_{\bar{\Delta}})$

$$R^*_T(\mathcal{B}_{\bar{\Delta}}) := \inf_{\pi\in\mathcal{C}} \sup_{\nu\in\mathcal{B}_{\bar{\Delta}}} R^{\nu}_{\pi}(T) ,$$

as,

$$c_\alpha \log(T) \sum_{a=1}^{K} \frac{1}{\Delta_a} ,$$

where c_α is a constant depending only upon alpha.

Remark 2. In the problem independent setting the UCB does not attain the minimax rate. A modified version of the UCB, termed MOSS, see [3], is minimax optimal in the problem independent setting. However, as this thesis doesn't concern the problem independent setting for cumulative regret, an analysis of the MOSS algorithm is not included.

1.3.3 Asymptotic optimality and the KL-UCB

While the original results of this thesis all concern finite time bounds, it is nevertheless important to mention results for asymptotic bounds, i.e. as $T \to \infty$, on cumulative regret, as much of the related literature will be linked to this concept. The presence of the $\log(T)$ term in the finite time lower bound, Theorem 8, ensures that for any possible policy π, as $T \to \infty$,

$$\sup_{\nu\in\mathcal{B}} R^{\nu}_{\pi}(T) \to \infty .$$

Therefore, given a problem $\nu \in \mathcal{B}$, it makes more sense to consider how the *scaled regret*, $R^{\nu}_{\pi}(T)/\log(T)$, behaves as $T \to \infty$. Let us first introduce an asymptotic concept, that of consistency. A policy π is said to be consistent on a class of problems $\bar{\mathcal{B}}$, if, $\forall p > 0$ and $\forall \nu \in \bar{\mathcal{B}}$,

$$\lim_{T\to\infty} \frac{R^{\nu}_{\pi}(T)}{T^p} = 0 .$$

For example, from Theorem 7 one can see that the UCB is consistent.

Theorem 9 (Theorem 16.2 [63]). *For a class of bandit problems $\bar{\mathcal{B}}$, and a bandit problem $\nu \in \bar{\mathcal{B}}$, and any consistent policy π,*

$$\lim_{T\to\infty} \frac{R_\pi^\nu(T)}{\log(T)} \geq \sum_{a:\mu_a \neq \mu^*} \frac{\Delta_a}{\inf_{\sigma\in\bar{\mathcal{B}}:\mathbb{E}[\sigma_a]>\mu^*} \mathrm{KL}(\nu_\mathrm{a}, \sigma_\mathrm{a})} .$$

To give some intuition behind the result of Theorem 9, let $\mathfrak{N}$ be the class of bandit problems, where all arms follow a normal distribution $\mathcal{N}(\mu, 1)$, for some unknown $\mu \in \mathbb{R}$. That is,

$$\mathfrak{N} := \{\nu \in \mathcal{B} : \forall a \in \mathcal{A}, \exists \mu \in \mathbb{R} : \nu_a \sim \mathcal{N}(\mu, 1)\} .$$

For a $\nu \in \mathfrak{N}$ and arm $a \in \mathcal{A}$, we remind the reader that,

$$\mathrm{KL}(\mathcal{N}(\mu_a, 1), \mathcal{N}(\mu^*, 1)) = \Delta_a^2/2 , \tag{1.9}$$

thus, for some bandit problem $\nu \in \mathfrak{N}$, and arm a,

$$\inf_{\sigma\in\mathfrak{N}:\mathbb{E}[\sigma_a]>\mu^*} \mathrm{KL}(\nu_a, \sigma_a) = \Delta_a^2/2 .$$

From Theorem 7 we have that,

$$\lim_{T\to\infty} \frac{R_{\texttt{UCB}}^\nu(T)}{\log(T)} \leq c \sum_{a:\mu_a \neq \mu^*} \frac{1}{\Delta_a} .$$

for some absolute constant $c > 0$. Therefore, in the case where we restrict to the class of problems $\mathfrak{N}$, the UCB will match the asymptotic lower bound of Theorem 9, up to a constant. As we see from Equation (1.9), for normal distributions the KL-divergence is equal to the squared gap, however, this is not always the case. Let us write $\mathfrak{B}$ as the class of bandit problems where all arms follow a Bernoulli distribution $\mathcal{B}er(\mu)$ for some unknown $\mu \in [0, 1]$, that is,

$$\mathfrak{B} := \{\nu \in \mathcal{B} : \forall a \in \mathcal{A}, \exists \mu \in [0, 1] : \nu_a \sim \mathcal{B}er(\mu)\} .$$

For some bandit problem $\nu \in \mathfrak{B}$, and arm a,

$$\inf_{\sigma\in\mathfrak{B}:\mathbb{E}[\sigma_a]>\mu^*} KL(\nu_a, \sigma_a) = \mathrm{KL}(\mathcal{B}er(\mu_a), \mathcal{B}er(\mu^*)) ,$$

and,

$$\mathrm{KL}(\mathcal{B}er(\mu_a), \mathcal{B}er(\mu^*)) = \mu_a \log\left(\frac{\mu_a}{\mu^*}\right) + (1 - \mu_a) \log\left(\frac{1-\mu_a}{1-\mu^*}\right) ,$$

which can differ significantly from Δ_a^2, specifically in cases where μ_a is far from $1/2$. Therefore, if we extend beyond the restricted Gaussian setting of $\mathfrak{N}$, the UCB algorithm can no longer be considered asymptotically optimal. To achieve asymptotic optimality, in the case where we restrict to Bernoulli distributed arms, a modified version of the UCB algorithm, the KL-UCB achieves asymptotic optimality. The principle of the KL-UCB is to construct the upper confidence bounds using the KL-divergence, as opposed to Hoeffding's. We define the KL-UCB index of an arm $a \in \mathcal{A}$ at time $t < T$ as,

$$\mathrm{KLUCB}(a,t) := \max\left\{q \in \mathbb{R} : N_a(t)\mathrm{KL}(\mathcal{B}er(\widehat{\mu}_{a,t}), \mathcal{B}er(q)) \leq \log(t) + 3\log\log(t)\right\}.$$

The KL-UCB algorithm is then as the `UCB`, the only change being that at each round we pick the arm maximising the KLUCB index, as opposed to the UCB index. The asymptotic regret of the KL-UCB is upper bounded as follows.

Theorem 10 (Theorem 1 [36]). *For all $\nu \in \mathcal{B}$, the KL-UCB satisfies,*

$$\lim_{T\to\infty} \frac{R^{\nu}(T)}{\log(T)} \leq \sum_{a:\mu_a \neq \mu^*} \frac{\Delta_a}{\mathrm{KL}(\nu_a, \mathcal{B}er(\mu^*))}.$$

Thus, Theorem 10 shows the asymptotic optimality of the KL-UCB for Bernoulli distributed arms, also optimal in the constant term. Furthermore as the result extends to the case of $[0,1]$ bounded support, the KL-UCB also has potential to outperform the UCB in this more general setting, see [36].

1.4 Pure exploration bandit problems

1.4.1 Best arm identification (BAI)

In pure exploration problems, the learner does not care about their cumulative regret, but instead wishes to uncover some underlying property of the arms. A natural objective is best arm identification (BAI). That is, at the end of T rounds the learner must output a prediction $\hat{a}$ of an optimal arm. One can then evaluate their performance on their expected simple regret - the expected difference in the mean of the predicted arm to the true optimal, see Equation (1.2), or probability of error - the probability the learner outputs a non optimal arm as their prediction, see Equation (1.3). In the fixed budget setting, simple regret is a more suitable measure of regret for the problem independent case and probability of error is more suited to the problem dependent case. To see this, imagine for the two armed case that we let $\Delta \to 0$, the probability of error of any algorithm will then tend to one. Thus, probability of error is not an informative measure of regret in the problem independent setting.

BAI in the fixed confidence setting To the best of the authors knowledge, BAI was first studied, under expected simple regret, in the fixed confidence regime. In [75] the authors propose an algorithm based on successive elimination of the arms, see also [69], [32] and [31]. Confidence bound based algorithms have also been studied, see [48]. In [57] the authors provide upper and lower bounds, for the expected stopping time of PAC(δ) algorithms, see Equation (1.5), for the two armed case, that match in the asymptotic, i.e. as $\delta \to \infty$. This result is extended to the K armed bandit in [37]. To the best of the authors knowledge, optimal non asymptotic bounds, for BAI in the fixed confidence setting, remains an open question.

BAI under simple regret for the fixed budget setting For the fixed budget setting, under expected simple regret, the paper [77] classifies minimax optimal rates for the restricted case of 2 arms. Moving on to the more general K-armed case, results follow almost immediately from existing results on cumulative regret. As highlighted in [12], a straightforward adaptation of the proof of Theorem 5.1 [5] leads to the following lower bound.

Theorem 11. *Assume $T > K$, then for any policy π there exists a problem $\nu \in \mathcal{B}$ such that on problem ν the policy π suffers the following expected simple regret,*

$$r_\pi^\nu(T) \geq c\sqrt{\frac{K}{T}} ,$$

where $c > 0$ is an absolute constant.

In [12] the authors also show the following upper bound.

Theorem 12 (Corollary 3 [12])**.** *Assume $T > 2K$, on all bandit problems $\nu \in \mathcal{B}$, the strategy of uniform allocation, then outputting the arm with highest empirical mean will satisfy,*

$$r^\nu(T) \leq c\sqrt{\frac{K\log(K)}{T}} ,$$

where $c > 0$ is an absolute constant.

The proof follows almost immediately from application of Hoeffding's and then integrating over all probabilities. The authors of [12] also show that, for large enough T, running the UCB and outputting the most played arm, will also achieve the upper bound on regret of $\sqrt{\frac{K\log(K)}{T}}$, up to a constant term. The $\log(K)$ term, in discrepancy with the lower bound of Theorem 11, can be removed, by recommending the most played arm, after instead running the later introduced MOSS algorithm, of [3]. The minimax rate for BAI under simple regret,

$$r_T^*(\mathcal{B}) := \inf_{\pi\in\mathcal{C}}\sup_{\nu\in\mathcal{B}} r_\pi^\nu(T) ,$$

is therefore of the order,

$$\sqrt{\frac{K}{T}} . \tag{1.10}$$

The take away from this result is that, in the worst case problem independent setting of simple regret, cumulative regret minimisation and BAI are essentially interchangeable, with algorithms such as MOSS being optimal in both cases. Intuitively this makes sense, as when the gaps become very small, the cost of exploration disappears.

BAI under probability of error for the fixed budget setting Of more interest for BAI, is to consider the problem dependent setting, taking regret as the probability of error. As we are in the problem dependent setting, our rates will depend upon the arm gaps. We define the following problem complexity,

$$H := \sum_{a:\mu_a\neq\mu^*} \frac{1}{\Delta_a^2} , \tag{1.11}$$

and for a specific complexity $\tilde{H}$, the set of bandit problems,

$$\mathcal{B}_{\tilde{H}} := \left\{ \nu \in \mathcal{B} : \sum_{a:\mu_a\neq\mu^*} \frac{1}{\Delta_a^2} \geq \tilde{H} \right\} . \tag{1.12}$$

Following the intuition in Section 1.2, for a sub optimal arm $a : \mu_a \neq \mu^*$, if we wish to distinguish it as sub optimal, with some constant probability, we must pull it approximately $\frac{1}{\Delta_a^2}$ times. Thus H is, broadly speaking, the minimum number of samples required, to find the optimal arm with constant probability. Our overview of BAI in the problem dependent setting, will mostly draw on results from the two seminal papers, [2] and [17], the former providing upper bounds and the latter lower bounds.

A UCB based approach For the finite K-armed bandit, where $|\mathcal{A}| = K$, in [2] the authors propose UCB based algorithm, with the index tuned by an exploration parameter $h > 0$. That is, for each arm $a \in \mathcal{A}$, at time $t \leq T$ define,

$$B(a,t) := \widehat{\mu}_{a,t} + \sqrt{\frac{h}{N_a(t)}} \,.$$

Their algorithm, the UCBE is then as follows,

```
Input: exploration parameter $h$
Pull each arm once
for $t = (K+1), ..., T$ do
    Pull arm $a_t$ in $\arg\max_{a\in\mathcal{A}}(B(a,t))$
end
Output: an arm in $\arg\max_{a\in\mathcal{A}}(N_a(T))$
```

Algorithm 4: Upper Confidence Bound Exploration UCB-E

To minimise cumulative regret, one would take the exploration parameter h of the order $\log(T)$, however, it is well known that in the problem dependent setting, algorithms achieving optimal cumulative regret perform poorly in BAI. Specifically, [12] show that an algorithm with at most logarithmic cumulative regret, will have a lower bound on its simple regret, of the order $n^{-\gamma}$, for some $\gamma > 0$. Thus, we need to let our exploration parameter grow much faster with T. The following theorem shows, that by letting h grow linearly with $(T-K)/H$, the UCBE performs well.

Theorem 13 (Theorem 1 [2]). *Let $H > 0$, $T > 5K$, on all bandit problems $\nu \in \mathcal{B}_H$, running the UCBE algorithm, with $h = \frac{T-K}{16H}$ will satisfy,*

$$e^{\nu}_{\mathtt{UCBE}}(T) \leq 2TK \exp\left(-\frac{T-K}{16H}\right) .$$

Therefore assuming $T > c(H\log(TK) + K)$, for some well chosen constant c, on all bandit problems $\nu \in \mathcal{B}_H$, the UCBE will satisfy,

$$e^{\nu}_{\mathtt{UCBE}}(T) \leq \exp(-c'T/H) \,,$$

where c' is a constant depending only on c.

Proof. Set $h = \frac{T-K}{16H}$ and let $a^* = \arg\max_{a\in\mathcal{A}}(\mu_a)$. For $t \leq T$, $a \in \mathcal{A}$, let $X_{a,t}$ be the sample received pulling the ath arm for the tth time. Assuming all rewards are generated in advance for the learner to uncover, for all arms $a \in \mathcal{A}$, $t \leq T$, $X_{a,t}$ is well defined. We will work on the following favourable event,

$$\xi = \left\{ \forall t < T, a \in \mathcal{A}, \left| \frac{1}{t}\sum_{s=1}^{t} X_{a,s} - \mu_a \right| \leq \sqrt{\frac{h}{4t}} \right\} .$$

Via Hoeffding's and a union bound we have,

$$\begin{aligned}\mathbb{P}(\xi) &\geq 1 - 2KT\exp(-h)\\ &\geq 1 - 2KT\exp\left(-\frac{T-K}{16H}\right) .\end{aligned}$$

Under event ξ we have that,

$$\forall t < T, B(a^*, t) > \mu^* . \tag{1.13}$$

Now assume there exists an arm $a \neq a^*$ and time s such that $a_s = a$ and $N_a(s) \geq \frac{4h}{\Delta_a^2}$. With this assumption, under event ξ,

$$\begin{aligned}B(a, s) &\leq \widehat{\mu}_{a,s} + \Delta_a/2\\ &\leq \mu_a + \Delta_a/4 + \Delta_a/2 < \mu^* ,\end{aligned}$$

contradicting Equation (1.13), and therefore under event ξ we have,

$$\forall a \neq a^*, N_a(T) \leq \frac{4h}{\Delta_a^2} + 1 . \tag{1.14}$$

Thus following Equation (1.14) we have that, under event ξ,

$$\begin{aligned}N_{a^*}(T) &\geq T - \sum_{a \neq a^*}\left(\frac{4h}{\Delta_a^2} + 1\right) ,\\ &\geq T - K - 4hH ,\\ &\geq T - K - \frac{T-K}{4} ,\\ &\geq T/2 ,\end{aligned}$$

where final line comes from the fact we assume $T > 5K$. Therefore, under event ξ, $\arg\max_{a\in\mathcal{A}}(N_a(T)) = a^*$ and the proof follows.

□

Problem dependent lower bound for BAI The authors of [2] also show that, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}_H$, such that,

$$e_\pi^\nu(T) \geq \exp(-c(T\log(K))/H) ,$$

for an absolute constant $c > 0$. The $\log(K)$ gap is tightened in the paper [17], where the authors provide the following lower bound.

Theorem 14 (Theorem 1 [17]). *Let $H > 0$ and assume $T > cH^2\log(TK)$, for a well chosen absolute constant c, then, for any policy π there exists a problem $\nu \in \mathcal{B}_H$, such that on problem ν the policy π suffers the following regret,*

$$e_\pi^\nu(T) \geq \frac{1}{6}\exp(-c'T/H) ,$$

where $c' > 0$ is a constant depending only on c.

Therefore in the case where $H > 0$ is known and $T > c(H^2 \log(TK) + K)$, for some absolute constant c, we see that via combination of Theorems 13 and 14, for BAI under probability of error, on the class of problems $\mathcal{B}_H$, we have upper and lower bounds of matching order, up to a constant term in the exponential depending on c.

Successive rejects algorithms in the case where H is unknown The issue remains, that the UCBE needs to know the problem complexity H to optimally tune its exploration parameter h. Is it possible to construct optimal algorithms and classify the minimax rate in the case where H is unknown? Algorithms based on successive elimination of arms, that do not require knowledge of H, have been proposed, see [2], [52]. The successive rejects (SR) algorithm of [2] is as follows, see SR. To overcome a technical issue the authors of [2] define,

$$\overline{\log}(K) := \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k} \,.$$

For $K > 2$ we have $\overline{\log}(K) \leq 2\log(K)$. The SR algorithm runs over K rounds

```
Initialise: $\mathcal{A}_1 = [K]$, $\overline{\log}(K) = \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k}$ and for $k \in [K]$ set,
    $$n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-k} \right\rceil$$
for $k \in [K-1]$ do
    For each $a \in \mathcal{A}_k$, pull arm $a$, $n_k - n_{k-1}$ times, let $\widehat{\mu}_{a,n_k}$ denote the sample mean
    Set $\mathcal{A}_{k+1} = \mathcal{A}_k \backslash \{a \in \arg\min_{a \in \mathcal{A}_k} \widehat{\mu}_{a,n_k}\}$[1]
end
Output: unique element of $\mathcal{A}_K$
```

Algorithm 5: Successive rejects algorithm (SR)

during which it maintains an active set of arms. Initially, the active set is the entire set of arms. Then, at each round k, each arm in the active set is pulled a fixed number of times, $n_k - n_{k-1}$. Thus at the end of round k, every remaining arm has been pulled n_k times. The single arm with minimal empirical mean is then discarded. Therefore, after K rounds only one arm will remain and importantly, on completion the algorithm, for each $k \in [K-2]$, exactly one arm has been pulled a total of n_k, and exactly 2 arms have been pulled a total of n_{K-1} times. Therefore, we can bound the total number of pulls of the SR as,

$$\begin{aligned}\sum_{k=1}^{K-1} \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-k} \right\rceil + \left\lceil \frac{T-K}{2\overline{\log}(K)} \right\rceil &\leq K + \frac{T-K}{\overline{\log}(K)} \left(\sum_{k=1}^{K-1} \frac{1}{K+1-k} + \frac{1}{2} \right) \\ &= T \,.\end{aligned}$$

The number of samples it allocates per arm per round increases, for instance, at the end of the first round, all arms have been pulled only $\left\lceil \frac{T-K}{\overline{\log}(K)K} \right\rceil$ times. Whereas $n_K = \left\lceil \frac{T-K}{2\overline{\log}(K)} \right\rceil$, so that at end of the final round, when only two arms remain both

[1] If multiple arms exist in $\arg\min_{a \in \mathcal{A}_k} \widehat{\mu}_{a,n_k}$, choose one to eliminate at random.

have pulled $\left\lceil \frac{T-K}{2\overline{\log}(K)} \right\rceil$ times. Intuitively this makes sense, as when few arms remain, we need to be more careful that we do not discard the optimal arm and on the other hand, when there are many arms, we do not want to waste our budget unnecessarily. For more intuition, see the proof of the following theorem, which upper bounds the probability of error of the- SR algorithm.

Theorem 15 (Theorem 2 [2]). *Let $H > 0$, on all bandit problems $\nu \in \mathcal{B}_H$, the SR algorithm satisfies,*

$$e^{\nu}_{SR}(T) \leq c' K(K-1) \exp\left(-c\frac{T-K}{\overline{\log}(K)H}\right) ,$$

for absolute constants $c, c' > 0$. Therefore, assuming $cT > 2c'H\overline{\log}(K)^2 + cK$, for all bandit problems $\nu \in \mathcal{B}_H$,

$$e^{\nu}_{SR}(T) \leq \exp\left(-c\frac{T}{2\overline{\log}(K)H}\right) .$$

Proof. Assuming all rewards are generated in advance for the learner to uncover, even if for some $k \in [K]$ there exists an arm $a \in \mathcal{A}$ which is not pulled n_k times, we can still define $\widehat{\mu}_{a,n_k}$. For $a \in [K]$ define the arm with ath highest mean as (a). If arms have equal means, let them be ordered randomly amongst themselves. Define the optimal arm as, $a^* := \arg\max_{a\in\mathcal{A}} \mu_a$. If the optimal arm is eliminated in the kth round, i.e.

$$a^* \notin \mathcal{A}_{k+1} ,$$

then the following must hold,

$$\widehat{\mu}_{a^*,n_k} \leq \max_{a\in\{(K),\dots,(K+1-k)\}} \widehat{\mu}_{a,n_k} .$$

Now via union bound,

$$e(T) \leq \sum_{k=1}^{K-1} \mathbb{P}(a^* \notin \mathcal{A}_{k+1}) ,$$

and then via Hoeffding's,

$$\begin{aligned} e(T) &\leq \sum_{k=1}^{K-1} \mathbb{P}(a^* \notin \mathcal{A}_{k+1}) \leq \sum_{k=1}^{K-1} \sum_{a=K+1-k}^{K} \mathbb{P}\big(\widehat{\mu}_{a^*,n_k} \leq \widehat{\mu}_{(a),n_k}\big) , \\ &\leq \sum_{k=1}^{K-1} \sum_{a=K+1-k}^{K} \exp\Big(-n_k\Delta^2_{(a)}/2\Big) \leq \sum_{k=1}^{K-1} k \exp\Big(-n_k\Delta^2_{(K+1-k)}/2\Big) . \end{aligned}$$

Now note that for all $k \in [K]$,

$$\frac{k}{\Delta^2_{(k)}} \leq \sum_{a=1}^{K} \frac{1}{\Delta^2_a} ,$$

and therefore,

$$\sum_{k=1}^{K-1} k \exp\left(-n_k \Delta^2_{(K+1-k)}/2\right) = \sum_{k=1}^{K-1} k \exp\left(-\frac{(T-K)}{2\overline{\log}(K)} \frac{\Delta^2_{(K+1-k)}}{K+1-k}\right) \tag{1.15}$$

$$\leq 2K(K-1) \exp\left(-\frac{T-K}{2\overline{\log}(K)H}\right) . \tag{1.16}$$

Thus completing the proof. □

As one can see from Theorem 15, the SR algorithm pays an additional multiplicative $\log(K)$ term, as the cost for being adaptive to unknown H, we remind the reader that for $K > 2$, $\overline{\log}(K) \leq 2\log(K)$. For some time, the necessity of $\log(K)$ term remained an open question. A $\log(K)$ term may seem insignificant, however, as it is in the exponential, it can have a severe effect on the regret as K grows. In [17] the authors also show that, in the case where one does not have a tight bound on H, the $\log(K)$ cost for adaptation is necessary, as the following theorem shows.

Theorem 16 (Theorem 1 [17]). *Assume $T > cK^2 \log(TK)$, for some well chosen constant $c > 0$, then for any policy π there exists a problem $\nu \in \{\mathcal{B}_H : H < c'K^2\}$, where c' is a constant depending only on c, such that on problem ν the policy π suffers the following regret,*

$$e^\nu_\pi(T) \geq \frac{1}{6} \exp(-c''T/(\log(K)H_\nu)) ,$$

where H_ν is the complexity, as defined in Equation 1.11, associated to problem ν and $c'' > 0$ is an absolute constant.

By combination of Theorems 15 and 16 we have demonstrated the optimality of the SR algorithm, in the case were one does not have too tight a bound on the problem complexity, H. This completes our overview of BAI.

1.4.2 TopM problem

Pure exploration bandit problems go beyond best arm identification. A natural extension is the TopM problem, here the objective of the learner is to identify the m best arms. In the TopM setting, the difficulty is in classifying arms as belonging to the m best arms or not. The problem hardness is thus now dependent on the distances of arms to the mth and $m+1$th best arms. With this in mind, we define the following notion of arm gaps. Without loss of generality, assume that the arms are ordered as follows,

$$\mu_1 \geq \mu_2 \geq ... \geq \mu_K{}^{2} ,$$

and for a given arm a define the following gap,

$$\Delta_a^{<m>} := \begin{cases} \mu_m - \mu_a & \text{if } a > m \\ \mu_a - \mu_{m+1} & \text{if } a \leq m . \end{cases}$$

TopM problem in the fixed confidence setting The TopM problem has been extensively studied in the fixed confidence PAC(ε, δ) setting and is relatively well understood. In this setting the learner must return a set $\hat{S}$ of m arms, where $|\hat{S}|$ must

[2] This assumption is made for simplicity, we remind the reader that the learner remains unaware of the ordering of the arms.

equal m and then suffers the regret,

$$(\mu_m - \min_{a \in \hat{S}} \mu_a) \vee 0 \,.$$

The objective of the learner is to then provide a PAC(δ, ε) algorithm with minimal stopping time. There have been approaches using uniform sampling and successive elimination, see [59] and [88], as well as sequential index based algorithms, see [59] and [51]. As it will be of use later on, in Section 1.7, we will briefly cover such an index based approach, the LUCB (lower upper confidence bound) algorithm. For each arm $a \in \mathcal{A}$, and some exploration parameter $\beta(t, \delta)$, we then define the following two indexes,

$$\overline{\text{UCB}}(a, t) = \widehat{\mu}_{a,t} + \sqrt{\frac{\beta(t, \delta)}{N_a(t)}} \,,$$

and

$$\text{LCB}(a, t) = \widehat{\mu}_{a,t} - \sqrt{\frac{\beta(t, \delta)}{N_a(t)}} \,.$$

Also, at time t define the set of arms $J_m(t)$ as the m best arms according to empirical mean, at time t. The LUCB algorithm is then as follows, see LUCB. At each iteration

```
Input: confidence level $\delta$, tolerance $\varepsilon$
Pull each arm once
Set $\psi(t) = \varepsilon$
while $\psi(t) \geq \varepsilon$ do
    Pull arm $a_t = \arg\max_{a \notin J_m(t)}(\overline{\text{UCB}}(a, t))$
    Pull arm $b_t = \arg\min_{a \in J_m(t)}(\text{LCB}(a, t))$
        $\psi(t) = \max_{a \notin J_m(t)} \overline{\text{UCB}}(a, t) - \min_{a \in J_m(t)} \text{LCB}(a, t)$
end
Output: $J_m(T)$
```

Algorithm 6: LUCB

the LUCB samples a pair of arms. One sampling rule of the LUCB,

$$\arg\min_{a \in J_m(t)}(\text{LCB}(a, t)) \,,$$

explores the m best arms, adaptive to their gaps and should sample the mth best arm many times. The other sampling rule of the LUCB,

$$\arg\max_{a \in J_m(t)}(\overline{\text{UCB}}(a, t)) \,,$$

explores the $m + 1$th to Kth best arms, adaptive to their gaps and should sample the $m + 1$th best arm many times. These sampling rules, allows it to define its stopping rule,

$$\max_{a \notin J_m(t)} \overline{\text{UCB}}(a, t) - \min_{a \in J_m(t)} \text{LCB}(a, t) < \varepsilon \,.$$

The performance of the LUCB is bounded in the following Theorem.

Theorem 17 (Combination of Theorem 1, Corollary 7 [51]). *Let $\varepsilon, \delta > 0$. For a suitable choice of exploration parameter, $\beta(t, \delta)$, the* LUCB *algorithm with output m arms, $\hat{S}$, is PAC(δ, ε), that is,*

$$\mathbb{P}((\mu_m - \min_{a \in \hat{S}} \mu_a) \leq \varepsilon) \geq 1 - \delta \ ,$$

and furthermore, for a constant $c > 0$, with probability $1 - c\delta$, has its stopping time τ bounded as,

$$\tau \leq c' H_\varepsilon^m \log(H_\varepsilon^m / \delta) \ ,$$

where $H_\varepsilon^m = \sum_{a \in \mathcal{A}} \frac{1}{\varepsilon^2 \wedge (\Delta_a^{<m>})^2}$ and $c' > 0$ is a constant depending only on c.

Remark 3. The complexity H_ε^m is the standard TopM problem complexity, with the additional property that when the gaps are smaller than ε, they no longer effect the complexity, which makes sense in the PAC(δ, ε) context.

TopM problem in fixed budget setting Compared to the fixed confidence setting, the fixed budget has seen less progress. To the best of the authors knowledge, the TopM problem was first considered in the fixed budget setting in [14]. Again, without loss of generality, assume the arms are ordered as follows,

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K \ .$$ [3]

As in the fixed confidence setting, the learner must output a list $\hat{S}$ of m arms, however, we now take regret as the probability of error,

$$\mathbb{P}\Big(\exists a \leq m : a \notin \hat{S}\Big) \ .$$

Let us first define our problem complexity for the fixed budget setting,

$$H_1^m := \sum_{a=1}^{K} (\Delta_a^{<m>})^{-2} \ .$$

H_1^m is essentially a direct parallel with (1.11) in the best arm identification setting. A slightly weaker version of problem complexity is,

$$H_2^m := \max_{a \in \mathcal{A}} \big(a(\Delta_a^{<m>})^{-2}\big) \ .$$

We say H_2^m is weaker than H_1^m, in the sense that,

$$H_2^m \leq H_1^m \ ,$$

however, they differ at most up to a multiplicative $\log(2K)$ term,

$$H_1^m \leq \log(2K) H_2^m \ .$$

The authors of [14] describe a successive accept and reject algorithm, which we will denote *SRM*, that for all $\nu \in \mathcal{B}$, for the TopM problem, has the following upper

[3] We make such an assumption only for simplicity of notation. As before, the learner does not know the ordering of the arms.

bound on its probability of error,

$$\mathbb{P}_{\nu,SRM}\left(\exists a \leq m : a \notin \hat{S}\right) \leq 2K^2 \exp\left(-c\frac{T-K}{8\log(K)H_2^m}\right) ,$$

for some absolute constant $c > 0$. The authors also conjecture a lower bound of the order,

$$\exp\left(-c'\frac{T}{H_1^m}\right) , \tag{1.17}$$

for some absolute constant c'. Even if one was able to prove the conjectured lower bound, there would remain the discrepancy between, H_1^m and H_2^m, although this would at most amount to a $\log(K)$ factor. To the best of the authors knowledge, no explicitly stated lower bound for the TopM problem, in the fixed budget setting, exists in the literature.

Another related problem is multi bandit best arm identification, (MB). In this setting the arm set is partitioned into several sets of arms. The learner has complete knowledge of the partition and aims to return the set wise optimal arms, for the MB in the fixed budget setting, see [35].

1.5 The thresholding bandit problem (TBP)

In the TopM problem, we wish to identify a subset of arms, the m best arms. Another, very natural, subset to consider is, for some given threshold τ, the set of arms with mean greater than τ. Recovery of this set, is the thresholding bandit problem. In [23] the authors describe a general framework for pure exploration bandit problems, which they term "Combinatoral Pure Exploration" (CPE) bandit problems, the thresholding bandit problem fits into the CPE framework and therefore, to the best of the authors knowledge, [23] can be considered as the first work to consider the thresholding bandit problem. In this setting the learner is given a threshold τ and aims to correctly classify all arms as above or below threshold, based on their mean. That is, if for an arm a, such that $\mu_a \geq \tau$, arm a is said to be above threshold, and below threshold otherwise. At the end of T rounds, the learner must output a list $\hat{Q} \in \{-1,1\}^K$ that classifies the arms as above or below threshold. Let Q encode the true classification, i.e.

$$Q_a = 2\mathbb{1}_{\{\mu_a \geq \tau\}} - 1 ,$$

with the convention $Q_a = 1$ if arm a is above the threshold and $Q_a = -1$ otherwise. For a policy $\pi \in \mathcal{C}$ and problem $\nu \in \mathcal{B}$, we can then define an equivalent simple regret in this setting as,

$$\bar{r}_\pi^\nu(T) := \mathbb{E}_{\nu,\pi}\left[\max_{a:\hat{Q}_a \neq Q_a} |\tau - \mu_a|\right] . \tag{1.18}$$

We can also measure the regret of the learner again by probability of error, i.e. the probability they misclassify at least one arm,

$$\bar{e}_\pi^\nu(T) := \mathbb{P}_{\nu,\pi}\left(\exists a : \mu_a \neq \tau : \hat{Q}_a \neq Q_a\right) .$$

1.5.1 The APT algorithm and problem dependent bounds for the TBP

In [68], the TBP is studied under the objective of minimising probability of error. The work of [68] is in the fixed budget finite armed bandit setting, with $|\mathcal{A}| = K$. They provide an algorithm, the APT and demonstrate an upper bound on its probability of error, along with a lower bound, on the probability of error for any possible algorithm, matching up to log terms. Before analysing the APT we must first define the gaps, specific to the thresholding bandit problem. The gap of an arm is now defined as its distance to the threshold, as opposed to the optimal arm, i.e. for any $a \in \mathcal{A}$ let,

$$\underline{\Delta}_a := |\mu_a - \tau| \, .$$

As for BAI, the problem complexity is then again taken as

$$\underline{H} := \sum_{a:\mu_a \neq \tau} \frac{1}{\underline{\Delta}_a^2} \, ,$$

and for a specific complexity $\tilde{H}$, we can define,

$$\underline{\mathcal{B}}_{\tilde{H}} : \{\nu \in \mathcal{B} : \underline{H} \geq \tilde{H}\} \, .$$

The APT algorithm The APT algorithm is index based, in that at each time step it calculates an index for each arm and pulls the arm with minimum index. As one would expect this index differs from the UCB index. Firstly, it is based on the empirical gaps of the arms, not the empirical means, that is, for an arm $a \in \mathcal{A}$ at time t define,

$$\hat{\underline{\Delta}}_{a,t} := |\widehat{\mu}_{a,t} - \tau| \, ,$$

we then define the index,

$$\underline{B}_\tau(a,t) = \sqrt{N_a(t)}\hat{\underline{\Delta}}_{a,t} \, .$$

```
Input: $\tau$
Pull each arm once
for $t = K+1, ..., T \in [T-K]$ do
    Pull arm $a_t = \arg\min_{a \in \mathcal{A}}(\underline{B}_\tau(a,t))$[4]
end
Output: $\hat{S} = \{a : \widehat{\mu}_{a,T} \geq \tau\}$
```

Algorithm 7: APT

Say we pull an arm $a \in \mathcal{A}$, a total of n times with empirical mean $\widehat{\mu}_a$. If we classify arm a according to its empirical mean, via Hoeffding's the probability of,

$$|\mu_a - \widehat{\mu}_a| \geq \underline{\Delta}_a \, ,$$

and thus the probability of incorrect classification will be roughly,

$$\exp\left(-cn\underline{\Delta}_a^2\right) \, ,$$

[4] If $\arg\min_{a \in \mathcal{A}}(\underline{B}_\tau(a,t))$ contains multiple arms, choose a single arm at random.

where $c > 0$ is an absoulte constant. Therefore if one wishes to achieve the regret bound of approximately,

$$\exp(-c'T/\underline{H}) \ ,$$

for some absolute constant $c' > 0$, mirroring that for BAI, for all $a \in \mathcal{A}$ we would need $N_a(T)$ of the order,

$$\frac{T}{\underline{\Delta}_a^2 \underline{H}} \ .$$

This is the principle behind taking the index of the APT as $\sqrt{N_a(t)}\hat{\underline{\Delta}}_{a,t}$.

Problem dependent upper bound for the APT The following theorem upper bounds the performance of the APT.

Theorem 18 (Theorem 2 [68]). *Let $H > 0$, $T > 2K$, on all bandit problems $\nu \in \underline{\mathcal{B}}_H$, the APT algorithm will satisfy,*

$$\bar{e}^{\nu}_{\texttt{APT}}(T) \leq \exp\left(-c\frac{T}{H} + c' \log(K(\log(T) + 1))\right) \ ,$$

where $c, c' > 0$ are absolute constants.

Proof. For $t \leq T$, $a \in \mathcal{A}$, let $X_{a,t}$ be the sample received pulling the ath arm for the tth time. Assuming all rewards are generated in advance for the learner to uncover, for all arms $a \in \mathcal{A}$, $t \leq T$, $X_{a,t}$ is well defined. During the proof we will work under a favourable event ξ,

$$\xi := \left\{ \forall a \in \mathcal{A}, \forall t \leq T, \left| \frac{1}{t}\sum_{s=1}^{t} X_{a,s} - \mu_a \right| \leq \alpha\sqrt{\frac{T}{Ht}} \right\} \ , \tag{1.19}$$

where we take $\alpha = \frac{1}{4\sqrt{2}}$. For all $u \leq \lfloor \log(T) \rfloor$ and $a \in \mathcal{A}$, we have that, via a martingale inequality,

$$\begin{aligned}\mathbb{P}\left(\exists v \in [2^u, 2^{u+1}] : \left| \frac{1}{v}\sum_{s=1}^{v} X_{a,s} - \mu_a \right| > \alpha\sqrt{\frac{2T}{Hv}} \right) &\leq 2\exp\left(-\frac{\alpha^2 T v}{H 2^{u+1}}\right) \\ &\leq 2\exp\left(-\frac{\alpha^2 T}{2H}\right) \ .\end{aligned}$$

Then considering a union bound across all $(\log(T) + 1)K$ combinations, we can show,

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

The next step of the proof is to recognise that after completion of our budget, at least one arm $\tilde{a} \in \mathcal{A}$ must have been pulled at least $\frac{T-K}{2H\underline{\Delta}_{\tilde{a}}^2} + 1$ times. To see this, if,

$$\forall a \in \mathcal{A}, N_a(T) < \frac{T-K}{H\underline{\Delta}_a^2} + 1 \ ,$$

then,

$$T < \sum_{a=1}^{K}\left(\frac{T-K}{H\underline{\Delta}_a^2} + 1\right) = T \ ,$$

a contradiction. Furthermore as $T > 2K$,

$$N_{\tilde{a}}(T) \geq \frac{T}{2H\underline{\Delta}_a^2} + 1 \; .$$

Let $\tilde{s}$ be the last time arm $\tilde{a}$ was pulled, we have $N_{\tilde{a}}(\tilde{s}) \geq \frac{T}{2H\underline{\Delta}_{\tilde{a}}^2}$. Now for all $a \in \mathcal{A}$, under event ξ,

$$\left(\underline{\Delta}_a - \alpha\sqrt{\frac{T}{N_a(\tilde{s})H}}\right)\sqrt{N_a(\tilde{s})} \leq B_\tau(a, \tilde{s}) \leq \left(\underline{\Delta}_a + \alpha\sqrt{\frac{T}{N_a(\tilde{s})H}}\right)\sqrt{N_a(\tilde{s})} \; . \tag{1.20}$$

and specifically for arm $\tilde{a}$,

$$B_\tau(\tilde{a}, \tilde{s}) \geq (1/\sqrt{2} - \alpha)\sqrt{\frac{T}{H}} \; . \tag{1.21}$$

Via action of the algorithm, and combination of Equations (1.20) and (1.21), for all $a \in \mathcal{A}$, at time $\tilde{s}$, on event ξ we must have,

$$\left(\underline{\Delta}_a + \alpha\sqrt{\frac{T}{N_a(\tilde{s})H}}\right)\sqrt{N_a(\tilde{s})} \geq (1/\sqrt{2} - \alpha)\sqrt{\frac{T}{H}} \; ,$$

and thus,

$$N_a(\tilde{s}) \geq (1 - 2\sqrt{2}\alpha)^2 \frac{T}{2\underline{\Delta}_a^2 H} \; .$$

For an arm $a \in \mathcal{A}$, as $N_a(T) \geq N_a(\tilde{s})$, we then have that, on event ξ,

$$|\widehat{\mu}_{a,T} - \mu_a| \leq \frac{\alpha\sqrt{2}\underline{\Delta}_a}{1 - 2\sqrt{2}\alpha} \; ,$$

taking $\alpha = \frac{1}{4\sqrt{2}}$, the above simplifies to $|\widehat{\mu}_{a,T} - \mu_a| \leq \underline{\Delta}_a/2$. Thus, on the event ξ, the APT will classify all arms correctly.

□

Remark 4. It is important to note, that in [68] their results are extended to a more general setting, where, for some $\varepsilon > 0$, the learner only aims to classify arms of distance greater than ε to the threshold. The analysis in such a setting remains unchanged, one simply makes an alternate definition of the gaps,

$$\underline{\Delta}_a^\varepsilon := |\mu_a - \tau| + \varepsilon \; .$$

For simplicity we will fix $\varepsilon = 0$. The results of [68] can also be extended to the sub Gaussian case, again for simplicity, we restrict to distributions with bounded support on $[0, 1]$.

Problem dependent lower bound for the TBP The authors of [68] also prove a lower bound, they show that for $H > 0$, for any policy π, there exists a problem $\nu \in \mathcal{B}_H$ of the order, such that,

$$\bar{e}_\pi^\nu(T) \geq \exp\left(-c\frac{T}{H} - c' \log(K \log(T))\right) \; ,$$

where $c, c' > 0$ are absolute constants, thus, the upper bound on the regret of the APT matches the corresponding lower bound, up to multiplicative constants and additive log terms in the exponential and interestingly one does not pay an additional $\log(K)$ term for adaptation, as in the BAI setting. The reason for this, is that in the TBP the learner has a key advantage over BAI, that is, they know the threshold τ. This would be akin to knowing the value of μ^* in the BAI setting. This advantage allows for the removal of the multiplicative $\log(K)$ term in the exponential. Indeed, by setting $\tau = \mu^*$ and outputting the most pulled arm, the APT is optimal in the BAI setting, see [68][Chapter 3.2]. The TBP was further studied in [70] and [86] where the authors consider a "variance aware algorithm", that bases its actions on the empirical variances, as well as the empirical means. Their problem complexity and then also their rates, then depend upon the variances of the arms. This gives their results a significantly different flavour to those of [68] and this thesis.

1.5.2 Problem independent bounds for the TBP

With the results of [68] in mind, a question of interest is whether one can recover optimal rates on the expected simple regret, that is, the expected maximum distance to threshold among mis-classified arms, see Equation (1.18), in the problem independent setting. Unlike the independent setting for BAI, simply running MOSS and outputting the most pulled arm will no longer work, as one now needs to separately classify each arm. The lower bound of [12], will hold for the TBP under expected simple regret, that is, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}$, such that,

$$\bar{r}_\pi^\nu(T) \geq c\sqrt{\frac{K}{T}} \, ,$$

for some absolute constant $c > 0$. The problem dependent upper bound of [68], on the probability of error for the `APT` algorithm, can be adapted to the problem independent setting to show that, for all $\nu \in \mathcal{B}, \delta > 0$,

$$\mathbb{P}_{\nu,\texttt{APT}}\left(\max_{a:\hat{Q}_a \neq Q_a} |\tau - \mu_a| \leq c'\sqrt{\frac{K\log(K\log(T/\delta))}{T}} \right) \geq 1 - \delta \, ,$$

for some absolute constant $c' > 0$.

Contribution

In the paper [26], co authored with Pierre Menard and Alexandra Carpentier, we tighten both of the above bounds so that they match. We show that algorithm, `Uniform`, which entails a uniform sampling of the arms, and then classification of arms as above or below threshold according to their sample means, on all problems $\nu \in \mathcal{B}$, has an upper bound on it's expected simple regret of the following order,

$$\bar{r}_{\texttt{Uniform}}^\nu(T) \leq c\sqrt{\frac{K\log(K)}{T}} \, ,$$

where $c > 0$ is an absolute constant, see Proposition 3. In itself this result may not be too surprising, however, we also prove that, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}$, such that,

$$\bar{r}^{\nu}_{\pi}(T) \geq c' \sqrt{\frac{K \log(K)}{T}} \ ,$$

where $c' > 0$ is an absolute constant, see Proposition 2. As our lower and upper bounds are of the same order, we can say that our uniform algorithm is optimal and we have identified the minimax rate, up to a multiplicative constant, for the problem independent TBP,

$$\bar{r}^{*}_{T}(\mathcal{B}) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}} \bar{r}^{\nu}_{\pi}(T) \ ,$$

as of the order,

$$\sqrt{\frac{K \log(K)}{T}} \ , \tag{1.22}$$

see Theorem 23.

The fact that a uniform sampling strategy can attain the minimax rate may seem surprising, however, remember we are in the *problem independent setting*. Essentially, what the result of Equation (1.22) tells us, is that the hardest problem is one such that,

$$\forall a \in \mathcal{A}, \underline{\Delta}_a = c \sqrt{\frac{K \log(K)}{T}} \ ,$$

where $c > 0$ is some absolute constant and in this case, uniform sampling is the best one can do.

1.6 Shape constrained thresholding bandit problem

The scope of [26] extends much further than the above result. We also consider the TBP under several shape constraints on the sequence of arm means.

1.6.1 Problem independent TBP under a monotone constraint

The first constraint we consider is a monotone constraint. In this setting the learner is given the information that the means form a monotonically increasing sequence. That is,

$$\mu_1 \leq \mu_2 \leq ... \leq \mu_K \ .$$

For the monotone constraint define the following class of bandit problems,

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

A monotone constraint is a natural extension to the TBP when considering potential applications. For example, in [39], this version of the TBP is considered in the context of drug dosing. When developing a new drug one wishes to maximise the dosage without having side effects exceed a certain tolerance. They consider the problem in the fixed confidence setting. That is, given some confidence level $\delta > 0$, they aim to provide a PAC$(\delta, 0)$ algorithm, see Equation (1.5), with minimal stopping time. They provide an algorithm which they then show to be optimal in the asymptotic sense, i.e.

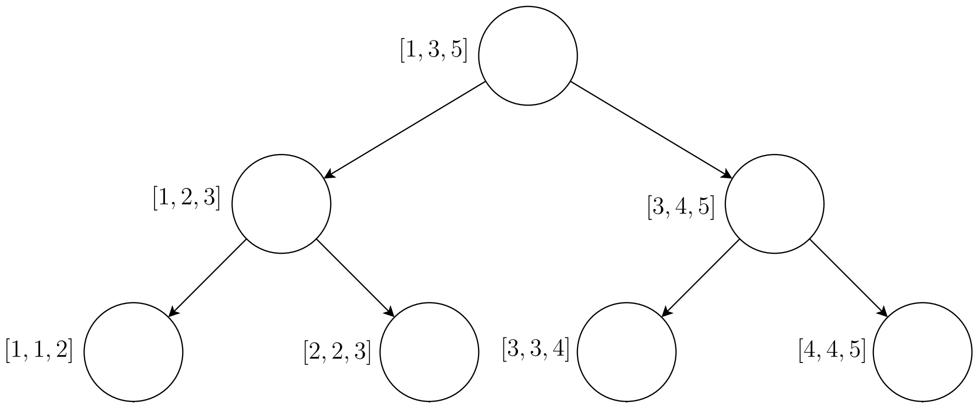

FIGURE 1.1: 5 Armed bandit problem viewed as a binary tree

as $\delta \to \infty$. There remains a large gap between their work and that of finding optimal rates on expected simple regret, or indeed probability of error, for a fixed T.

Noisy binary search An immediate solution under the monotone constraint, is to use a binary search to find the point at which the arms cross the threshold, as noted in [53] Section 1.2. To do this we first map our arm set onto 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 $\lceil\log(K)\rceil$.

Now consider the following algorithm, Naive Binary Search, that moves down the binary tree. For $i \le \lceil\log(K)\rceil$, let v_i be the current node the algorithm samples from at the ith step of the search, with $v_1 = \texttt{root}$. At each step $i \le \lceil\log(K)\rceil$ we sample $T/\lceil\log(K)\rceil$ times, the arm corresponding to the middle index of the current node, $v_i(m)$, let $\widehat{\mu}_i$ denote it's empirical mean. We then progress to the right or left child, setting $v_{i+1} = L(v_i)$ or $v_{i+1} = R(v_i)$, depending on whether the empirical mean $\widehat{\mu}_i$ is above or below τ respectively.

```
for $i = 1, ..., \lceil\log(K)\rceil$ do
    Sample $T/\lceil\log(K)\rceil$ times arm $v_i(m)$, let $\hat{\mu}_i$ denote the empirical mean
    if $\hat{\mu}_i < \tau$ then
        $v_{i+1} = R(v_i)$
    end
    if $\hat{\mu}_i \ge \tau$ then
        $v_{i+1} = L(v_i)$
    end
end
Output: $\widehat{Q} = 2\mathbb{1}_{a \ge v_i(r)} - 1$
```

Algorithm 8: Naive binary search

The Naive Binary Search moves down through the binary tree until it reaches a leaf. At each step it decides to progress to the left or right child of the current node, however, only one direction is the correct decision. As the decision is made solely based on whether or not $\widehat{\mu}_i > \tau$, if $\widehat{\mu}_i$ is close to the true mean $\mu_{v_i(m)}$, specifically,

$$|\widehat{\mu}_i - \mu_{v_i(m)}| < \underline{\Delta}_{v_i(m)} \ ,$$

we can be sure the algorithm progresses to the correct child. If we have that,

$$\forall i \leq \lceil \log(K) \rceil, |\widehat{\mu}_i - \mu_{v_i(m)}| < \underline{\Delta}_{v_i(m)} \ ,$$

we can be sure to correctly identify the point at which the arms cross the threshold. However, if there is potential for the algorithm to make a single mistake, i.e.

$$\exists i : |\widehat{\mu}_i - \mu_{v_i(m)}| \geq \underline{\Delta}_{v_i(m)} \ ,$$

we have no guarantee on our regret bound. This is demonstrated in the following theorem.

Theorem 19 (Following intuition of Section 1.2 [53])**.** *Assume $T > \log(K)$, on all bandit problems $\nu \in \mathcal{B}_m$, running the `Naive Binary Search` will satisfy,*

$$\bar{r}^{\nu}(T) \leq c\sqrt{\frac{\log(K)(\log\log(K) \vee 1)}{T}} \ ,$$

where $c > 0$ is an absolute constant.

Proof. For $i < \lceil \log(K) \rceil$, let $\mathcal{F}_i$ be the information available up to and including the ith step of the algorithm and define the event,

$$\xi_i := \{|\widehat{\mu}_i - \mu_{v_i(m)}| \geq \varepsilon\} \ .$$

For $i < \lceil \log(K) \rceil$, via Heoffding's we have that,

$$\mathbb{P}(\xi_i|\mathcal{F}_i) \leq 2\exp\left(\frac{-2T\varepsilon^2}{\lceil \log(K) \rceil}\right) \ .$$

Therefore,

$$\mathbb{P}\left(\bigcap_{i \leq \lceil \log(K) \rceil} \xi_i^c\right) \geq 1 - 2\lceil \log(K) \rceil \exp\left(\frac{-2T\varepsilon^2}{\lceil \log(K) \rceil}\right)$$

and thus,

$$\mathbb{P}(r(T) \geq \varepsilon) \leq 2\lceil \log(K) \rceil \exp\left(\frac{-2T\varepsilon^2}{\lceil \log(K) \rceil}\right) \wedge 1 \ .$$

Now set $\varepsilon_0 = \sqrt{\frac{\lceil \log(K) \rceil \log\lceil \log(K) \rceil}{2T}}$ and note that

$$\lceil \log(K) \rceil \exp\left(\frac{-2T\varepsilon^2}{\lceil \log(K) \rceil}\right) = \exp\left(\frac{-2T(\varepsilon^2 - \varepsilon_0^2)}{\lceil \log(K) \rceil}\right) \ .$$

We can now integrate over the probabilities to bound expected simple regret,

$$
\begin{aligned}
r(T) &\leq \varepsilon_0 + \int_{\varepsilon_0}^{+\infty} \exp\left(-2(\varepsilon^2 - \varepsilon_0^2)\frac{T}{\lceil \log(K) \rceil}\right) d\varepsilon \,, \\
&\leq \varepsilon_0 + \int_{0}^{+\infty} \exp\left(-2((\varepsilon + \varepsilon_0)^2 - \varepsilon_0^2)\frac{T}{\lceil \log(K) \rceil}\right) d\varepsilon \,, \\
&\leq \varepsilon_0 + \int_{0}^{+\infty} \exp\left(-2(\varepsilon^2 + 2\varepsilon\varepsilon_0)\frac{T}{\lceil \log(K) \rceil}\right) d\varepsilon \,, \\
&\leq \varepsilon_0 + \int_{0}^{+\infty} \exp\left(-2\varepsilon^2\frac{T}{\lceil \log(K) \rceil}\right) d\varepsilon \,, \\
&= \sqrt{\frac{\lceil \log(K) \rceil \log\lceil \log(K) \rceil}{2T}} + c'\sqrt{\frac{\lceil \log(K) \rceil}{T}} \,, \\
&\leq c\sqrt{\frac{\log(K)(\log\log(K) \vee 1)}{T}} \,.
\end{aligned}
$$

where $c, c' > 0$ are absolute constants. □

Auto correcting binary search Exploiting the monotone constraint with a naive binary search, is already a substantial improvement on the unconstrained TBP, however, there is potential for further improvement, specifically where it concerns the $\log\log K$ term. Essentially, the reason we pay a $\log\log(K)$, is that, if the binary search makes a single mistake at any of it's $\log K$ steps and progresses to the incorrect node, we have no guarantee on our regret. Therefore we need to apply a union bound to ensure the algorithm makes the correct decision at each of its $\log(K)$ steps, this then leads to the additional $\log\log(K)$ term in our upper bound. If one were to allow for mistakes in the binary search, the union bound would become unnecessary and the $\log\log K$ term could be removed. There are clear hints to this also in the literature and a technique used by many is a binary search with corrections, [33], [8], and [30], see also [53]. However, not only are the above papers in the fixed confidence setting, they also consider severely restrictive structural assumptions and their results are not applicable to our setting. In [53], the authors consider a setting where each arm $a \in \mathcal{A}$ is restricted to follow a Bernoulli distribution with parameter p_a such that,

$$p_1 < p_2 < ... < p_K \,,$$

we denote the set of such problems, $\mathfrak{B}_m$, that is,

$$\mathfrak{B}_m := \mathfrak{B} \cap \mathcal{B}_m \,.$$

Their results are in the fixed confidence setting, in that for a fixed $\delta, \varepsilon > 0$ they provide a PAC(δ, ε) algorithm, see Equation (1.5), with an upper bound on its expected stopping time. To remind the reader, in this setting, for a problem $\nu \in \mathcal{B}$, an algorithm is PAC(δ, ε), if it correctly classifies all arms of distance greater than ε to the threshold, with probability greater than $1 - \delta$. For a specific $\varepsilon = 1/3$ and $\tau = 1/2$, they propose the following algorithm `Auto-correcting Binary Search`.

At each step i, instead of only sampling the arm $v_i(m)$, the `Auto-correcting Binary Search` also samples the arms $v_i(l)$ and $v_i(r)$. If an inconsistency is detected then the algorithm backtracks to the parent node. At each step, with probability $1/\log(K)$, it performs a confirmation check as to whether the algorithm should terminate on the current node. The expected number of steps to perform a confirmation

```
Initialise: $k = \lceil 300 \log(K) \rceil$
for $i = 1, 2, \ldots$ do
    Sample arm $v_i(l)$, two times, let $X^l_{1,i}, X^l_{2,i}$ denote the samples.
    Sample arm $v_i(r)$, two times, let $X^r_{1,i}, X^r_{2,i}$ denote the samples.
    if $X^l_{1,i} \wedge X^l_{2,i} > 0$ or $X^r_{1,i} \vee X^r_{2,i} < 1$ then
        $v_{i+1} = P(v_i)$
    end
    else
        Sample arms $v_i(m)$, once, let $X_{m,i}$ denote the sample.
        if $X_{m,i} = 0$ then
            $v_{i+1} = R(v_i)$
        end
        if $X_{m,i} = 1$ then
            $v_{i+1} = L(v_i)$
        end
    end
    Draw single sample $U_i$ from Uniform$(0, 1)$
    if $U_i \leq 1/\log(K)$ then
        Sample arm $v_i(r)$, $k$ times, let $\widehat{\mu}_{i,r}$ denote sample mean.
        Sample arm $v_i(l)$, $k$ times, let $\widehat{\mu}_{i,l}$ denote sample mean.
        if $\widehat{\mu}_{i,l} \in [1/4, 3/4]$ then
            Output: $\widehat{Q} = 2\mathbb{1}_{a > v_i(l)} - 1$
        end
        if $\widehat{\mu}_{i,r} \in [1/4, 3/4]$ then
            Output: $\widehat{Q} = 2\mathbb{1}_{a \geq v_i(r)} - 1$
        end
        if $r = l + 1$ and $\widehat{\mu}_{i,l} < \frac{1}{2} < \widehat{\mu}_{i,r}$ then
            Output: $\widehat{Q} = 2\mathbb{1}_{a \geq v_i(r)} - 1$
        end
    end
end
```

Algorithm 9: Auto correcting binary search

check is $\log(K)$, the depth of tree. The performance of the algorithm is bounded in the following theorem.

Theorem 20 (Combination of Propositions 5.1, 3.1 [53]). *Setting $\varepsilon_0 = 1/3, \delta_0 = 1/4$ and threshold $\tau = 1/2$, for all $\nu \in \mathfrak{B}_m$, the* `Auto-correcting Binary Search` *algorithm is PAC$(\varepsilon_0, \delta_0)$ and its stopping time $\tilde{\tau}$ satisfies,*

$$\mathbb{E}[\tilde{\tau}] \leq c \log(K) \ ,$$

for some absolute constant $c > 0$. Furthermore, given $\delta > 0$, there exists an algorithm which, for all $\varepsilon > 0$ and $\nu \in \mathfrak{B}_m$, is PAC(ε, δ), with stopping time $\tilde{\tau}'$ satisfying,

$$\mathbb{E}\left[\tilde{\tau}'\right] \leq c' \log\log(\varepsilon) \log(\delta) \log(K)/\varepsilon^2 \ ,$$

for some absolute constant $c' > 0$.

Remark 5. The proof of the second statement of Theorem 20, involves a slightly modified version of the `Auto-correcting Binary Search` algorithm, which, instead of sampling arms a constant number of times at each step, tunes the number of samples according to δ.

If one wished to translate the result of Theorem 20 to our fixed budget setting, there are two non trivial issues.

- With an auto correcting binary search, the main issue is knowing when to stop and output an arm as the point at which the arms cross the threshold. The `Auto-correcting Binary Search` of [53] gets around this by performing a confirmation check at each step with a certain probability. However, this technique relies on the fact that they only bound the stopping time in expectation, and not with high probability, and is therefore not suitable in the fixed budget setting.
- The second result of Theorem 20 only holds for a fixed δ, such a result is not suitable to bound simple regret in our fixed budget setting, as one would need a bound on all δ simultaneously, see the proof of Theorem 19.

In [8] they describe a similar result to that of [53], for $\delta, \varepsilon > 0$, they provide a PAC(δ, ε) algorithm with expected stopping time bounded above by $c(1-\delta)\log(K)/\varepsilon^2$, with $c > 0$ an absolute constant, however, as with [53], their algorithm takes δ as a parameter and therefore, the same issue remains when translating their results to our setting, where a more nuanced treatment will be needed.

Contribution

In [26] we describe an algorithm, the `MTB`, which is an auto correcting binary search, working under the more general assumption of distributions with supports bounded on $[0, 1]$, with a novel method of choosing a cutting point. Our analysis is also novel and allows us to provide bounds holding simultaneously across all probabilities. We show that for all problems $\nu \in \mathcal{B}_m$, the expected simple regret of our algorithm is upper bounded as,

$$\bar{r}^{\nu}_{\texttt{MTB}}(T) \leq c\sqrt{\frac{\log(K)}{T}} \ ,$$

where $c > 0$ is an absolute constant, see Corollary 5. We also prove that, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}_m$, such that,

$$\bar{r}^{\nu}_{\pi}(T) \geq c'\sqrt{\frac{\log(K)}{T}} \ ,$$

where c' is an absolute constant, see Proposition 5. Thus, we identify the minimax rate for expected simple regret, on the set of monotone bandit problems,

$$\bar{r}^{*}_{T}(\mathcal{B}_m) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}_m} \bar{r}^{\nu}_{\pi}(T) \ ,$$

as of the order,

$$\sqrt{\frac{\log(K)}{T}} \ .$$

see Theorem 24.

1.6.2 Problem independent TBP under concave and unimodal constraints

Alternative to the monotone constraint on the means, we also consider a concave constraint and a uni-modal constraint. For the concave constraint define the class of bandit problems,

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

To the best of our knowledge, such a setting had not yet been considered in the literature.

Contribution

For the TBP under a concave constraint we describe an algorithm, the CTB, which utilises an auto corrective binary search but now on a well chosen $\log(K)$ sized subset of the arms. We show that for all problems $\nu \in \mathcal{B}_c$, the expected simple regret of our algorithm is upper bounded as,

$$\bar{r}^{\nu}_{\text{CTB}}(T) \leq c\sqrt{\frac{\log\log(K)}{T}} ,$$

where $c > 0$ is an absolute constant, see Proposition 11. We also prove that, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}_c$, such that,

$$\bar{r}^{\nu}_{\pi}(T) \geq c'\sqrt{\frac{\log\log(K)}{T}} ,$$

where c' is an absolute constant, see Proposition 10. Thus, we identify the minimax rate for expected simple regret, on the set of concave bandit problems,

$$\bar{r}^{*}_{T}(\mathcal{B}_c) := \inf_{\pi\in\mathcal{C}} \sup_{\nu\in\mathcal{B}_c} \bar{r}^{\nu}_{\pi}(T) ,$$

as of the order,

$$\sqrt{\frac{\log\log(K)}{T}} ,$$

see Theorem 26. Here, our result is tied to the distributions of the arms being supported on $[0, 1]$. The related settings of optimisation or estimation of a convex function, are almost exclusively studied in the continuous setting and focus on achieving good dependency in the dimension d.

For the unimodal constraint define the class of problems,

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

Contribution

For the TBP under a unimodal constraint, we provide an algorithm, the UTB, and for all $\nu \in \mathcal{B}_u$, show that,

$$\bar{r}^{\nu}_{\texttt{UTB}}(T) \leq c\sqrt{\frac{K}{T}} \, ,$$

where $c > 0$ is an absolute constant, see Proposition 9. For the unimodal constraint, if we were to know the location of an optimal arm, i.e. one in $\arg\max_{a \in \mathcal{A}}(\mu_a)$, we could split the arm set into a monotonically increasing set and monotonically decreasing set, on which we could then run the MTB[5]. The strategy of the UTB algorithm, is therefore, to first identify an optimal arm, $\hat{a}$, and then run the MTB on the arm sets $[1, \hat{a}]$ and $[\hat{a}, K]$. To output a prediction of an optimal arm it utilises a well known approach, see [12] and Theorem 11. The minimax rate for BAI under simple regret is of the order $\sqrt{\frac{K}{T}}$, see Equation (1.10) and therefore, this is the dominating term in the regret of the UTB.

We also prove that, for any policy $\pi \in \mathcal{C}$, there exists a problem $\nu \in \mathcal{B}_u$, such that,

$$\bar{r}^{\nu}_{\pi}(T) \geq c'\sqrt{\frac{K}{T}} \, ,$$

where c' is an absolute constant, see Proposition 8. Thus, we identify the minimax rate for expected simple regret, on the set of unimodal bandit problems,

$$\bar{r}^{*}_{T}(\mathcal{B}_u) := \inf_{\pi \in \mathcal{C}} \sup_{\nu \in \mathcal{B}_u} \bar{r}^{\nu}_{\pi}(T) \, ,$$

as of the order,

$$\sqrt{\frac{K}{T}} \, ,$$

see Theorem 25. Much of the related literature concerns the problem dependent setting, [85] and [69] are in the problem independent setting, however, they work in the continuous case and have smoothness assumptions on the arm means around the optimal arm, which does not translate to our discrete setting, where one can have jumps between the means of the arms.

Before moving on to the shape constrained TBP in the problem dependent setting, let us present our identified minimax rates for simple regret, under each of the considered constraints, see Table 1.1.

1.6.3 Problem dependent rates for the shape constrained TBP

Our work in [26] was from the problem independent perspective, in that our rates did not depend on the distance of the arms to the threshold. In the paper [25], co authored with Pierre Menard and Alexandra Carpentier, we again study the thresholding bandit problem under shape constraints but now from a problem dependent perspective. In such a setting, probability of error is a more suitable measure of the learner's performance than expected simple regret. For a sequence of gaps $\bar{\Delta} \in [0, 1]^K$, define the set of bandit problems,

[5] So far we have bounded the regret of the MTB on monotonically increasing sequences of arms, however, it can be trivially modified to give identical guarantees under a monotonically decreasing constraint, see DEC-MTB.

Results	Unstructured *TBP*	Monotone *TBP*	Unimodal *TBP*	Concave *TBP*
Minimax rate	$\sqrt{\frac{K\log K}{T}}$	$\sqrt{\frac{\log K}{T}}$	$\sqrt{\frac{K}{T}}$	$\sqrt{\frac{\log\log K}{T}}$

TABLE 1.1: Order of the minimax rate for expected simple regret in TBP, $\bar{r}(T)$, see Equation (1.18), in the case of all four structural assumptions, on the means of the arms, considered in this thesis. All results are given up to universal multiplicative constants.

$$\mathcal{B}_m^{\bar{\Delta}} = \{\nu \in \mathcal{B}_m : \forall a \in \mathcal{A},\ |\mu_a - \tau| = \bar{\Delta}_a\} \ .$$

Contribution

In the monotone setting we describe an algorithm, ProbDep-Explore-again based on an auto correcting binary search but with several key differences, that, for a sequence of gaps $\bar{\Delta} \in [0,1]^K$ and $T > 36\log K$, for all problems $\nu \in \mathcal{B}_m^{\bar{\Delta}}$, satisfies,

$$\bar{e}^{\nu}_{\texttt{ProbDep-Explore}}(T) \leq \exp(-cT\min(\bar{\Delta})^2 + c'\log(K)) \ , \tag{1.23}$$

where $c, c' > 0$ are absolute constants, see Theorem 28. We also prove a lower bound, such that for any sequence of gaps $\bar{\Delta}$ and any policy π, there exist a problem $\nu \in \mathcal{B}_m^{\bar{\Delta}}$, such that,

$$\bar{e}^{\nu}_{\pi}(T) \geq \exp(-c''T\min(\bar{\Delta})^2) \ ,$$

where $c'' > 0$ is an absolute constant, see Theorem 27.

One will notice that for $T \geq c\frac{\log(K)}{\min\Delta^2}$, for some absolute constant $c > 0$, our bound (1.23) matches the result of Theorem 2, for BAI in the 2 armed setting. Essentially this means, that in this regime, the TBP is no harder than classifying the single arm with minimal gap as above or below threshold, a result that was surprising to us! Furthermore the learning rate for the concave setting is also of the same order. Precisely, for a sequence of gaps $\bar{\Delta} \in [0,1]^K$, define the following set of problems,

$$\mathcal{B}_c^{\bar{\Delta}} := \left\{\nu \in \mathcal{B}_c : \forall a \in \mathcal{A}, |\mu_a - \tau| \in \left[\frac{\bar{\Delta}_a}{2}, 3\frac{\bar{\Delta}_a}{2}\right]\right\} .$$

Contribution

In the concave setting, we provide an algorithm, ProbDep-CTB, that, for a sequence of gaps $\bar{\Delta} \in [0,1]^K$ and $T > 108\log(K)$, for all problems $\nu \in \mathcal{B}_c^{\bar{\Delta}}$, satisfies,

$$\bar{e}^{\nu}_{\texttt{ProbDep-CTB}}(T) \geq 3\exp(-cT\min(\Delta_k)^2 + c'\log(K)) \ ,$$

where $c, c' > 0$ are absolute constants, see Theorem 30. We also demonstrate a lower bound, such that for any sequence of gaps $\bar{\Delta} \in [0,1]^K$, and any policy π, there exist a problem $\nu \in \mathcal{B}_c^{\bar{\Delta}}$ such that on problem ν algorithm π has the following lower bound on its probability of error,

$$\bar{e}_\pi^\nu(T) \geq \exp(-c'' T \min(\Delta_k)^2) \ .$$

where $c'' > 0$ is an absolute constant, see Theorem 30.

The lack of difference in rates between the monotone and concave constraints is in sharp contrast to the problem independent setting where the rates for the monotone and concave constraints differ.

1.7 Best arm identification under many optimal arms

We now turn to a slightly different topic, that of best arm identification, under the possibility of many optimal arms. In the classical setting of best arm identification one often assumes a single optimal arm, e.g. [2], [48], however, in many practical applications this is simply not the case. For example, when recommending a film to a user there will, most likely, be multiple potential titles they would respond favourably to. When there are many optimal arms, one would expect the learner to be able to leverage this to their advantage. Classical regret bounds for best arm identification scale poorly with many optimal arms, we recall the bound of [2], Theorem 13,

$$e(T) < \exp\left(-\frac{cT}{\sum_{a:\mu_k \neq \mu^*} \Delta_a^{-2}}\right) ,$$

with $c > 0$ an absolute constant. Here, even if half the arms were optimal, we would see an improvement only in the constant term in the exponential. Furthermore, if we fix the proportion of optimal arms to a constant fraction, say $\frac{1}{2}$, and let K grow to infinity, the bound quickly becomes nonsensical. One would hope that with a large, or even infinite set of arms, if a large proportion are optimal, the learner should still be able to recover comparable bounds on their regret. With this in mind we will introduce a new formulation of the MAB, one that extends beyond the K armed bandit.

1.7.1 Bandits with an infinite reservoir

We consider a setting with a (potentially infinite) set of arms $\mathcal{A}$, which we call the *reservoir*. Each arm $a \in \mathcal{A}$ is associated with a probability distribution ν_a, which we assume to be supported on $[0, 1]$, and we denote its mean by μ_a. Again, write $\mu^* = \sup_{a \in \mathcal{A}} \mu_a$ for the highest mean and write,

$$\mu_{sub} = \sup_{a \in \mathcal{A}:\mu_a \neq \mu^*} \mu_a \ ,$$

for the mean of the largest non optimal arm. We further assume that there exists a partition $\mathcal{A} = \mathcal{A}^* \cup \mathcal{A}_{sub}$ such that each arm $a \in \mathcal{A}^*$ is optimal, i.e.

$$\forall a \in \mathcal{A}, \mu_a = \mu^* \ ,$$

and each arm $a \in \mathcal{A}_{sub}$ strictly is sub-optimal, i.e.

$$\forall a \in \mathcal{A}_{sub}, \mu_a < \mu_{sub} \ .$$

We assume that the agent can pick arms at random from the reservoir, according to some distribution on $\mathcal{A}$, which we shall denote $\bar{\mathbb{P}}_{\mathcal{A}}$. The arms $\mathcal{A}^*$ form a $p^\star$ proportion of the reservoir, that is, for an arm a drawn from the reservoir,

$$\bar{\mathbb{P}}_{\mathcal{A}}(a \in \mathcal{A}^*) = p^\star \ ,$$

and,

$$\bar{\mathbb{P}}_{\mathcal{A}}(a \in \mathcal{A}_{sub}) = 1 - p^\star \ ,$$

i.e. this arm belongs either to the set $\mathcal{A}^*$ with probability $p^\star$, or it belongs to the set $\mathcal{A}_{sub}$ with probability $1 - p^\star$. We again fix the time horizon at T and the learner interacts with the environment in several rounds $t = 1, 2, \ldots, T$. At each round $t \leq T$, the learner chooses an arm a_t by either picking a new arm from the reservoir $\mathcal{A}$, or playing a past arm and gets a reward $Y_t \sim \nu_{a_t}$. As before, the arm choice depends only on the past observations, the past arm choices, and possibly some exogenous randomness. The rewards for each arm a are i.i.d. random variables with mean μ_a, unknown to the learner. Let $\mathfrak{D}$ denote the global class of bandit problems with a potentially infinite reservoir $\mathcal{A}$, with associated distribution $\bar{\mathbb{P}}_{\mathcal{A}}$, admitting partition, $\mathcal{A} = \mathcal{A}^* \cup \mathcal{A}_{sub}$.

We will again work in the problem dependent regime, we denote

$$\Delta_{min} = \mu^* - \mu_{sub} \ ,$$

for the associated minimal gap. For larger Δ_{min} the optimal arms are further from the rest and therefore easier to identify. For $\Delta > 0$, $p \in (0, 1]$ we let $\mathfrak{D}_{\Delta,p}$, define the set of bandit problems whose reservoir distribution is such that $p^\star \geq p$ and $\Delta_{min} \geq \Delta$, that is,

$$\mathfrak{D}_{\Delta,p} := \{\nu \in \mathfrak{D} : p^\star \geq p, |\mu^* - \mu_{sub}| \geq \Delta\} \ .$$

Ideally, what we would hope to achieve in this setting is something comparable to the classical bound for BAI with K arms, see Theorem 13, but with the dependency on K replaced with one on $p^\star$.

Literature relating to BAI in the infinite reservoir setting under fixed budget Formulating the MAB with a potentially infinite reservoir is well studied in the fixed budget literature. A classical assumption in this setting is that, for $\Delta > 0$, the proportion of Δ - near optimal arms is of order Δ^α for some $\alpha > 0$. This was first considered by [9] with further work by [83] and [10], all for cumulative regret, see Section 1.8 for further discussion of their results. In the case of BAI, under the assumption that, for $\Delta > 0$, the proportion of Δ - near optimal arms is of proportion larger than Δ^α for some $\alpha > 0$, [18] identify the minimax rate as of the order,

$$T^{-1/2} \vee T^{-1/\alpha} \ ,$$

up to a $\log\log(T)$ term. Their assumption on the reservoir is much weaker than assuming a fixed proportion of optimal arms $p^\star$, however, their rates are also considerably weaker than what we would hope for. They essentially identify a sub optimal arm, whose distance to the optimal is bounded polynomially with T, while we would aim for an exponential decay in the regret with respect to T.

1.7.2 BAI for for bandits with infinite reservoir, under fixed confidence

While much of the fixed budget literature differs greatly to our setting, in terms of their assumptions on the reservoir, in the fixed confidence regime there are several works much closer to our own. These works consider various settings, which can all be related to BAI. We will address each setting in turn.

Quantile estimation In the works of [7, 20] they are in the fixed confidence setting and their objective is to identify an arm $\hat{a}$ with mean greater than a quantile of known order, with respect to the reservoir distribution, with high probability. For simplicity, when covering their results we will restrict to the setting where for all arms $a \in \mathcal{A}$, ν_a is Bernoulli. In [7] their results extend beyond this assumption to general reward distributions on the arms. For some $i > 0$, let $\mathcal{G}_i$ denote the ipth quantile, that is, for an arm a drawn from the reservoir,

$$\mathcal{G}_i = \sup\big(x : \bar{\mathbb{P}}_{\mathcal{A}}(\mu_a \geq x) \geq ip \wedge 1\big) \, .$$

Given $p > 0$, we can then measure the learners regret as,

$$(\mathcal{G}_1 - \mu_{\hat{a}})_+ \, .$$

The objective of the learner is then, for $\delta, \varepsilon > 0$, to describe a PAC(ε, δ), see Equation (1.5), algorithm with minimal expected stopping time. In [7] they propose a method that first draws an appropriately sized sample from the reservoir, and then proceeds to run the KL-LUCB algorithm. Essentially, the KL-LUCB algorithm is to the `LUCB` what the KL-UCB is to the `UCB`, in that its confidence bounds are based on the KL divergence as opposed to Hoeffdings. We will treat the KL-LUCB as a black box, see [59] for a detailed analysis. For a given $p, \varepsilon, \delta > 0$, the algorithm of [7] is then as follows, see (p, δ, ε)-`KL-LUCB`. The (p, δ, ε)-`KL-LUCB` first draws a sub sample of size $\frac{1}{p}\log(\frac{2}{\delta})$ from the reservoir. Due to its well chosen size, there will be an arm with mean above the pth quantile contained in said sample, with probability roughly greater than $1-\delta$. One can then run known techniques for finite bandits, in this case the KL-LUCB, and maintain PAC(δ, ε). To understand the rates of [7] we must understand their definition of problem complexity. Let $m = \lceil \frac{1}{p} \rceil$, they then define the following problem complexity,

```
Input: target quantile $p$, confidence level $\delta$, tolerance $\varepsilon$
Draw a $\frac{1}{p}\log(\frac{2}{\delta})$ sized sub sample
Run KL-LUCB on sub sample
```

Algorithm 10: (p, δ, ε)-`KL-LUCB`

$$H_{p,\varepsilon} := \frac{1}{\varepsilon^2} + \sum_{i=2}^{m} \frac{1}{\max(\varepsilon^2/2, d^*(\mathcal{G}_i, \mathcal{G}_1))} \, ,$$

where, for $d, p \in [0, 1]$, $d^*(p, q)$ is the Chernoff information and is equal to $\mathrm{KL}(\mathcal{B}er(z^*), \mathcal{B}er(p))$ where z^* is the unique solution in z to $\mathrm{KL}(\mathcal{B}er(z), \mathcal{B}er(p)) = \mathrm{KL}(\mathcal{B}er(z), \mathcal{B}er(q))$. Naturally the Chernoff information is closely related to the KL divergence and also the arm gaps, for an arm $a \in \mathcal{A}$ we have,

$$\frac{\Delta_a^2}{2} \leq d^*(\mu_a, \mu^*) \leq \mathrm{KL}(\mathcal{B}er(\mu_a), \mathcal{B}er(\mu^*)) \leq \frac{\Delta_a^2}{\mu^*(1-\mu^*)} \, .$$

The complexity $H_{p,\varepsilon}$ can then be seen as somewhat analogous to the complexity H in the finite armed case, see equation (1.11). There is the additional effect that, when the confidence level ε becomes greater than an arm's gap, that gap's effect on the complexity disappears.

Theorem 21 (Theorem 6 [7])**.** *Let $\delta < p < 1/3$, on all problems $\nu \in \mathfrak{D}$, such that $\forall a \in \mathcal{A}$, ν_a is Bernoulli, the (p, δ, ε)-KL-LUCB algorithm is PAC(δ, ε), that is,*

$$\mathbb{P}_{\nu,(p,\delta,\varepsilon)\text{-}\mathtt{KL}\text{-}\mathtt{LUCB}}((\mathcal{G}_1 - \mu_{\hat{a}})_+ \geq \varepsilon) \leq \delta \ ,$$

and furthermore, with probability greater than $1 - 7\delta$, it's stopping time τ is upper bounded as follows,

$$\tau \leq cH_{p,\varepsilon} \log(1/\delta)^2 \ ,$$

for some absolute constant $c > 0$.

Remark 6. In [7] the authors extend the result of Theorem 21 to the one parameter exponential family of distributions.

For $\varepsilon = 0$[6] and target quantile $p > 0$, in [7], they also provide a lower bound on the expected stopping time of any PAC$(\delta, 0)$ algorithm of the order,

$$\frac{c'}{\mathrm{KL}(\mathcal{B}er(\mu^*), \mathcal{B}er(\mathcal{G}_2))} + \log(c/\delta) \sum_{i=2}^{m-1} \frac{1}{\mathrm{KL}(\mathcal{B}er(\mathcal{G}_i), \mathcal{B}er(\mu^*))} \ , \tag{1.24}$$

for some absolute constants $c', c > 0$.

Remark 7. The authors of [7] extend the lower bound of Equation (1.24) to distributions continuously parameterised by their mean, with some additional assumptions, typical for the one parameter exponential family of distributions, see [7][Assumption 1].

Note the $\log(1/\delta)$ discrepancy in the bounds of Theorem 21 and Equation (1.24). In the case where $\forall a \in \mathcal{A}_{sub}, \mu_a = \mu_{sub}$, we have $H_{p,\varepsilon} = \frac{1}{\varepsilon^2} + \frac{m}{\max(\varepsilon^2/2, d^*(\mathcal{G}_2, \mathcal{G}_1))}$. Thus, in the aforementioned case, taking $\varepsilon = \Delta_{min}$ and $p = p^*$, the bound of Theorem 21, roughly translates to our fixed budget setting as,

$$\exp\left(-c\sqrt{Tp^*\Delta_{min}^2}\right) \ ,$$

for some absolute constant $c > 0$. This is far larger than what we would hope to achieve. Essentially, the additional multiplicative $\log(1/\delta)$ in Theorem 21 is sub optimal, this also highlighted in gap between the UB and LB in [7]. The authors therein wonder if the additional multiplicative $\log(1/\delta)$ term is necessary for such two stage algorithms. They note that in the first stage, clearly one cannot avoid drawing at least $c\frac{1}{p}\log(\frac{1}{\delta})$ arms from the reservoir, for some absoulte constant $c > 0$. However, in the second phase one only needs to find an arm in the top p fraction, and in expectation there should be roughly $\log(1/\delta)$ such arms in the sub sample. One should, therefore, be able to do better than the standard rates for BAI in this case.

Epsilon good arm identification Also directly comparable is the problem of ε-good arm identification. This can be seen as the PAC(δ, ε), see Equation (1.5), setting

[6] Their lower bound also extends to the $\varepsilon > 0$ case, see [7][Remark 4]

for BAI under simple regret. That is, given a prediction of the learner $\hat{a}$, we consider the simple regret,

$$|\mu_{\hat{a}} - \mu^*| \,,$$

and, for $\delta, \varepsilon > 0$, the aim of the learner is to provide a PAC(δ, ε) algorithm with minimal stopping time. Let us write $p^\star_\varepsilon$ as the proportion of epsilon good arms, that is, for an arm a drawn from the reservoir,

$$\bar{\mathbb{P}}_{\mathcal{A}}(\mu^* - \mu_a \leq \varepsilon) = p^\star_\varepsilon \,.$$

We now take the gap as the distance between the worst epsilon good arm and the best non-epsilon good arm, that is, we define,

$$\Delta_{<\varepsilon>} := \inf_{a:\mu_a \geq \mu^* - \varepsilon} (\mu_a) - \sup_{a:\mu_a < \mu^* - \varepsilon} (\mu_a) \,.$$

The paper [55] considers such a setting and proposes a UCB type algorithm that runs over increasingly large brackets of arms taken from the reservoir. For a given arm a at time t, with confidence level $\delta > 0$, they define their UCB index as follows,

$$\widetilde{\mathrm{UCB}}_\delta(a, t) = \widehat{\mu}_{a,t} + c\sqrt{\frac{\log(\log(N_a(t))/\delta)}{N_a(t)}} \,,$$

for some absolute constant $c > 0$. Their algorithm is then as follows, see BUCB.

As mentioned, if the learner does not know ε or the proportion of ε good arms, they cannot tune the size of a single sub sample. The BUCB algorithm overcomes this by running over multiple sizes of sub sample simultaneously, drawing increasingly larger brackets of arms from the reservoir as it runs. That is, for $l = 1, 2, 3..$ at time $l2^l$ it opens a new bracket $\mathcal{A}_l$ of size 2^l. In the meantime it runs its UCB type algorithm over all open brackets. It is important to note that the BUCB algorithm does not have a explicit stopping time, instead the user must decide a time $t > 0$ at which to stop and take output O_t. With this in mind, the performance of the BUCB is upper bounded in the following theorem.

```
Input: confidence level $\delta > 0$
Initialise: $l = 0$, $R_0 = 0$
for $t = 1, 2, ...$ do
    if $t > 2^l l$ then
        Draw sample $\mathcal{A}_{l+1}$ of size $2^l$ from reservoir
        $l = l + 1$
    end
    $R_t = R_{t-1}\mathbb{1}_{R_t<l} + 1$
    Pull arm $a_t = \arg\max_{a\in\mathcal{A}_{R_t}} \left(\widetilde{\mathrm{UCB}}_\delta(a,t)\right)$
    $O_t = \arg\max_{a\in\mathcal{A}_r \text{for some } r\leq l} \left(\widetilde{\mathrm{UCB}}_{\delta/(|A_r|r^2)}(a,t)\right)$
end
```

Algorithm 11: Bracketing UCB (BUCB)

Theorem 22 (Specific version Theorem 2 [55])**.** *Let $\delta < 0.025$, $\varepsilon > 0$, their exists a stopping time τ such that, for all problems $\nu \in \mathfrak{D}$, when running the* BUCB*,*

$$\mathbb{P}(\exists s > \tau : O_s < \mu^* - \varepsilon) \leq 2\delta \ ,$$

and furthermore,

$$\mathbb{E}[\tau] \leq c\tilde{H}_\delta \log(\tilde{H}_\delta) \ ,$$

where $\tilde{H}_\delta = \frac{1}{p^*_\varepsilon \Delta^2_{<\varepsilon>}} \log(1/\delta)$ and $c > 0$ is an absolute constant.

At a glance, Theorem 22 appears to be a clear improvement over Theorem 21, as the additional multiplicative $\log(1/\delta)$ term is removed and the BUCB algorithm is adaptive on ε. However, we see Theorem 22 is a fundamentally different result and not comparable to Theorem 21 for two key reasons.

- Firstly, the stopping time τ such that $\mathbb{E}[\tau] \leq c\tilde{H}_\delta \log(\tilde{H}_\delta)$, for some absolute constant $c > 0$, is not explicitly stated and indeed the stopping time used in the proof requires knowledge of the true means of the arms, information not available to the learner.

- Furthermore the main result in Theorem 22 is in expectation, not high probability. This is a far weaker result and means that Theorem 22 cannot be directly applied to the fixed budget setting. If the authors of [55] wished to obtain a result which holds in high probability they would need to sample far more arms and would end up paying the $\log(1/\delta)^2$. To try and build some intuition for this statement, consider $\tilde{l} > 1 : \frac{1}{p^*_\varepsilon} = \tilde{l}2^{\tilde{l}}$. Thus when the BUCB algorithm opens a bracket of size $\tilde{l}2^{\tilde{l}}$ one can *expect* to have at least one epsilon good arm in said bracket. However, if one wanted to have an epsilon good arm in the bracket with *probability greater than* δ, one would need to open a much larger bracket, of size $\log(1/\delta)\tilde{l}2^{\tilde{l}}$. See also Remark 4 in [56] and page 15 in the appendix of the full version [55].

This is not to disparage the work of [55], the focus of their paper is instead to get more complete gap dependent bounds, considering also the gaps within the epsilon good arms.

Most biased coin problem The most biased coin problem is a restricted version of our setting where,

$$\forall a \in \mathcal{A}^*, \nu_a = \mathcal{B}\text{er}(\mu^*) \ ,$$

that is, all optimal arms follow a $\mathcal{B}\text{er}(\mu^*)$ distribution and furthermore,

$$\forall a \in \mathcal{A}_{sub}, \nu_a = \mathcal{B}\text{er}(\mu_{sub}) \ ,$$

that is, all sub optimal arms are distributed according the same $\mathcal{B}\text{er}(\mu_{sub})$ distribution. The goal of the learner is then again to return an optimal arm, which in this setting can be seen as identifying "the most biased coin". This setting has been studied in the fixed confidence regime, see [19], [49]. Specifically, [49] provide an algorithm that, in the above setting, for $\delta > 0$ will, with probability at least $1 - \delta$ output an optimal arm and furthermore, with probability greater than $1 - \delta$, have it's stopping time upper bounded by,

$$c \log(1/(p^\star \Delta^2_{min})) \frac{\log(1/\delta)}{p^\star \Delta^2_{min}} \ ,$$

where $c > 0$ is an absolute constant. Translating this result to the fixed budget setting would yield a bound of the order,

$$\exp\left(-cTp^\star\Delta_{min}^2 / \log(1/(p^\star\Delta_{min}^2))\right) ,$$

for some absolute constant $c > 0$. This is very close to what we aim to achieve, however, specifying that all sub optimal arms must have the same distribution fundamentally changes the nature of the problem, as in this, restricted setting one is able to estimate Δ_{min}.

1.7.3 BAI under fixed proportion of many optimal arms in the fixed budget setting

The algorithms employed in the above literature typically approach the challenge of an infinite arm set by drawing a finite sub sample (or samples) from the reservoir and then using classical techniques for best arm identification on said sub sample(s). If we assume we have $p^\star$ proportion of optimal arms in the reservoir and, for some $\delta > 0$, draw a sub sample of order of size $\frac{\log(1/\delta)}{p^\star}$ we will, roughly speaking, have an optimal arm in our sub sample with probability greater than $1 - \delta$. Thus, if one were to know $p^\star, \Delta_{min}$ one could draw a suitably sized sub sample and run classical bandit algorithms. As we have seen in the fixed confidence literature there are two main issues with this,

- What if we do not know $p^\star, \Delta_{min}$, how do we tune the size of our sub sample?
- How do we exploit the fact that in expectation we will have multiple optimal arms in our sub sample?

In the paper [45] we overcome both these issues by utilising a successive elimination algorithm, see the SR and also the elimination algorithm of [52], with a novel modification.

Contribution

We show that for any $\Delta, p > 0$, on all problems in $\mathfrak{D}_{\Delta,p}$, our provided alglorithm, `Elimination`, has the following upper bound on its probability of error,

$$e^{\nu}_{\texttt{Elimination}}(T) \leq \exp\left(\frac{-cT\Delta^2 p}{\log(T)}\right) , \tag{1.25}$$

for some absolute constant $c > 0$, see Theorem 34. We also show a lower bound, such that for any policy π and $\Delta, p > 0$, there exists a bandit $\nu \in \mathfrak{D}_{\Delta,p}$, such that on problem ν, policy π has the following lower bound on its probability of error,

$$e^{\nu}_{\pi}(T) \geq \exp\left(-c'T\Delta^2 p\right) , \tag{1.26}$$

for some absolute constant $c' > 0$, see Theorem 35.

Without any knowledge of Δ_{min} or p^*, removing the $\log(T)$ term in our upper bound, for us, seems unfeasible. Perhaps, with an alternate algorithm not based on successive elimination, an improvement in the $\log T$ term would be possible, this is a question for future research.

1.8 Cumulative regret under many optimal arms

Following BAI, it is then natural to also consider cumulative regret minimisation under many optimal arms. The setting is as in the previous Section 1.7, the only difference being that we now consider the cumulative regret,

$$R(T) := T\mu^* - \mathbb{E}\left[\sum_{t<T} Y_t\right] ,$$

as opposed to probability of error. As in BAI, classical regret bounds for cumulative regret do not scale well under many optimal arms. Recall the upper bound of Theorem 7. For the K armed bandit, on all problems $\nu \in \mathcal{B}$,

$$R^{\nu}_{\mathtt{UCB}}(T) \leq c \sum_{a\in[K]} \Delta_a + c' \log(T) \sum_{a\in[K]} \frac{1}{\Delta_a} .$$

where $c, c' > 0$ are absolute constants, even if half the arms were optimal the bound would still be of the order $\log(T) \sum_{a\in\mathcal{A}_{sub}} \frac{1}{\Delta_a}$ and becomes nonsensical as we let K grow very large. As in BAI, what we would hope to achieve, is to replace the dependency on K with one on $p^\star$, leading to an ideal bound of the order,

$$\frac{\log(T)}{p^\star \Delta_{min}} .$$

To the best of our knowledge cumulative regret has not been studied in the specific case of an infinite arm set with $p^\star$ proportion of optimal arms, however, it has been studied in several related bandit settings with an infinite reservoir. A common assumption, first considered in [9] and then [83], is that there exists a $\beta > 0$ such that for $\varepsilon > 0$, for an arm a drawn from the reservoir,

$$\bar{\mathbb{P}}_{\mathcal{A}}(\mu_a \geq \mu^* - \varepsilon) = c\varepsilon^\beta , \tag{1.27}$$

where $c > 0$ is an absolute constant. The approach of [83] is to draw a finite sample of size K from the reservoir and then run a UCB variant on said sub sample. Their UCB variant is tuned to have a shorter than normal exploration phase to exploit the possibility of multiple near optimal arms in the sub sample. Assuming the learner knows β and μ^* the upper bound on the expected cumulative regret of their algorithm is of the order,

$$\log(T)(\sqrt{T} \vee T^{\beta/(1+\beta)}) ,$$

this is far worse than what we would hope to achieve, however, as in the case of [18] their assumptions are far weaker than our own. The authors in [29] also consider a infinitely armed bandit under very similar assumptions to [83]. Their tail assumption is slightly more general but contains (1.27). The main difference of their paper is that the rewards are no longer stochastic but are now deterministic. This means that after pulling an arm once the learner knows the exact value of its mean. In this setting, again assuming (1.27), they are able to achieve an upper bound on the expected cumulative regret of their algorithm of the order,

$$\log(T)(\sqrt{T} \vee T^{\beta/(1+\beta^2)}) ,$$

without knowledge of β or μ^*. As mentioned, the above papers consider much weaker assumptions on the reservoir and we would expect to significantly out perform their results. Our approach will be similar, at least in spirit, to that of [83], that is we will run a UCB variant on a appropriately sized sub sample. However, our particular variant of the UCB differs greatly to that of [83], as it must exploit our much stronger assumptions.

Contribution

Given $\Delta, p \in (0,1]$, we provide an algorithm, Sampling-UCB, and for all problems $\nu \in \mathfrak{D}_{\Delta,p}$, demonstrate the following upper bound on its regret,

$$R^{\nu}_{\texttt{Sampling-UCB}}(T) \leq \frac{c\log(T)\log(1/\Delta)}{p\Delta} \,,$$

where $c > 0$ is an absolute constant, see Theorem 31. We also prove that, given $\Delta, p \in (0,1]$, for any policy π, there exists a problem $\nu \in \mathfrak{D}_{\Delta,p}$, such that,

$$\mathbb{R}^{\nu}_{\pi}(T) \geq \frac{c'\log(T\Delta^2)}{p\Delta} \,,$$

where c' is an absolute constant, see Theorem 32. Importantly, Sampling-UCB does not require knowledge of Δ and although it does take p as a parameter, we show that is inevitable. Specifically, let $p \leq 1/4, \Delta > 0$, with $T \geq 4\left(\frac{c''\log(T)}{p\Delta^2}\right)^2$, with $c'' > 0$ an absolute constant. We show that, for any policy π, such that for all problems $\nu \in \mathfrak{D}_{\Delta,p}$,

$$R^{\nu}_{\pi}(T) \leq \frac{c''\log(T)}{p\Delta} \,,$$

for all $q \leq \frac{4p}{c''}$, there exists $\nu' \in \mathfrak{D}_{\Delta,q}$, such that,

$$R^{\nu'}_{\pi}(T) \geq \frac{\sqrt{T}\Delta}{4} \,,$$

see Theorem 33.

It is a common phenomenon that one is able to be adaptive to parameters in pure exploration problems while adaptive algorithms are impossible in the setting of cumulative regret. Indeed, if we are in a pure exploration bandit problem we can allocate a constant fraction of our budget to estimate parameters, if we were to do this in the cumulative regret setting, our regret would be linear in T.

Chapter 2

The Influence of Shape Constraints on the TBP

In this chapter we present the following work, "The Influence of Shape Constraints on the Thresholding Bandit Problem" [26], authored by James Cheshire, Pierre Ménard and Alexandra Carpentier.

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.

2.1 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. [23]. 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 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 [23, 24] -see also e.g. [32, 22, 78, 37] 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. [23], Locatelli, Gutzeit, and Carpentier [68], and Mukherjee et al. [71], Zhong, Huang, and Liu [87] - but also see e.g. [12, 2, 34, 17] 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 23 for a more thorough comparison to this literature, and Appendix 2.5.1 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 23. For a discussion on the performance of the uniform-sampling strategy in the *problem dependent* regime, see Appendix 2.5.1.

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. [39]. 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 [33]. 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. [84], 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, [33], Ben-Or and Hassidim [8] and Emamjomeh-Zadeh, Kempe, and Singhal [30] solve this issue by introducing a noisy binary search *with corrections* - see also [73], [53]. 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 Theorem 24 for a more thorough comparison to this literature and see Appendix 2.5.1 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

[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.

MTBP, specifically the difference coming from the monotone assumption, can be found in Appendix 2.5.1 and see the algorithm `Explore` and the associated text in Section 2.4 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 2.5.1.

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 [28, 27, 74, 85]. [74, 27] focus on the *problem dependent regime*, and are not transferable - at least to the best of our knowledge - to the problem independent setting. [85, 28] 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 25 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. [79], 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. [72][Chapter 9], and [82, 1, 66] to name a few, all of them in a continuous setting and in dimension d. The focus of this 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 26 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 $\sqrt{\text{loglog}(K)/T}$. 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.

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

TABLE 2.1: 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.

Organisation of the paper Our results are summarized in Table 2.1. See also Appendix 2.6.1 for an adaptation of these results in the $\mathcal{X}$-armed bandit setting. In Section 2.2 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 2.3. In Section 2.4 we describe algorithms attaining the minimax rates of Section 2.3, 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.

2.2 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 $\widehat{Q} \in \{-1, 1\}^K$ 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 $\hat{Q} := \hat{Q}^\pi$:

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

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 $R_T^{\underline{\nu},\pi} = \max_{\{k\in[K]:\ \widehat{Q}_k^\pi \neq Q_k\}} \Delta_k$. 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,

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

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,

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

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,

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

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

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

2.3 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 2.4. 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 $ca_n \le b_n \le Ca_n$.

Theorem 23 provides the minimax rate of the *TBP*. The proof can be found in Appendix 2.7.1, i.e. Proposition 2, and Proposition 4.

Theorem 23. *It holds that*

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

The algorithm **`Uniform`** *described in Sections 2.4 (see also Appendix 2.7.1) attains this rate.*

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 [68] an upper bound of order $\sqrt{K \log(K \log T/\delta)/T}$, and from [24] a lower bound of order $\sqrt{K/T}$, which are both slightly sub-optimal.

[2] In particular, independent of T, K.

Theorem 24 provides the minimax rate of the *MTBP*. The proof can be found in Appendix 2.7.2, i.e. Proposition 5, and Corollary 5.

Theorem 24. *It holds that*

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

The algorithm `MTB` *described in Section 2.4 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 [53]. 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 [8] 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 25 provides the minimax rate of the *UTBP*. The proof can be found in Appendix 2.7.3, i.e. Proposition 8, and Proposition 9.

Theorem 25. *It holds that*

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

The algorithm `UTB` *described in Section 2.4 attains this rate.*

Most related papers consider the problem dependent setting. However the papers [85, 28] 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 $T^{-\frac{\beta}{2\beta+1}}$, see Appendix 2.6.1. 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 [28]. And interestingly, a direct corollary to Theorem 24 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 2.6.1, which is very much aligned with the findings in [28].

Theorem 26 provides the minimax rate of the *CTBP*. The proof can be found in Appendix 2.7.4, i.e. Proposition 10, and Proposition 11.

Theorem 26. *It holds that*

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

The algorithm `CTB` *described in Section 2.4 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 26 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 2.6.1 for details.

In [79], 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 [72][Chapter 9], and [82, 1, 66] 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 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.

2.4 Minimax optimal algorithms

In this section we present algorithms that match minimax regret rates in Section 2.3 up to multiplicative constants for *TBP*, *MTBP*, *UTBP* and *CTBP*.

2.4.1 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 17 in Appendix 2.7.1.

Surprisingly the naive `Uniform` algorithm is optimal in the unstructured case with respect to the lower bound of Theorem 23. See the proof of Proposition 4 in Appendix 2.7.1. This contrasts with the related TOP-1 bandit problem where the minimax regret rate is $\sqrt{K/T}$, see [12, 3] for hints toward this. This is not very surprising as in the TOP-1 problem we are interested in finding one arm only, namely

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

the arm with highest mean, while in our problem we search for *all arms above threshold* and for this we pay an additional $\sqrt{\log K}$.

2.4.2 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 $\mu_{K+1} = +\infty$.

We introduce the `MTB` (Monotone Thresholding Bandits) algorithm. It is composed of two sub-algorithms, `Explore` and `Choose`. The first algorithm, `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 $\hat{a}$ which leads to the estimator $\widehat{Q}$, where $\widehat{Q}: \ \widehat{Q}[k] = -1 \ \forall k < \hat{a}, \ \widehat{Q}[k] = 1 \ \forall k \geq \hat{a}$.

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.

```
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
```

Algorithm 12: Explore

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 13.

```
Input: $\varepsilon$, $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \leq T_1}$, $(v_t)_{t \leq T_1}$
Initialization: $S_1 = [\,]$
for $t = 1 : T_1$ do
    $S_{t+1} = S_t$
    if $\{\exists k \in \{l, m, r\} : |\widehat{\mu}_{k,t} - \tau| \leq \varepsilon\} \vee$
     $\{k = v_t(r) = v_t(l) + 1;\ \hat{\mu}_{l,t} + \varepsilon < \tau \leq \hat{\mu}_{r,t} - \varepsilon\}$ 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$(S_{T_1+1})$.
```

Algorithm 13: Choose

Remark 8. 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,

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

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 24 in Appendix 2.7.2. Morally it allows to select a majority of "good" arms (i.e that provide a low regret classification $\widehat{Q}$) in S_{T_1+1} such that the median $\hat{a}$ is also a "good" arm, see Algorithm 14.

```
run algorithm Explore
    • Output: $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \le T_1}$, $(v_t)_{t \le T_1}$
run algorithm Choose
    • Input: $\hat{\varepsilon}$, $(\hat{\mu}_{l,t}, \hat{\mu}_{m,t}, \hat{\mu}_{r,t})_{t \le T_1}$, $(v_t)_{t \le T_1}$
    • Output: arm index $\hat{a}$
return $(\hat{a}, \widehat{Q})$ :    $\widehat{Q}_k = 2\mathbb{1}_{\{k \ge \hat{a}\}} - 1$
```

Algorithm 14: MTB

The MTB algorithm will achieve the minimax rate on expected simple regret given in Theorem 24, see the proof of Theorem 24, in Appendix 2.7.2, for details.

Remark 9 (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.

2.4.3 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 [12]. 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

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

where $c_{SR} > 0$ is a universal constant. Note that taking MOSS from [3] 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 $\widehat{\mu}_k$ 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 15.

This intuitively makes sense as $\hat{m}$ is an estimator of the maximum k^* of the mean sequence and unimodality implies that $(\mu_k)_{k \le k^*}$ is non-decreasing, and that $(\mu_k)_{k \ge k^*}$ 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.

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

Algorithm 15: UTB

2.4.4 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 $i < \tilde M$ 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 $6 \sum_{i \leq M} T_2^{(i)}(M) \leq T$. 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 $\widehat{\mu}_{p,i}$ 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

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

```
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.
    $l_{i+1} =$ output $\hat{k}$ of MTB with arms $\mathcal{S}^{\log}_{l_i,r_i}$, threshold $\tau_i$, budget $T_2^{(i)}$
    $r_{i+1} =$ output $\hat{k}$ of DEC-MTB with arms $-S^{\log}_{-r_i,-l_i}$, threshold $\tau_i$, 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 $\hat{l} = \max\{l_i : i \in \mathcal{I}_l, l_i \leq \min_{j\in\mathcal{I}_m}(m_j)\}$
        Set $\hat{r} = \min\{r_i : i \in \mathcal{I}_r, r_i \leq \max_{j\in\mathcal{I}_m}(m_j)\}$
        Set $\hat{S} = \{\hat{l}, \ldots, \hat{r}\}$
    end
end
return $\widehat{Q}$ :    $\widehat{Q}_k = 2\mathbb{1}_{\{k\in\hat{S}\}} - 1$
```

Algorithm 16: CTB

2.5 Discussion

2.5.1 Supplementary discussion concerning the *TBP* and *MTBP*

2.5.1.1 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 2.1 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*.

2.5.1.2 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 [68]. A corollary to the lower bound in [68] 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 [12] 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 [68]. 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*.

2.5.1.3 Supplementary details of the related works: *MTBP*

The papers [33], Ben-Or and Hassidim [8] and Emamjomeh-Zadeh, Kempe, and Singhal [30] 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 [73] 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 [53].

Karp and Kleinberg [53] 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 [8] 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 $(1 - \delta)\log(K)/\varepsilon^2$.

2.5.1.4 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 [53] is specific to the assumption that all arms follow a Bernoulli distribution - and related literature makes even more constraining assumptions [33, 8, 30]. 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 [53] - as well as in [33, 8, 30] 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 2.2.

- *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 [53, 8, 30, 23], 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 [53] 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 2.2 for a comprehensive summary of state of the art rates, as well as of our rates.

2.5.1.5 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 2.2 with the exception that we consider a modified version the simple regret

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

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 $\exp\big(-c\frac{1}{K}\sum \Delta_i^2 \frac{T}{K} + c' \log(\log(T)K)\big)$ is provided from [68], 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 [68], up to universal constants in the exponential. In the same setting we can upper bound the simple regret of the `Uniform` algorithm by $\sum_k \exp\big(-c\Delta_k^2 \frac{T}{K}\big)$, 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 $\exp\big(-cT \min_{k \in [K]} \Delta_k^2\big)$ - where $c > 0$ is some universal constant - while the `MTB` algorithm achieves an upper bound of the order $\exp\Big(-c\frac{T}{\log(K)} \min_{k \in [K]} \Delta_k^2\Big)$ - 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.

2.5.2 Supplementary discussion

2.5.2.1 Parameters of the algorithms

The `Uniform` algorithm only takes T as a parameter, see Subsection 2.5.2.2 for a discussion on how to make it anytime. The `MTB` algorithm takes only σ, K, T as parameters. Again, see Subsection 2.5.2.2 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. Note however that it is a very mild assumption. Indeed σ comes from Definition 1.In many case, natural choices for σ

	State of the art		Our results	
	LB	UB	LB	UB
TBP FB [1] [68]	$\sqrt{\frac{K}{T}}$	$\sqrt{\frac{K\log(K\log T)}{T}}$	$\sqrt{\frac{K\log(K)}{T}}$ [4]	$\sqrt{\frac{K\log(K)}{T}}$ [5]
TBP FC [23]	$\frac{K\log(\delta^{-1})}{\varepsilon^2}$	$\frac{K\log(K^2\varepsilon^{-2}\delta^{-1})}{\varepsilon^2}$	$\frac{K\log(K)(1-K^{-1}-\delta)}{\varepsilon^2}$ [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] [53]	$\frac{\underline{c}_\delta\log(K)}{\varepsilon^2}$ [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 2.2: 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.

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.

2.5.2.2 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, 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 6 and Corollary 5 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.

[4] See also Bubeck, Munos, and Stoltz [12] 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 [53].

[6] See also Ben-Or and Hassidim [8] 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, Gutzeit, and Carpentier [68] 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, Munos, and Stoltz [12].

[9] And combining this with the lower bound in [23], 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 [8] 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.

2.5.2.3 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.

2.6 Extension of results

2.6.1 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 $\{1, \ldots, K\}$ 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 $\beta \le 1$ and a constant $L > 0$ such that $\forall x, y \in [0, 1]$, $|\mu_x - \mu_y| \le L|x - y|^\beta$. In this case, straightforward corollaries of our results imply the minimax regret rates in Table 2.3.

In order 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 literature (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.

2.6.2 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 $\underline{\nu} = \{\nu_1, \ldots, \nu_K\}$, where ν_k is the unknown

Our results	Unstructured	Monotone	Unimodal	Convex
	TBP	*MTBP*	*UTBP*	*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 2.3: 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).

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 1 (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 accommodate this case in the MTB algorithm is to define

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

2.6.3 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 $1-\delta$, $R_T^{\nu,\pi} \leq \varepsilon$. 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.

2.6.4 Lower Bounds

The following corollary is a direct extension to Proposition 2 which provides a lower bound in the unstructured case.

Corollary 1. *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 2. 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 2 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 [23], 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 5 which provides a lower bound in the monotone case.

Corollary 2. *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 [53] , but for Bernoulli random variables in the lower bound, and without providing an explicit dependence on δ. In the paper [8], 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.*

2.6.5 Upper Bounds

The following Corollary is a direct extension to Proposition 4, which provides an upper bound on regret of the `Uniform` algorithm.

Corollary 3. *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 1 combined with the one in [23].

The following Corollary is a direct extension to Corollary5 which provides an upper bound on the regret of the MTB algorithm,

Corollary 4. *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 2. 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*.

2.7 Proofs

2.7.1 Proof of Theorem 23

In the proof of all results in this section, we assume that the more general sub-Gaussian assumption described in Section 2.6.2 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 $\mathcal{B}er(p)$ and $\mathcal{B}er(q)$ (with the usual conventions) by

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

```
for $k = 1 : K$ do
    Sample arm $k$ a total of $\lfloor \frac{T}{K} \rfloor$ times.
    Compute $\widehat{\mu}_k$ the sample mean of arm $k$.
end
return
    $\widehat{Q} : \quad \widehat{Q}_k = \begin{cases} -1 & \text{if } \widehat{\mu}_k < \tau \\ 1 & \text{if } \widehat{\mu}_k \geq \tau \end{cases}$
```

Algorithm 17: Uniform

During this section we will prove Theorem 23 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 2. For any $T \geq 1$ and any strategy π, there exists a unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, such that

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

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 $Q \in \{-1,1\}^K$

$$\underline{\nu}^Q = \big(\mathcal{N}(Q_1\varepsilon,\sigma^2),\dots,\mathcal{N}(Q_K\varepsilon,\sigma^2)\big)\,,$$

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

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

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 $\underline{\nu}^Q$,

$$\max_Q \mathbf{R}_T^{\underline{\nu}^Q,\pi} \geq \varepsilon\left(1-\frac{1}{2^K}\sum_Q \mathbb{E}_Q \mathbb{1}_{\{\widehat{Q}=Q\}}\right). \tag{2.1}$$

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, Ménard, and Stoltz [41], we have

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

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, Ménard, and Stoltz [41]) $\mathrm{kl}(x,y) \geq (x-y)^2 \max\big(2,\log(1/y)\big)$, we obtain

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}\mathbb{1}_{\{\widehat{Q}=Q^k\}} \leq \frac{1}{K} + \sqrt{\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{Q^k}[N_k(T)]\frac{\varepsilon^2}{2\sigma^2\max\big(2,\log(K)\big)}}\,. \tag{2.2}$$

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

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

Now it remains to remark that by symmetry

$$\sum_{Q}\sum_{k=1}^{K}\mathbb{E}_{Q^k}\mathbb{1}_{\{\widehat{Q}=Q^k\}}=\sum_{Q'}\sum_{k=1}^{K}\mathbb{E}_{Q'}\mathbb{1}_{\{\widehat{Q}=Q'\}}=K\sum_{Q}\mathbb{E}_{Q}\mathbb{1}_{\{\widehat{Q}=Q\}}\,,$$

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

Hence from (2.2) we get

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

and then from (2.1) we obtain

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

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 3. For any unstructured bandit problem $\underline{\nu}\in\mathcal{B}$, any $T\ge K$, any $0<\delta<1$, Uniform satisfies

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

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,

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

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

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

and the returning classification is

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

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

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

□

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

Proposition 4. For any unstructured bandit problem $\underline{\nu} \in \mathcal{B}$, and any $T \geq K$, Uniform satisfies

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

Proof. By application of Theorem 3, for $\varepsilon > 0$ we have,

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

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

$$\begin{aligned}
\mathbf{R}_T &\leq \sqrt{2}\varepsilon_0 + \int_{\sqrt{2}\varepsilon_0}^{+\infty} \exp\left(-(\varepsilon^2 - \varepsilon_0^2)\frac{T}{2\sigma^2 K}\right) \mathrm{d}\varepsilon \\
&\leq \sqrt{2}\varepsilon_0 + \int_0^{+\infty} \exp\left(-\varepsilon^2 \frac{T}{8\sigma^2 K}\right) \mathrm{d}\varepsilon \\
&= \sqrt{\frac{8\sigma^2 \log(2K)K}{T}} + \sqrt{\frac{2\pi\sigma^2 K}{T}} \\
&\leq 7\sqrt{\frac{\sigma^2 \log(2K)K}{T}} .
\end{aligned}$$

□

Setting $\sigma = 1$, Theorem 23 follows directly from a combination of Propositions 4 and 2.

2.7.2 Proof of Theorem 24

In the proofs of all results in this section, we assume that the more general sub-Gaussian assumption described in Section 2.6.2 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 2.6.2. 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 24 by first demonstrating a lower bound upon expected regret in the *MTBP* setting, Proposition 5. We will then go on to provide an upper bound on the regret of the MTB with high probability, Proposition 6 which will be used to finally prove Corollary 5 which provides a optimal bound for the MTB in expected regret. Setting $\sigma = 1$ Theorem 24 will then follow directly from Proposition 5 and Corollary 5.

Proposition 5. For any $T \geq 1$ and any strategy π, there exists a structured bandit problem $\underline{\nu} \in \mathcal{B}_m$, such that

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

Proof. We will proceed as in the proof of Proposition 2. 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 $k \in \{0,\dots,K\}$, $l \in [K]$,

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

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

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

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 2 one can lower bound the maximum of the expected regrets over all the bandit problems introduced above,

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

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

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

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

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

Hence combining the last three inequalities we get

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

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 5 to establish an upper bound on the expected regret of the MTB algorithm. For Proposition 6 we consider a more general set of problems, given $\varepsilon > 0$, define,

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

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 6. 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,

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

Corollary 5. *For any problem $\underline{\nu} \in \mathcal{B}_m$ and any $T \geq 12\log(K)$, the MTB algorithm will achieve the following bound on expected regret,*

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

The proof of Proposition 6 and Corollary 5 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 7 we prove these nodes form a "consecutive tree", see Definition 3. 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 (2.10). In Lemma 4 we prove that on every good decision we move towards the set of good arms or remain within them. Lemma 5 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 6, to give an upper bound on regret. This in combination with a second upper bound, Lemma 8, will give our result.

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

Definition 2. 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 3. A consecutive tree U with root $u_{\texttt{root}}$ is a set of nodes such that $u_{\texttt{root}} \in U$ and

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

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

$$Z_1^\varepsilon := \{v : \exists k \in \{l, m, r\} : |\mu_{v(k)} - \tau| \leq \varepsilon\}\,, \tag{2.3}$$

$$Z_2^\varepsilon := \{v : v(r) = v(l) + 1;\ \mu_{v(l)} \leq \tau \leq \mu_{v(r)}\}\backslash Z_1^\varepsilon\,, \tag{2.4}$$

that is

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

It is important to note that

$$Z_2^\varepsilon \neq \emptyset \Rightarrow |Z^\varepsilon| = 1 \; . \tag{2.5}$$

Proposition 7. Z^ε is a consecutive tree with root $z_{\text{root}}^\varepsilon$ the unique element $v \in Z^\varepsilon$, such that $P(v) \notin Z^\varepsilon$.

Proof. If $Z_2^\varepsilon \neq \emptyset$ by (2.5) 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

$$\exists k \in \{l, m, r\} : \left|\mu_{v(k)} - \tau\right| \leq \varepsilon \; . \tag{2.6}$$

Now since $v(l), v(r) \in P(v)$ and $P(v) \notin Z^\varepsilon$, it follows that, thanks to (2.6),

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

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 2. *The node w_t^ε is unique and*

$$w_t^\varepsilon = \underset{\{w : ST(w) \cap Z^\varepsilon \neq \emptyset \;\&\; v_t \in ST(w)\}}{\arg\min} \left(|v_t| - |w| + (|z_{root}^\varepsilon| - |w|)^+\right) \, . \tag{2.7}$$

Proof. At time t consider, a node q_t^ε which also satisfies 2.7, giving

$$|v_t| - |w_t^\varepsilon| + (|z_{\mathtt{root}}^\varepsilon| - |w_t^\varepsilon|)^+ = |v_t| - |q_t^\varepsilon| + (|z_{\mathtt{root}}^\varepsilon| - |q_t^\varepsilon|)^+ \; .$$

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,

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

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

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

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

$$ST(v_t) \cap Z^\varepsilon \neq \emptyset \Rightarrow v_t = w_t^\varepsilon \ , \tag{2.8}$$

$$D_t = 0 \Rightarrow v_t = w_t^\varepsilon \text{ And } w_t^\varepsilon, v_t \in Z^\varepsilon \ . \tag{2.9}$$

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 $S_t^{2\varepsilon}$,

$$G_t^\varepsilon := \left| \left\{ k \in S_t^{2\varepsilon} : \ k \in W_{3\varepsilon} \right\} \right| . \tag{2.10}$$

Note that if $\hat{a}$ 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,

$$\xi_t^\varepsilon := \left\{ \forall k \in \{l, m, r\}, \left| \hat{\mu}_{k,t} - \mu_{v_t(k)} \right| \leq \varepsilon \right\} . \tag{2.11}$$

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.

$$v_{t+1} \in \{L(v_t), R(v_t), P(v_t)\} , \tag{2.12}$$

- for $t \leq T_1$,

$$|S_t^{2\varepsilon}| \leq |S_{t+1}^{2\varepsilon}| \leq |S_t^{2\varepsilon}| + 1 . \tag{2.13}$$

Lemma 3. *On execution of algorithm Explore and algorithm Choose with parameter $\varepsilon > 0$ for all $t \leq T_1$ we have the following,*

$$D_{t+1}^\varepsilon \leq D_t^\varepsilon + 1, \tag{2.14}$$

$$G_{t+1}^\varepsilon \geq G_t^\varepsilon . \tag{2.15}$$

Proof. As the algorithm moves at most 1 step per iteration, see (2.12), for $t \leq T_1$, it holds

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

Noting that,

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

where the third line comes from the definition of w^{ε}_{t+1}, see (2.7), we obtain $D^{\varepsilon}_{t+1} \leq D^{\varepsilon}_t + 1$. By (2.13) we have, for $t \leq T_1$,

$$|S^{2\varepsilon}_t| \leq |S^{2\varepsilon}_{t+1}| \leq |S^{2\varepsilon}_t| + 1\,,$$

hence $G^{\varepsilon}_{t+1} \geq G^{\varepsilon}_t$. □

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

Lemma 4. *On execution of algorithm* `Explore` *and algorithm* `Choose` *with parameter $\varepsilon > 0$ for all $t \leq T_1$, on ξ^{ε}_t, we have the following,*

$$D^{\varepsilon}_{t+1} \leq \max(D^{\varepsilon}_t - 1, 0)\,, \tag{2.16}$$
$$G^{\varepsilon}_{t+1} \geq G^{\varepsilon}_t + \mathbb{1}_{\{D^{\varepsilon}_t = 0\}}\,. \tag{2.17}$$

Proof. We first prove (2.17). Note that if the arm $v_t(k)$ is added in $S^{2\varepsilon}_{t+1}$ then either $|\widehat{\mu}_{k,t} - \tau| \leq 2\varepsilon$ or $v_t(k) = v_t(r) = v_t(l) + 1$ and $\widehat{\mu}_{l,t} + \varepsilon \leq \tau \leq \widehat{\mu}_{r,t}$. Thus, on ξ^{ε}_t, 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 $|\widehat{\mu}_{k,t} - \tau| \leq 2\varepsilon$. 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

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

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

Now we prove (2.16). Note that on the favorable event ξ^{ε}_t, we have $\forall k \in \{l, m, r\}$,

$$\mu_{v_t(k)} \geq \tau + \varepsilon \Rightarrow \hat{\mu}_{k,t} \geq \tau\,, \tag{2.18}$$
$$\mu_{v_t(k)} \leq \tau - \varepsilon \Rightarrow \hat{\mu}_{k,t} \leq \tau\,. \tag{2.19}$$

We consider the following three cases:

- If $\tau \notin \left[\mu_{v_t(l)} + \varepsilon,\, \mu_{v_t(r)} - \varepsilon\right]$. From (2.18) and (2.19), 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 2, under ξ^{ε}_t,

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

- 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 (2.18) and (2.19), 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 2, under ξ_t^ε,

$$D_{t+1}^\varepsilon = (|w_{t+1}| - |z_{\texttt{root}}^\varepsilon|)^+ = (|w_t| - |z_{\texttt{root}}^\varepsilon|)^+ - 1 = D_t^\varepsilon - 1\,.$$

- If $\tau \in \big[\mu_{v_t(l)} + \varepsilon,\, \mu_{v_t(r)} - \varepsilon\big]$ and $v_t \in Z^\varepsilon$. We distinguish two cases: Z_2^ε is empty or not. In both cases we will show that, under ξ_t^ε, $v_{t+1} \in Z^\varepsilon$ and thus

  $$D_{t+1}^\varepsilon = D_t^\varepsilon = 0\,.$$

  Hence it remains to consider these two cases:

  - If $Z_2^\varepsilon \neq \emptyset$. Via the definition of Z_2^ε, see (2.4), and the fact $Z_1^\varepsilon = \emptyset$, v_t is a leaf with $\mu_{v_t(r)} \leq \tau - \varepsilon$ and $\mu_{v_t(l)} \geq \tau + \varepsilon$. Hence from (2.18) and (2.19) 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 $G_{T_1+1}^\varepsilon$ We denote by $\bar\xi_t^\varepsilon$ the complement of ξ_t^ε.

Lemma 5. *For any execution of algorithm `Explore` and subsequent execution of `Choose` with parameter $\varepsilon \geq \varepsilon_0$,*

$$G_{T_1+1}^\varepsilon \geq \frac{3}{4} T_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\bar\xi_t^\varepsilon}\,.$$

Proof. Combining (2.16) and (2.14) from Lemma 3 and Lemma 4 respectively we have

$$\begin{aligned} D_{t+1}^\varepsilon &\leq D_t^\varepsilon + \mathbb{1}_{\bar\xi_t^\varepsilon} - \mathbb{1}_{\xi_t^\varepsilon}\mathbb{1}_{\{D_t^\varepsilon > 0\}} \\ &= D_t^\varepsilon + 2\mathbb{1}_{\bar\xi_t^\varepsilon} - 1 + \mathbb{1}_{\xi_t^\varepsilon}\mathbb{1}_{\{D_t^\varepsilon = 0\}}\,. \end{aligned}$$

Using this inequality with (2.17) we obtain

$$\begin{aligned} G^{\varepsilon}_{T_1+1} &= \sum_{t=1}^{T_1} G^{\varepsilon}_{t+1} - G^{\varepsilon}_t \\ &\geq \sum_{t=1}^{T_1} \mathbb{1}_{\xi^{\varepsilon}_t} \mathbb{1}_{\{D^{\varepsilon}_t = 0\}} \\ &\geq \sum_{t=1}^{T_1} \left(D^{\varepsilon}_{t+1} - D^{\varepsilon}_t - 2\mathbb{1}_{\xi^{\varepsilon}_t} + 1\right) \\ &\geq T_1 - D^{\varepsilon}_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\xi_{t,\varepsilon}} \\ &\geq \frac{3}{4} T_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\xi_{t,\varepsilon}} , \end{aligned}$$

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

$$\mathbb{P}(R_T > 3\varepsilon) \leq e^{-3\log(K)/4} . \tag{2.20}$$

Before proving Lemma 6 we need to show that the number of times a favorable events $\xi^{\varepsilon_0}_t$ occurs is not to small with high probability. Precisely in the following lemma we upper bound the probability of the event

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

Lemma 7. *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 $\mathbb{P}(\bar{\xi}^{\varepsilon_0}_t | \mathcal{F}_{t-1}) \leq 1/8$. Then the Azuma-Hoeffding inequality applied to the martingale

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

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

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \left[\mathbb{1}_{\bar{\xi}^{\varepsilon_0}_t} - \mathbb{P}(\bar{\xi}^{\varepsilon_0}_t | \mathcal{F}_{t-1})\right] \geq \frac{T_1}{4}\right) \leq e^{-2T_1/16} \leq e^{-3\log(K)/4} , \tag{2.21}$$

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

We are now ready to prove Lemma 6.

Proof of Lemma 6. We first prove it for $\varepsilon = \varepsilon_0$. Thanks to Lemma 5 on the event ξ^{ε_0} we have

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

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

$$S^{2\varepsilon_0}_{T_1+1} \subset S^{\hat{\varepsilon}}_{T_1+1}\,.$$

Thus there is more than the half of the arms of $S^{\hat{\varepsilon}}_{T_1+1}$ in $W_{3\varepsilon_0}$, since this list is at most of size T_1. In particular this implies that $\hat{a} = \mathrm{Median}(S^{\hat{\varepsilon}}_{T_1+1}) \in W_{3\varepsilon_0}$. Indeed $W_{3\varepsilon_0}$ is a segment in $[K]$, see (2.6). Therefore, on the event ξ^{ε_0} we have

$$R_T \leq 3\varepsilon_0.$$

Lemma 7 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 8. *For all $\varepsilon \geq \varepsilon_0$, following the execution of algorithm* MTB*,*

$$\mathbb{P}(R_T > 3\varepsilon) \leq 72\log(K)\exp\!\left(-\frac{T\varepsilon^2}{36\sigma^2\log(K)}\right). \tag{2.22}$$

Proof. We consider the event where all the favorable events ξ^{ε}_t occur,

$$\xi^{\varepsilon}_a := \bigcap_{t=1}^{T_1} \xi^{\varepsilon}_t\,.$$

On this event ξ^{ε}_a thanks to Lemma 5 we have

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

hence $S^{2\varepsilon}_{T_1+1} \neq \emptyset$ is not empty. Furthermore following the same arguments of the beginning of the proof of Lemma 4 all arms in the list $S^{2\varepsilon}_{T_1+1} \neq \emptyset$ are also in $W_{3\varepsilon}$. Then noting that by construction

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

we get $\hat{\varepsilon} \leq 2\varepsilon$ therefore $S^{\hat{\varepsilon}}_{T_1+1} \subset S^{2\varepsilon}_{T_1+1}$. Thanks to the remarks above we know that $\hat{a} \in W_{3\varepsilon}$ thus on ξ^{ε}_a,

$$R_T \leq 3\varepsilon\,.$$

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

$$\mathbb{P}(\bar{\xi}_a^{\varepsilon}) \le \sum_{t=1}^{T_1} \mathbb{E}\big[\mathbb{P}(\bar{\xi}_t^{\varepsilon}|\mathcal{F}_{t-1})\big] \le 72\log(K)\exp\Big(-\frac{T_2\varepsilon^2}{2\sigma^2}\Big) \tag{2.23}$$

$$\le 72\log(K)\exp\Big(-\frac{T\varepsilon^2}{36\sigma^2\log(K)}\Big)\,. \tag{2.24}$$

□

Conclusion The proof of Proposition 6 is straightforward combining Lemma 6 and Lemma 8. Thus we obtain for all $\varepsilon \ge 3\varepsilon_0$,

$$\mathbb{P}(R_T \ge \varepsilon) \le \min\bigg(\exp\Big(-\frac{3\log(K)}{4}\Big), 72\log(K)\exp\Big(-\frac{T\varepsilon^2}{324\sigma^2\log(K)}\Big)\bigg)\,.$$

We can integrate the high probability upper bound obtained in Proposition 6 to prove Corollary 5.

Proof of Corollary 5. Thanks to Proposition 6, for $\varepsilon_1 = \log(72\log(K))\sqrt{324\sigma^2\log(K)/T}$, we have

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

□

Setting $\sigma = 1$ Theorem 24 follows directly from Proposition 5 and Corollary 5.

2.7.3 Proof of Theorem 25

To prove Theorem 25 we first demonstrate, in Proposition 8, a lower bound on the expected regret of any strategy on the *UTBP*. We will then show, with Proposition 9, that the `UTB` achieves said lower bound. The proof of Theorem 25 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 8. For any $T \ge 1$ and any strategy π, there exists an unimodal bandit problem $\underline{\nu} \in \mathcal{B}_u$, such that

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

Proof. We will proceed as in the proof of Proposition 2. 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 2 one can lower bound the maximum of the expected regrets over all the bandit problems introduced above,

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

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

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

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

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

Hence combining the last three inequalities we get

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

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

Proposition 9. There exists a universal constant $c_{\text{uni}} > 0$ such that for any unimodal bandit problem $\underline{\nu} \in \mathcal{B}_u$, UTB satisfies

$$\mathbf{R}_T^{\mathtt{CTB}, \underline{\nu}} \leq c_{\text{uni}} \sqrt{\frac{K}{n}} .$$

Proof. **Step 1: Definitions** Write

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

and

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

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

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

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

$$R_T \leq R^{(l)} \mathbb{1}_{\{\mathcal{E}\}} + (\mu_{\hat{m}} - \tau)_+ \mathbf{1}\{\mathcal{E}^C\} + \hat{\Delta}. \tag{2.25}$$

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

$$R^{(l)} \mathbb{1}_{\{\mathcal{E}\}} \leq R^{(l)} \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon}. \tag{2.26}$$

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 $R^{(l)} \geq R^{(r)}$

$$(\mu_{\hat{m}} - \tau)_+ \mathbb{1}_{\{\mathcal{E}^C\}} \leq R^{(l)} \wedge (\mu_{\hat{m}} - \tau)_+ + 2\hat{\varepsilon}. \tag{2.27}$$

So we have combining (2.26) and (2.27) all cases in (2.25) 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 6 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.

□

2.7.4 Proof of Theorem 26

For the proof of Proposition 11 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 10. For any $T \geq 1$, $K \geq e^{12}$ and any strategy π, there exists a structured bandit problem $\underline{\nu} \in \mathcal{B}_c$, such that

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

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

$$\mu_k^l = \begin{cases} \frac{k}{k_l}\varepsilon & \text{if } k \leq 2k_l := 2^{l+1} \\ 2\varepsilon & \text{else}\,. \end{cases}.$$

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

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,

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

where we denote by $\mathbb{E}^l$ the expectation and by Q^l the true classification in the problem $\underline{\nu}^l$. In particular we have $Q^l = [-1, \dots, -1, 1, \dots, 1]$ where the first one is at position k_l. Let $\hat{l} = \arg\min\{j \in [L] :\ \forall i \geq j,\ \widehat{Q}_{k_i} = 1\}$ be an estimate for the index of the problem with the convention $\hat{l} = L$ if the set is empty. Then we have

$$\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widetilde{Q}_k \neq Q_k^l\}} \geq \frac{\varepsilon}{2} \mathbb{1}_{\left\{\hat{l} \notin \{l, l+1\}\right\}}.$$

Indeed if $\hat{l} < l$ then we know that $\widehat{Q}_{k_{\hat{l}}} = 1 \neq -1 = Q_{k_{\hat{l}}}$, thus we obtain

$$\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widehat{Q}_k \neq Q_k^l\}} \geq \Delta_{k_{\hat{l}}}^l = \varepsilon - \frac{k_{\hat{l}}}{k_l}\varepsilon \geq \frac{\varepsilon}{2}.$$

Else $\hat{l} > l+1$, and similarly we get, because $\widehat{Q}_{k_{\hat{l}-1}} = -1$ and $\hat{l} - 1 > l$ (or $\widehat{Q}_k = -1$ for some $k > \hat{l}$ if we choose $\hat{l} = L$ in the case where the set defining $\hat{l}$ is empty),

$$\max_{k\in[K]} \Delta_k^l \mathbb{1}_{\{\widehat{Q}_k \neq Q_k^l\}} \geq \Delta_{k_{\hat{l}-1}}^l = \min\left(\frac{k_{\hat{l}-1}}{k_l}\varepsilon - \varepsilon, \varepsilon\right) \geq \frac{\varepsilon}{2}.$$

Using the previous inequality we obtain

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

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

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

Hence combining the last three inequalities we get

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

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

Proposition 11. There exists a universal constant $c_{\text{conv}} > 0$ such that for any convex bandit problem $\underline{\nu} \in \mathcal{B}_c$, CTB satisfies

$$\mathbf{R}_T^{\texttt{CTB},\underline{\nu}} \le c_{\text{conv}} \sqrt{\frac{\log \log K \vee 1}{T}} \,.$$

For a sketch of proof of Proposition 11, see Section 2.7.5. Before going on to prove Proposition 11 we first show the following.

Lemma 9. *Consider* $1 \le p \le q \le K$, $\tilde{\varepsilon} > 0, \tilde{\tau} \in \mathbb{R}$. *Consider any* $1 \le p < q \le K$, *such that,*

$$\mu_{\lfloor \frac{p+q}{2} \rfloor} \ge \tilde{\tau} + \frac{1}{8}\tilde{\varepsilon} \,. \tag{2.28}$$

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

- If $k^* \in [p, \lfloor \frac{p+q}{2} \rfloor]$, and since (2.28) holds, we have that $\forall k \in [k^*, \lfloor \frac{p+q}{2} \rfloor]$, $\mu_k - \tilde{\tau} \ge \tilde{\varepsilon}/8$. 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 (2.28) that $\forall k \in [p, \lfloor \frac{p+q}{2} \rfloor]$, $\mu_k - \tilde{\tau} \ge \tilde{\varepsilon}/8$. This implies the result.

□

Lemma 10. *Let* $\tilde{\varepsilon} > 0, \tilde{\tau} \in \mathbb{R}$. *For any* $1 \le p \le q \le K$, *such that,*

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

we have that, $\forall k \in \{p, \dots, q\}$ *that* $\mu_k \le \tilde{\tau} - \frac{1}{8}\varepsilon$.

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 10 we have $(\mu_k - \mu_{\lfloor \frac{p+q}{2} \rfloor}) > \frac{1}{2}\tilde{\varepsilon}$. However, via the convex property $(\mu_k - \mu_{\lfloor \frac{p+q}{2} \rfloor}) \le (\mu_{\lfloor \frac{p+q}{2} \rfloor} - \mu_q)$, a contradiction as it implies with the forelast equation that $\mu_q < \tilde{\tau} - \frac{1}{8}\tilde{\varepsilon}$. □

We now define the event,

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

Consider the event

$$\mathcal{E}_i = \{\mu_{m_i} \geq \tau_i + \frac{1}{8}\varepsilon_i\}. \tag{2.29}$$

Proposition 12. 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 $\mathcal{S}_{l_i,r_i}^{\log}$. 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 $-\mathcal{S}_{-r_i,l_i}^{\log}$.

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

Proof. A straightforward corollary of Proposition 6 is as follows.

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

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

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, $|\mathcal{S}_{l_i,r_i}^{\log}| \leq \log K$ so that we apply MTB on a problem that has less than $\log K$ arms,
- that on $\mathcal{E}_i$, we have that $(\min(|\mu_k - \tau_i|, \varepsilon_i/8)\mathrm{sign}(\mu_k - \tau_i))_{k\in[l_i,m_i]}$ is increasing (respectively, $(\min(|\mu_k - \tau_i|, \varepsilon_i/8)\mathrm{sign}(\mu_k - \tau_i))_{k\in[m_i,r_i]}$ is decreasing) - see Lemma 9.

- Moreover $\varepsilon_i \geq \varepsilon_M \geq \sqrt{\frac{2\log(48)6\log\log(K)}{T}}$. And so since $\mathcal{S}^{\log}_{l_i,r_i} \subset [l_i, m_i]$ (resp. $-\mathcal{S}^{\log}_{-r_i,-l_i} \subset [m_i, r_i]$) and $|S^{\log}_{l_i,r_i}| \leq \log(K)$, the conditions of Corollary 6 are satisfied, for the set $\mathcal{S}^{\log}_{l_i,r_i}$ of arms.

Therefore we can apply Corollary 6 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 $\hat{a}$ such that setting $l_{i+1} = \hat{a}$ satisfies our result with probability greater than $1 - \delta'_i$.

□

Proposition 13. We have that for $i \leq M$

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

and

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

Proof. We prove this proposition only for $\xi^{(L)}_{i+1}$ as the proof for $\xi^{(R)}_{i+1}$ is similar. Consider the high probability event of Proposition 12, 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,

$$\mu_{l'_{i+1}} + \varepsilon_i/8 < \tau_i < \mu_{l_{i+1}} - \varepsilon_i/8 \ , \tag{2.30}$$

where l'_{i+1} is defined in Proposition 12. 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 $\xi^{(L)}_{i+1}$

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 9 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 14. We have that for $i \leq M$

$$\mathbb{P}\Big(\xi_{i+1}^{(A)}\Big|\xi_i \cap \mathcal{E}_i^c\Big) = 1.$$

Proof. On $\xi_i \cap \mathcal{E}_i^c$, 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 10 we conclude that for any $k \leq K$, $\mu_k < \tau - \frac{1}{8}\varepsilon_i$. And so $\xi_{i+1}^{(A)}$ holds.

□

Corollary 7. *We have that*

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

Proof. This holds by combining Propositions 13 and Proposition 14. □

Hence by Corollary 7 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

$$\eta_i^I := \Bigg\{ |\widehat{\mu}_{m,i} - \mu_{m_i}| \vee |\widehat{\mu}_{l,i} - \mu_{l_i}| \vee |\widehat{\mu}_{r,i} - \mu_{r_i}| \vee \\ |\widehat{\mu}_{l-1,i} - \mu_{l_i-1}| \vee |\widehat{\mu}_{r+1,i} - \mu_{r_i+1}| \leq \frac{1}{16}\varepsilon_i \vee \varepsilon_I \Bigg\}, \tag{2.31}$$

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

$$1 - 10\exp\left(-\frac{1}{2}T_2^{(i)}\varepsilon_i^2\right) \geq 1 - 10\delta_i. \tag{2.32}$$

Proposition 15. 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 $\{k : \mu_k \geq \tau\} \subset \{l_I, \ldots, r_I\} \subset \{k : \mu_k \geq \tau - \varepsilon_I\}$.

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 7 and the definition of ξ_I. Now assume $\{k : \mu_k = \tau\} \neq \emptyset$. Let k^* be as in the proof of Lemma 9. 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 16. 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 17. Fix $I \leq M$. On $\xi_I \cap \left(\cap_{i\leq M} \eta_i^I \right)$, we have that

$$\{m_i : i \in \mathcal{I}_m\} \subset \{l_I, \dots, r_I\},$$

and also

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

Proof. On $\cap_{i\leq I}\eta_i^I$, we have that $\{m_i : i \in \mathcal{I}_m\} \subset \{k : \mu_k \geq \tau\} \cup \{l_I, \dots, r_I\}$, and so from Propositions and 15 and 16, we have on $\cap_{i\leq I}\eta_i^I \cap \xi_I$, that $\{m_i : i \in \mathcal{I}_m\} \subset \{l_I, \dots, r_I\}$.

The proof that $I \in \mathcal{I}_l$ on $\xi_I \cap \eta_I^I$ - as well as the fact that $I \in \mathcal{I}_r$ - follows immediately by combining the definition of $\mathcal{I}_l$ - resp. $\mathcal{I}_r$ - with Proposition 15 and 16, and the definition of η_I^I. □

Proposition 18. Fix $I \leq M$, and assume that $I \notin \mathcal{I}_m$. 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 17 that $\{m_i : i \in \mathcal{I}_m\} \subset \{l_I, \dots, r_I\}$ and that $I \in \mathcal{I}_l, I \in \mathcal{I}_r$. 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 15 and 16 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 18 that $\mu_{m_I} \leq \tau + \frac{17}{8}\varepsilon_I$. Together with Proposition 15 and 16 and Lemma 10, 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 19. Fix $I \leq M$, and assume that $I \in \mathcal{I}_m$. 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 18, 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 $\hat{l} \in \{l_i : i \in \mathcal{I}_l\}, \hat{r} \in \{r_i : i \in \mathcal{I}_r\}$ 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 19 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 of Proposition 11. Let $I \leq M$. Combining Propositions 18 and 19, 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 $i \in \{0, \ldots, M\}$ we have

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

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. □

2.7.5 Sketch proof of Proposition 11

Broadly speaking the CTB runs the MTB iteratively on $\log K$ size subsets of arms to gradually refine our arm set. To better understand the mechanism at work, let us first consider a more restricted setting, where the sequence of means is both monotone and concave.

An illustrative example in the case of a monotone and concave constraint Consider a bandit problem whose sequence of means is both monotone and concave, i.e. a problem $\nu \in \mathcal{B}_m \cap \mathcal{B}_c$. Say we wish to find an arm $\tilde{k}$ such that

$$\tilde{k} \geq \tau - \frac{1}{2} \qquad \text{and} \qquad \forall k < \tilde{k}, \mu_k \leq \tau\ . \tag{2.33}$$

To do this we can run the MTB on the set $S^{\log}_{1,K}$, with threshold $\tilde{\tau} = \tau - \frac{1}{2}$. Consider the arm $a = \min\{k \in S^{\log}_{1,K} : \mu_a \geq \tau - \frac{1}{2}\}$, that is, a is the arm we hope to output running the MTB on $S^{\log}_{1,K}$ with threshold $\tilde{\tau} = \tau - \frac{1}{2}$. If $a \in \{1, 2\}$ then it trivially satisfies the properties of Equation (2.33). Assume $a \geq 4$, if

$$\mu_a \geq \tau - \frac{1}{2}, \qquad \mu_{a/2} < \tau - \frac{1}{2}\ ,$$

then via the concavity of the sequence of means, we have the additional guarantee that $\mu_a \leq \tau$, as otherwise, via concavity $\mu_{a/4} \leq \tau - 1$, contradicting the assumption that $\forall k \in [K], \mu_k \in [0, 1]$. Thus, by running the MTB on a $\log(K)$ sized subset of our arms, we can find, with high probability, an arm $\tilde{k}$ satisfying Equation (2.33) and can effectively refine our arm set by removing all arms $k : k \neq \tilde{k}, \mu_k \leq \tau - \frac{1}{2}$. The CTB will make use of repeated application of this phenomenon, to incrementally refine our arm set, although, as we only assume the sequence of means is concave, not necessarily monotone, there are several additional technical difficulties.

The CTB for the concave setting The CTB runs for M phases, at each phase $i \in [M]$ it maintains an active set of arms $\{l_i, ..., r_i\}$, with $m_i = \lfloor \frac{l_i + r_i}{2} \rfloor$. We will refine our set as follows, run MTB on $S^{\log}_{l_i, r_i}$, respectively DEC-MTB on $-\mathcal{S}^{\log}_{-r_i,-l_i}$, with threshold τ_i, to find arm l_{i+1}, respectively r_{i+1}. Our probabilistic guarantees on l_{i+1} and r_{i+1} depend upon m_i. Firstly, assume

$$\mu_{l_i} \geq \tau - \varepsilon_i \ , \forall k < l_i : \mu_k \leq \tau - \frac{1}{2}\varepsilon_i \ ,$$

and,

$$\mu_{r_i} \geq \tau - \varepsilon_i \ , \forall k > r_i : \mu_k \leq \tau - \frac{1}{2}\varepsilon_i \ .$$

Case 1: $\mu_{m_i} \geq \tau_i + \frac{1}{8}\varepsilon_i$**.** In this case, where arm m_i is significantly greater than τ_i, the following will hold with high probability,

$$\mu_{l_{i+1}} \geq \tau - \varepsilon_{i+1} \ , \forall k < l_{i+1} : \mu_k \leq \tau - \frac{1}{2}\varepsilon_{i+1} \ ,$$

and

$$\mu_{r_{i+1}} \geq \tau - \varepsilon_{i+1} \ , \forall k > r_{i+1} : \mu_k \leq \tau - \frac{1}{2}\varepsilon_{i+1} \ .$$

This result is contained in Proposition 13. The reason the above holds is that we exploit the phenomenon observed in the first paragraph of this sketch of proof. Of course, things are a little more complicated without the additional monotonic assumption, as now we must insure that l_{i+1},r_{i+1} remain either side of the max of our concave sequence, that is, $\max \mu_k \in [l_i, r_i]$, however, this is merely a technical issue and is resolved using the concavity of the sequence of means and the assumption $\mu_{m_i} \geq \tau_i + \frac{1}{8}\varepsilon_i$ of **Case 1**, while utilising several technical Lemmas, see Lemmas 9 and 10.

Case 2: $\mu_{m_i} \leq \tau_i + \frac{1}{8}\varepsilon_i$**.** In this case, the arm m_i is not significantly greater than the threshold τ_i and we can leverage the concave property to show that,

$$\forall k \leq K, \mu_k \leq \tau - \frac{1}{8}\varepsilon_i \ ,$$

this result is contained in Proposition 14. Essentially, this case isn't an issue as we can simply classify all arms as below threshold, however, the CTB must recognise when **Case 2** occurs. Proposition 16 ensures that if, for some phase $i < M$, **Case 2** occurs, with high probability we will set $\hat{S} = \emptyset$ and classify all arms below threshold.

Choosing $\hat{S}$ The number of phases M is chosen carefully such that $\varepsilon_M \leq \sqrt{\frac{\log\log(K)}{T}}$. However, assuming $\hat{S} \neq \emptyset$, instead of simply outputting $\hat{S} = \{l_M, r_M\}$, we set $\hat{S} = \{\hat{l}, ..., \hat{r}\}$ for a carefully chosen $\hat{l}, \hat{r}$. This final step is to ensure that we have suitable bounds on the probability that the regret of the CTB exceeds some ε, for all $\varepsilon \geq \varepsilon_M$, as such a result is necessary to bound the expected simple regret.

Conclusion As we run the MTB iteratively on $\log(K)$ sized subsets, as opposed to the entire arm set, we can expect tighter error bounds. The trade off is that we run the MTB many times and must divide our budget accordingly, however, in the initial phases, we deal with arms very far from threshold and can be relatively loose with our probabilistic guarantees. For this reason, when allocating our budget across the M phases, we give more budget to later phases. This can be seen in our definition of $T_2^{(i)}$.

Chapter 3

Problem Dependent View on Structured TBP

In this chapter we present the following work, "Problem Dependent View on Structured Thresholding Bandit Problems" [25], authored by James Cheshire, Pierre Ménard and Alexandra Carpentier.

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 SFI Sachsen-Anhalt for the project RE-BCI ZS/2019/10/102024 by the Investitionsbank Sachsen-Anhalt. 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.

3.1 Introduction

Stochastic multi-armed bandit problems model 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. [23]. The learner is presented with $[K] = \{1, \dots, K\}$ arms, each following an unknown distribution ν_k with unknown mean μ_k. We focus on the *fixed budget* variant of this problem. 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 known threshold $\tau \in \mathbb{R}$. We will measure the learner's performance by the *probability of error* - i.e. the probability that the learner mis-classifies at least one arm - and consider therefore the *problem dependent regime.*

The focus of this paper is on *structured, shape constrained TBP*. More precisely, we study the influence of some classical *structures, in the form of a shape constraint* on the *sequence of means of the arms*, on the *TBP* problem. That is, we study how classical shape constraints influence the probability of error. A related study was performed by Cheshire, Ménard, and Carpentier [26] for the problem independent (overall worst-case) regime, and we aim at extending this study to the *problem dependent regime.* We will aim at finding the problem dependent quantities that have an impact on the optimal probability of error, and at providing matching upper and lower bounds.

We will discuss three structured *TBP*s in this paper; among those, we recall existing results of one, and provide results for two. Here is a short overview.

Vanilla, unstructured case *TBP* The vanilla, unstructured case is the simplest *TBP* where we only assume that the distributions of the arms are sub-Gaussian - also related to the TOP-M[1] setting. The *TBP* is already well studied in the literature - both in a fixed budget and in a fixed confidence context - and we only introduce it here to provide a benchmark for later structured problems. We recall here results in the problem dependent, fixed budget, setting, which is most relevant for this paper. Locatelli, Gutzeit, and Carpentier [68] prove that up to multiplicative constants, and additives $\log(TK)$ terms, in the exponential, the optimal probability of regret in this problem is $\exp(-\frac{T}{\sum_{i:\Delta_i>0}\Delta_i^{-2}})$, where $\Delta_i = |\tau - \mu_i|$. We present their results for completeness and comparison to the bounds under additional shape constraints in Table 3.1 - see also Subsection 3.3.1. The *TBP* in the problem dependent regime is also studied by Mukherjee et al. [71] and Zhong, Huang, and Liu [87], however they consider a problem complexity based also upon variance making their results not so relevant to our setting. The *problem independent* regime for the *TBP* is studied by Cheshire, Ménard, and Carpentier [26], we also present their results in Table 3.1 for comparison across the different regimes.

Monotone constraint, *MTBP*. We then consider the problem where on top of assuming that the distributions are sub-Gaussian, we assume that the sequence of means $(\mu_k)_{k\in[K]}$ is monotone - this is problem *MTBP*. This specific instance of the *TBP* is introduced within the context of drug dosing by Garivier et al. [39]. In this paper, the authors provide an algorithm for the fixed confidence setting that is optimal asymptotically, in the fixed confidence regime. However the definition of the algorithms, as well as the provided optimal error bound, are defined in an implicit way and not so easy to relate in a simple way to the gaps Δ_i moreover it is not clear how to translate a result from the fixed confidence setting to the fixed budget one. On the other hand, 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. [33]. They describe an algorithm structurally similar to ours, using a binary tree with infinite extension however they consider a simpler setting where the probability of correct labeling is fixed as some $\delta > \frac{1}{2}$ and go on to show that there exists an algorithm that will correctly insert an element with probability at least $1-\delta$ in $\mathcal{O}\big(\log\big(\frac{K}{\delta}\big)\big)$ steps. For further literature on the related yet different problem of noisy binary search, see Feige et al. [33], Ben-Or and Hassidim [8], Emamjomeh-Zadeh, Kempe, and Singhal [30], Nowak [73]. Again, these papers consider settings with more structural assumptions than our own and are focused on the problem independent, fixed confidence regime. The *problem independent* regime for the *MTBP* is studied by Cheshire, Ménard, and Carpentier [26], we also present their results in Table 3.1 for comparison across the different regimes.

In this work, we prove that, up to universal multiplicative constants and additive $\log(K)$ terms in the exponential, the optimal error probability is $\exp(-T\min_k \Delta_k^2)$, which highlights the somewhat surprising fact that this structured monotone *TBP* problem is akin to a one armed *TBP*- see Subsection 3.3.2. We provide the Problem

[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 or "best arm identification" problem where the aim is to find the arm that realises the maximum.

Dependent Monotone *TBP* (ProbDep-Explore) algorithm that matches this bound, see Section 3.4.

Concave constraint, *CTBP*. We next consider the problem where on top of assuming that the distributions are sub-Gaussian, we assume that the sequence of means $(\mu_k)_{k\in[K]}$ is concave - this is problem *CTBP*. Again, in the problem independent regime the *CTBP* has been studied by Cheshire, Ménard, and Carpentier [26]. In the problem dependent regime however, to the best of our knowledge, the *CTBP* has not been studied in the literature. However the related problems of estimating a concave function and optimising a concave function are well studied in the literature. Both problems are considered primarily in the continuous regime which makes comparison to the K-armed bandit setting difficult. 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. [79], where 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. [72][Chapter 9], and Wang et al. [82], Agarwal et al. [1], and Liang, Narayanan, and Rakhlin [66] to name a few, all of them in a continuous setting with dimension d. The focus of this literature is however very different to ours and Cheshire, Ménard, and Carpentier [26], as the main difficulty under their assumption is to obtain a good dependence in the dimension d, and with this in mind logarithmic factors are not very relevant.

In this work, we prove that, up to universal multiplicative constants and additive $\log(K)$ terms in the exponential, the optimal error probability is $\exp(-T\min_k \Delta_k^2)$, which highlights the somewhat surprising fact that this structured concave *TBP* problem is also akin to a one armed *TBP*- see Subsection 3.3.3. We provide the Problem Dependent Concave *TBP* (CTB) algorithm that matches this bound, see Section 3.4.

Organisation of the paper This paper is structured as follows. In Section 3.2 we formally introduce the *TBP* setting along with the monotone and concave shape constraints. We also describe the performance criterion - probability of error, we will be primarily using for the duration of the paper. Following this, upper and lower bounds on probability of error for all shape constraints are presented in Section 3.3. Descriptions of algorithms achieving said upper bounds can be found in Section 3.4. The results are discussed and compared to related work in Section 3.5. In Appendix 3.9 we conduct some preliminary experiments to explore how our theoretical results translate in practice. All proofs are found in the Appendix.

3.2 Setting

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 $\sigma^2 \geq 0$. We remind the learner that 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.

Let $\mathcal{B} := \mathcal{B}(K, \sigma^2)$ be the set of all bandit problems as presented above, i.e. where the distributions ν_k of the arms are all σ^2 sub-Gaussian.

In what follows, we assume that all $\underline{\nu} \in \mathcal{B}$, and we write μ_k for the mean 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 based on their means. 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 $k_t \in [K]$ 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 $\widehat{Q} \in \{-1, 1\}^K$ and the aim is that it matches the unknown vector Q as well as possible.

Unstructured case *TBP* In the *problem dependent* regime, for $\bar{\Delta} \in \mathbb{R}_+^K$, we consider the following class of problems

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

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. We denote by $\Delta\mathcal{B}_m = \{\bar{\Delta} \in \mathbb{R}_+^K : \exists \nu \in \mathcal{B}_m, \forall k \in [K],\ |\mu_k - \tau| = \bar{\Delta}_k\}$ the set of possible vectors of gaps in $\mathcal{B}_m$ - i.e. the set of sequences $\bar{\Delta}$ that would correspond to at least one problem in $\mathcal{B}_m$. In the *problem dependent* regime, for $\bar{\Delta} \in \Delta\mathcal{B}_m$, we consider the following class of problems

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

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. We denote by $\Delta\mathcal{B}_c = \{\bar{\Delta} \in \mathbb{R}_+^K : \exists \nu \in \mathcal{B}_c, \forall k \in [K],\ |\mu_k - \tau| = \bar{\Delta}_k,\ \exists l : \mu_l \geq \tau\}$ the set of possible vectors of gaps in $\mathcal{B}_c$ where at least one arm is above threshold - i.e. the set of sequences $\bar{\Delta}$ that would correspond to at least one problem in $\mathcal{B}_c$ where at least one arm is above threshold. In the *problem independent* regime, for $\bar{\Delta} \in \Delta\mathcal{B}_c$, we consider the following class of problems

$$\mathcal{B}_c^{\bar{\Delta}} := \left\{\nu \in \mathcal{B}_c : \forall k < K, |\mu_k - \tau| \in \left[\frac{\bar{\Delta}_k}{2}, 3\frac{\bar{\Delta}_k}{2}\right]\right\}.$$

Remark 10. The classes of problems $\mathcal{B}^{\bar{\Delta}}, \mathcal{B}_m^{\bar{\Delta}}, \mathcal{B}_c^{\bar{\Delta}}$ contain bandit problems in resp. $\mathcal{B}, \mathcal{B}_m, \mathcal{B}_c$ that are 'local' around $\bar{\Delta}$ in the sense that while the sign of $\mu_k - \tau$ is arbitrary - although severely restricted by the shape constraint when it comes to $\mathcal{B}_m^{\bar{\Delta}}, \mathcal{B}_c^{\bar{\Delta}}$ - the gap of arm k is fixed to being - approximately, for the concave case set $\mathcal{B}_c^{\bar{\Delta}}$ - $\bar{\Delta}_k$. This implies that in each case and on top of the respective shape constraint, we restrict ourselves to a small class of problems whose complexity is entirely characterised by $\bar{\Delta}$, in a *problem dependent sense.*

Strategy A strategy is a sequence of functions that maps the information gathered in the past to an arm and finally to a classification. Precisely, if we denote by I_t the information available to the player at time t, that is $I_t = \{Y_1, Y_2, \dots, Y_t\}$, with the convention $I_0 = \emptyset$. Then a strategy $\pi = \big((\pi_t)_{t\in[T]}, \widehat{Q}^\pi\big)$ is given by a sampling rule $\pi_t(I_{t-t}) = k_t \in [K]$ and a classification rule $\widehat{Q}^\pi(I_T) = \widehat{Q} \in \{-1,1\}^K$.

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

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

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. However, the focus of this paper is on the *problem dependent* regime where, as usual, we consider as a performance criterion rather the related *probability of error*

$$e_T^{\underline{\nu},\pi} = \mathbb{P}_{\underline{\nu}}\Big(\exists k \in [K] : \widehat{Q}_k^\pi \neq Q_k\Big)\,.$$

When it is clear from the context we will remove the dependence on the bandit problem $\underline{\nu}$ and/or the strategy π. Note that if we denote by $\bar{\Delta}_{\min} = \min_{k\in[K]} \bar{\Delta}_k$ the minimum of the gaps then

$$r_T^{\underline{\nu},\pi} \geq \bar{\Delta}_{\min} e_T^{\underline{\nu},\pi}\,.$$

Consider a set of bandit problems $\tilde{\mathcal{B}} \subset \mathcal{B}$. The minimax optimal probability of error on $\tilde{B}$ is then

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

We will study this quantity over the local classes $\mathcal{B}^{\bar{\Delta}}, \mathcal{B}_m^{\bar{\Delta}}, \mathcal{B}_c^{\bar{\Delta}}$.

Remark 11. As argued above, the classes $\mathcal{B}^{\bar{\Delta}}, \mathcal{B}_m^{\bar{\Delta}}, \mathcal{B}_c^{\bar{\Delta}}$ contain only bandit problems that satisfy their respective shape constraint and whose complexity is entirely characterised by $\bar{\Delta}$, in a *problem dependent sense.* Studying the minimax probability of error over these very restricted classes is therefore a very meaningful way of studying the problem dependent regime of structured *TBP* problems - and we expect this probability of error to heavily depend on $\bar{\Delta}$. The focus of this paper is to characterise this dependence in a tight manner.

3.3 Minimax rates

In this section we present upper and lower bounds on probability of error for all three shape constraints. Given a vector $\bar{\Delta} \in \mathbb{R}_+^K$ we denote $\bar{\Delta}_{\min} = \min_{k\in[K]} \bar{\Delta}_k$.

3.3.1 Problem dependent unstructured setting *TBP*

The unstructured thresholding bandit in the problem dependent regime has already been considered in the literature. We remind results from Locatelli, Gutzeit, and

Carpentier [68], where they provide tight upper and lower bounds over $e^*_T(\mathcal{B}^{\bar{\Delta}})$, for any $\bar{\Delta} \in \mathbb{R}^K_+$. In our context they prove that

$$\exp\Big(-\frac{3}{\sigma^2}\frac{T}{H} - 4\sigma^{-2}\log(12(\log T+1)K)\Big) \le e^*_T(\mathcal{B}^{\bar{\Delta}})$$
$$\le \exp\Big(-\frac{1}{64\sigma^2}\frac{T}{H} + 2\log((\log T+1)K)\Big),$$

where $H = \sum_{i:\bar{\Delta}_i>0} 1/\bar{\Delta}_i^2$ - see Theorems 1 and 2 by Locatelli, Gutzeit, and Carpentier [68]. This implies that up to multiplicative universal constants and whenever $T \ge H\sigma^2\log(\log(T)+K)$, it holds that

$$-\log\Big(e^*_T(\mathcal{B}^{\bar{\Delta}})\Big) \asymp \frac{1}{\sigma^2}\frac{T}{H},$$

and upper and lower bound match up to universal multiplicative constants in the exponential of the error probability. The quantity H is therefore the problem dependent quantity that characterises the difficulty of the problem. Note that of course, the APT algorithm by Locatelli, Gutzeit, and Carpentier [68] does not take any information on the class - $\bar{\Delta}$, but also σ^2 - as parameters, and is essentially parameter free.

In this paper, we won't therefore discuss further this unstructured setting - the reminder provided here is only to be taken as a benchmark for the rest of the paper. We will on the other hand focus on the structured problems - monotone and concave and study how the minimax error probability evolves, in particular depending on the problem-dependent quantities $\bar{\Delta}$.

3.3.2 Problem dependent monotone setting

Given a class of problems $\mathcal{B}^{\bar{\Delta}}_m$ for some $\bar{\Delta} \in \Delta\mathcal{B}_m$, the following theorem provides a lower bound on the probability of error for any strategy π. The proof of Theorem 27 can be found in Appendix 3.7.

Theorem 27. *Let $\bar{\Delta} \in \Delta\mathcal{B}_m$. For any strategy π there exists a monotone bandit problem $\underline{\nu} \in \mathcal{B}^{\bar{\Delta}}_m$ such that*

$$e^{\nu,\pi}_T \ge \frac{1}{4}\exp\Big(-\frac{T\bar{\Delta}^2_{\min}}{\sigma^2}\Big).$$

Now the following theorem gives an upper bound on the probability of error for the `ProbDep-Explore` algorithm. The proof of Theorem 28 can be found in Appendix 3.7.

Theorem 28. *Let $\nu \in \mathcal{B}_m$ associated with arm gaps Δ, and assume that $T > 36\log(K)$. The algorithm `ProbDep-Explore` satisfies the following bound on error probability:*

$$e^{\nu,\texttt{ProbDep-Explore}}_T \le \exp\Big(-c_{\text{mon}}\frac{T\Delta^2_{min}}{\sigma^2} + c'_{\text{mon}}\log(K)\Big)$$

where $c_{\text{mon}} = 1/48$ and $c'_{\text{mon}} = 12$.

The parameter free algorithm `ProbDep-Explore` is described in Sections 3.4 - see also Appendix 3.7.

The assumption on T is reasonable as in the monotone setting it is clear no algorithm can gain enough information in less than $\log(K)$ pulls, see Cheshire, Ménard, and Carpentier [26]. Note that combining both bounds yields that whenever $T > 36\log(K)/\bar{\Delta}^2_{\min}$:

$$-\log\left(e_T^*(\mathcal{B}_m^{\bar{\Delta}})\right) \asymp \frac{1}{\sigma^2} T \bar{\Delta}_{\min}^2,$$

and upper and lower bound match up to universal multiplicative constants in the exponential of the error probability. Perhaps surprisingly, the number of arms plays no role in this rate - as long as we assume that $T > 36\log(K)/\bar{\Delta}_{\min}^2$. Only the minimal arm gap appears, and this amounts to saying that when $T > 36\log(K)/\bar{\Delta}_{\min}^2$, this problem is not more difficult - in order, up to universal multiplicative constants in the exponential - than a one-armed *TBP* with gap $\min_k \Delta_k$! And that in a sense, even if we knew in our monotone problem the position of all means but one - the arm with minimal gap - with respect to the threshold, the problem would not be significantly easier.

3.3.3 Problem dependent concave setting

Given a class of problems $\mathcal{B}_c^{\bar{\Delta}}$ for some $\bar{\Delta} \in \Delta\mathcal{B}_c$ the following theorem provides a lower bound on the probability of error for any strategy π. The proof of Theorem 29 can be found in Appendix 3.8.

Theorem 29. *Let $\bar{\Delta} \in \Delta\mathcal{B}_c$. For any strategy π there exists a problem $\nu \in \mathcal{B}_c^{\bar{\Delta}}$ such that*

$$e_T^{\nu,\pi} \geq \frac{1}{4}\exp\left(-9\frac{T\bar{\Delta}_{\min}^2}{\sigma^2}\right).$$

Now the following theorem gives an upper bound on the probability of error for the CTB algorithm. The proof of Theorem 30 can be found in Appendix 3.8.

Theorem 30. *Let $\nu \in \mathcal{B}_c$ with associated gaps Δ and assume $T > 108\log(K)$. The algorithm CTB has the following bound on error,*

$$e_T^{\nu,\mathit{CTB}} \leq 3\exp\left(-c_{\text{con}}\frac{T\Delta_{\min}^2}{\sigma^2} + c'_{\text{con}}\log(K)\right)$$

where $c_{\text{con}} = 1/576$ and $c'_{\text{con}} = 12$.

The parameter free algorithm CTB is described in Sections 3.4 - see also Appendix 3.8.

The assumption on T is reasonable as in the monotone setting it is clear no algorithm can gain enough information in less than $\log(K)$ pulls, see Cheshire, Ménard, and Carpentier [26]. Note that combining both bounds yields that whenever $T > 108\frac{\log(K)}{\Delta_{\min}^2}$:

$$-\log\left(e_T^*(\mathcal{B}_m^{\bar{\Delta}})\right) \asymp \frac{1}{\sigma^2} T \bar{\Delta}_{\min}^2,$$

and upper and lower bound match up to universal multiplicative constants in the exponential of the error probability. Similar comments can be made here as in the case of the monotone *TBP* in Section 3.3.2: the convex *TBP* is also as difficult as a one-armed *TBP* with gap $\min_k \Delta_k$.

3.4 Optimal algorithms in the problem dependent regime

3.4.1 Monotone case *MTBP*

We assume in this section, without loss of generality, instead of considering K arms, we consider for technical reasons $K + 2$ arms adding 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 σ^2-sub-Gaussian the addition of two such arms will not invalidate our proofs, see Appendix 3.7. We do this to ensure that, after re-indexing of the arms and adapting the number of arms, $\tau \in [\mu_1, \mu_K]$.

To match a minimax rate as described in Section 3.3 we will utilise a modified version of the MTB algorithm described by Cheshire, Ménard, and Carpentier [26]. The algorithm `ProbDep-Explore` performs a random walk on the set of arms $[K]$ as a binary tree. We consider the binary tree as Cheshire, Ménard, and Carpentier [26] with an specific extension akin to that by Feige et al. [33].

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 $P(\texttt{root}) = \texttt{root}$.

Extended Binary Tree We extend the above Binary tree in the following manner. For a leaf v we replace the condition $R(v) = L(v) = \emptyset$ with the following: for any leaf $v = \{L, M, R\}$ we set $R(v) = \tilde{v}$ where $\tilde{v} = \{L, M, R\}$ and set $L(v) = \emptyset$. Note that $\tilde{v}$ is also a leaf therefore iterative application this relation will lead to an infinite extension. The result being that each leaf in our original binary tree is now the root of an infinite chain of identical nodes, see Figure 3.1. For practical purposes we need only consider such an extension up to depth T and can simply cut the tree at this depth.

Remark 12. We set $L(v) = \emptyset$ for some leaf v during the extension of the binary tree as by construction all leaves of the original binary tree are of the form $\{v = \{L, M, R\} : R = L+1 \text{ and } M = L\}$.

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 `ProbDep-Explore` to backtrack and this is why `ProbDep-Explore` performs a binary search *with corrections*.

`ProbDep-Explore` algorithm First, define the following integers

$$T_1 := \lceil 6 \log(K) \rceil \qquad T_2 := \left\lfloor \frac{T}{3T_1} \right\rfloor . \tag{3.1}$$

The algorithm `ProbDep-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 $\{v_t(j) : j \in \{l, m, r\}\}$ each T_2 times. Let the sample mean of arm $v_t(j)$ be denoted $\hat{\mu}_{j,t}$. `ProbDep-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 does not contain the threshold, then the algorithm backtracks to the parent of the current node, otherwise ProbDep-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, that is $v(r) = v(l) + 1$, then due to the extension of our binary tree $R(v_t) = L(v_t) = \tilde{v}_t$ where $\tilde{v}$ is a duplicate of v_t. Hence $v_{t+1} = \tilde{v}_t$. Via this mechanism the ProbDep-Explore algorithm essentially gives additional preference the the node v_t. See ProbDep-Explore for details. We now formally state the parameter free ProbDep-Explore algorithm (Problem Dependent Monotone Thresholding Bandit Algorithm). We rely on the assumption $T > 36\log(K)$, see Theorem 28 to ensure $T_2 \geq 1$.

```
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
        $v_{t+1} = P(v_t)$
    end
    else
        if $\hat{\mu}_{m,t} \leq \tau \leq \hat{\mu}_{r,t}$ then
            $v_{t+1} = R(v_t)$
        end
    end
    else
        if $\hat{\mu}_{l,t} \leq \tau \leq \hat{\mu}_{m,t}$ then
            $v_{t+1} = L(v_t)$
        end
    end
end
Set $\hat{a} = v_{T_1+1}(r)$
Output: $(\hat{a}, \widehat{Q})$ :   $\widehat{Q}_k = 2\mathbb{1}_{\{k \geq \hat{a}\}} - 1$
```

Algorithm 18: ProbDep-Explore

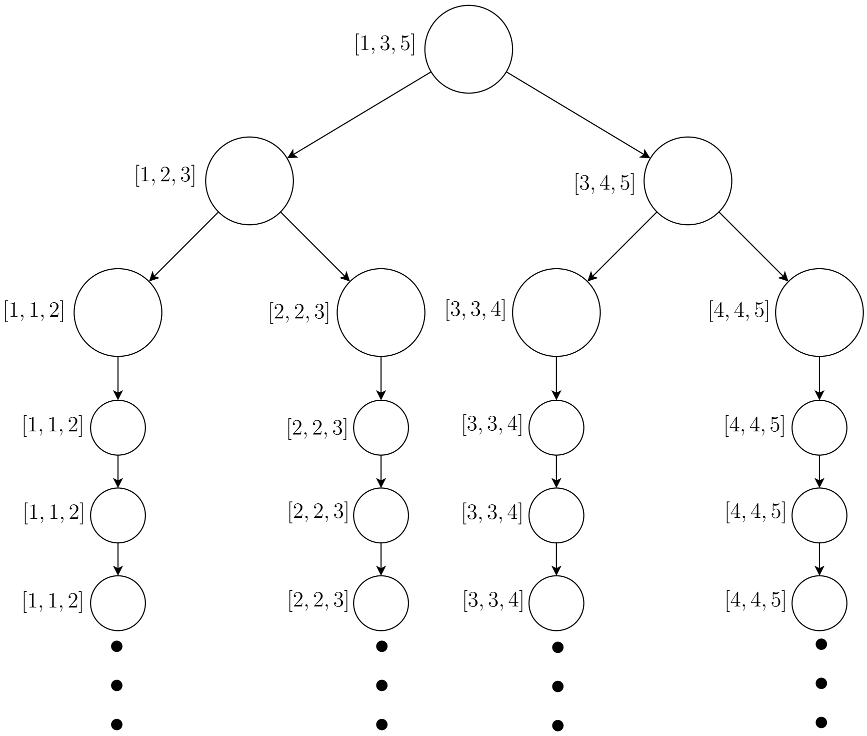

FIGURE 3.1: Extended binary tree for $K = 5$

Remark 13 (Adaptation of ProbDep-Explore to a non-increasing sequence, PD-DEC-MTB). ProbDep-Explore is applied for a monotone non-decreasing sequence $(\mu_k)_{k\in[K]}$, and it is easy to adapt it to a monotone non-increasing sequence $(\mu_k)_{k\in[K]}$. In this case, we transform the label of arm k into $K - k$, and apply ProbDep-Explore to the newly labeled problem - where the mean sequence in now non-decreasing. We refer to this modification as PD-DEC-MTB.

Remark 14 (Relaxing the monotone assumption). By inspecting the proof of Theorem 28 in Appendix 3.7 we can obtain the same guarantee for a larger class of problem than one with increasing means. Indeed we only need that there exists an arm for which all the arms before it have a mean below the threshold and all arm after have a mean above the threshold. Precisely the bound of Theorem 28 holds also for problems that belongs to

$$\begin{aligned}\mathcal{B}_{rm} :=\{\nu \in \mathcal{B} :\ \exists k \in [1, K], \forall j \le k\ \mu_j \le \tau,\\ \forall j \ge k+1\ \mu_j \ge \tau\}\ .\end{aligned}$$

Note the same remark also applies for problems with monotone non-increasing sequence.

3.4.2 Concave case *CTBP*

We assume in this section, without loss of generality, instead of considering K arms, we consider for technical reasons $K + 2$ arms adding two deterministic arms 0 and $K + 1$ with respective means $\mu_0 = \mu_{K+1} = -\infty$. While we assume that the distributions of the original K arms are σ^2-sub-Gaussian the addition of two such arms will not invalidate our proofs, see Appendix 3.8. We do this to ensure that after re-indexing $\tau > \mu_1, \mu_K$.

As in the monotone case we construct a binary tree to span the arms of the bandit problem. The construction of this tree is identical to that described in Section 3.4.1 but without the infinite extension. We will use a variant off the ProbDep-Explore Algorithm, Grad-Explore to move around the tree. The difference is that Grad-Explore bases its movement off the estimated gradients of the arms as opposed to their sample means. The objective of Grad-Explore is to find an arm with corresponding mean above threshold. Once such an arm has been identified we split our problem into two “relaxed monotone" bandit problems - see Remark 14, one increasing and one decreasing. We then run ProbDep-Explore and PD-DEC-MTB respectively. We split our budget evenly across the three algorithms: Grad-Explore, ProbDep-Explore and PD-DEC-MTB.

Grad-Explore algorithm As with ProbDep-Explore the algorithm Grad-Explore is essentially a random walk on the 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 $\{v_t(l), v_t(l)+1, v_t(m), v_t(m)+1, v_t(r), v_t(r)+1\}$ each T_2 times. As in Section 3.4.1, we adopt the convention that the arm $K+1$ is a Dirac distribution at $-\infty$. Let the sample mean of arm $v_t(j)$ be denoted $\hat{\mu}_{j,t}$ and the sample mean of arm $v_t(j)+1$ be denoted $\widehat{\mu}_{j+1,t}$. Let the estimated local gradient at arm j, that is $\widehat{\mu}_{j,t} - \widehat{\mu}_{j+1,t}$ denote $\widehat{\nabla}_{j,t}$. Grad-Explore will use these estimates to decide which node to explore next. If an error is detected - i.e. the left most or right most gradient is negative or positive respectively, then the algorithm backtracks to the parent of the current node, otherwise Grad-Explore acts as the deterministic binary search for the maximum mean, $\max_{i\in[K]} \mu_i$. More specifically, if there is an anomaly, $\left(\widehat{\nabla}_{l,t}, \widehat{\nabla}_{r,t}\right) \notin (\mathbb{R}_+, \mathbb{R}_-)$, then the next node is the parent $v_{t+1} = P(v_t)$, otherwise if $\widehat{\nabla}_{m,t} < 0$ the next node is the left child $v_{t+1} = L(v_t)$ and if $\widehat{\nabla}_{m,t} \geq 0$ the next node is the right child $v_{t+1} = R(v_t)$. See Algorithm 19 for details.

```
Initialization: $v_1 = \texttt{root}$
for $t = 1 : T_1$ do
    $S_{t+1} = S_t$
    for each $k \in v_t$ sample $\frac{T_2}{12}$ times the arms $k, k+1$ if $\exists k \in \{l, m, r\} : \widehat{\mu}_k > \tau$
     then
        Append arm $k$ to the list $S_{t+1}$
        $v_{t+1} = v_t$
    end
    else
        If$\left(\widehat{\nabla}_{l,t}, \widehat{\nabla}_{r,t}\right) \notin (\mathbb{R}_+, \mathbb{R}_-)$ $v_{t+1} = P(v_t)$
    end
    else
        If$\widehat{\nabla}_{m,t} \geq 0$ $v_{t+1} = R(v_t)$
    end
    else
        if $\widehat{\nabla}_{m,t} < 0$ then
            $v_{t+1} = L(v_t)$
        end
    end
end
```

Algorithm 19: Grad-Explore

```
run Grad-Explore
Output: list $S_{T_1}$ if $|S_{T_1}| \leq \frac{T_1}{4}$ then
    Output: $\widehat{Q} = \{-1\}^K$
end
else
    $\hat{a} = \text{Median}(S_{T_1})$
    $l$ = output of PD-DEC-MTB on set of arms $[1, \hat{a}]$ budget: $\frac{T}{3}$
    $r$ = output of Explore on set of arms $[\hat{a}, K]$ budget: $\frac{T}{3}$
    Output: $\widehat{Q}$ :    $\widehat{Q}_k = 1 - 2\mathbb{1}_{k<l} - 2\mathbb{1}_{k>r}$
end
```

Algorithm 20: ProbDep-CTB

For the arms whose means are below threshold, due to the concave property gradients are essentially greater than $\bar{\Delta}_{\min}$ and can easily be estimated. Above threshold however gradients are less than $\bar{\Delta}_{\min}$ and are relatively hard to estimate. Therefore, although on the face Grad-Explore is in part a binary search for the arm with maximum mean, in reality this is not feasible. The true utility of Grad-Explore to the learner is to act as a binary search for the "set" of arms above threshold. If we refer to nodes containing an arm $k : \mu_k > \tau$ as "good nodes" the idea behind Grad-Explore is to spend a sufficient amount of time in exploring this set of nodes and adding "good arms" - i.e ones with a corresponding mean above threshold, to the list S. We can then output such an arm with high probability when outputting the median of S_{T_1}.

Once we have identified our arm above threshold we split our problem into two bandit problems where the classification can be done by binary search, see Remark 14 and 13. We can thus then apply ProbDep-Explore and PD-DEC-MTB. Precisely, the complete procedure, namely ProbDep-CTB (Problem Dependent- Concave Threshold Bandits), is detailed in Algorithm 20.

3.5 Discussion

3.5.1 Algorithms Explore and ProbDep-CTB

Both the ProbDep-Explore and ProbDep-CTB are based upon a binary search with corrections, this allows them to exploit the structure of the shape constraints reducing the problems to sets of arms with cardinally of order $\log(K)$, something in sharp contrast to existing algorithms for the vanilla setting. The difference between ProbDep-Explore and ProbDep-CTB is that while ProbDep-Explore works exclusively on a binary tree based upon the classification of an arms mean above or below threshold, the sub algorithm Grad-Explore of ProbDep-CTB bases a binary tree on positive or negative gradient. Therefore ProbDep-Explore acts as a search for the point the arms cross threshold while Grad-Explore acts as a search for the arm $k^* = \arg\max_k(\bar{\Delta}_k)$. Another more subtle difference is that on a "good decision" at time t - i.e when the sample means are well concentrated up to $\bar{\Delta}_{\min}$, ProbDep-Explore will make a step in the right direction. The same cannot be said for Grad-Explore as we can only guarantee that the increments between arms are greater than $\bar{\Delta}_{\min}$ for arms below threshold, this is a direct result of the concave property. Therefore the true utility of Grad-Explore is not to find k^* but to find any arm $k : \mu_k > \tau$.

It is worth noting that both algorithms described in this paper are parameter free, being adaptive not only to the hardness of the problem characterised by the gaps $\bar{\Delta}$, but also to the underlying sub-Gaussian assumption parameter σ^2.

3.5.2 Problem classes and optimality

In the monotone and concave settings we consider a very narrow class of problems and argue our classes are relevant for characterising the problem dependent regime - i.e. are narrow enough.

- In the monotone setting this is obvious as the class of problems is defined by a specific vector $\bar{\Delta} \in \mathbb{R}_+^K$, so that all problems in this class have a similar complexity, bear in mind that our algorithms do not need to know $\bar{\Delta}_{\min}$ or any aspect of $\bar{\Delta}$. In fact, when constructing our lower bound, we just need a class with two problems where, given a first problem, we simply switch the arm with minimal gap $\bar{\Delta}_{\min}$ from below to above threshold in order to obtain the second problem - see the proof of Theorem 27.

- In the concave setting this approach is unfeasible as under the concave constraints the class of problems defined by a specific vector of gaps $\bar{\Delta} \in \mathbb{R}_+^K$ has very often cardinality 1 which is nonsensical for a lower bound. Instead, given a specific vector $\bar{\Delta} \in \mathbb{R}_+^K$ we consider a class of problems with gaps within a proportional tolerance of $\bar{\Delta}$. This class is designed to be as narrow as possible while still containing multiple problems which disagree on the placement of certain arms above or below threshold. In fact, when constructing our lower bound, we just need a class with two problems where, starting from a first problem, we simply flip the arm with minimal gap and translate other means vertically in such a way to preserve concavity - see the proof of Theorem 27.

In both cases, we prove that for T large enough, the problem dependent optimal probability of error is of order

$$\exp(-T\bar{\Delta}_{\min}^2/\sigma^2),$$

up to universal multiplicative constants inside and outside the exponential. This implies that from a problem dependent perspective, both problems are as difficult as a one armed bandit problem where we just want to decide whether the arm with minimal gap $\bar{\Delta}_{\min}$ is up or down the threshold, which is quite surprising - as the number of arms plays therefore no role asymptotically. While the lower bounds are relatively simple, the upper bounds are more interesting and challenging.

3.5.3 Comparison of rates between settings

Table 3.1 presents a comparison of results across the problem independent and dependent regimes. Although the results are not immediately comparable between the regimes, of particular interest is the difference in rates across the monotone and concave settings in the problem independent regime compared to the lack of difference between said rates in the problem dependent regime.

problem:	independent	dependent
Unconstrained	$\sqrt{\frac{K\log K}{T}}$	$\exp\left(-\frac{T}{H}\right)$
Monotone	$\sqrt{\frac{\log K\vee 1}{T}}$	$\exp\left(-T\bar{\Delta}_{\min}^2\right)$
Concave	$\sqrt{\frac{\log\log K\vee 1}{T}}$	$\exp\left(-T\bar{\Delta}_{\min}^2\right)$

TABLE 3.1: Order of the optimal problem dependent probability of error, and of the problem independent expected simple regret for the three structured *TBP*, 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 both in and outside the exponential. The first line concerns the problem independent setting and the simple regret, see Cheshire, Ménard, and Carpentier [26]. The second line concerns the problem dependent setting and the probability of error, the main focus of this paper. The results for the monotone and concave are novel and can be found in this paper, see Section 3.3. The results for the unstructured setting are by Locatelli, Gutzeit, and Carpentier [68], where they take $H = \sum_{i=1}^{K} \bar{\Delta}_i^{-2}$

In both the monotone and concave setting an initial lower bound is one which does not depend upon K - imagine the setting in which a learner places their entire budget on the two arms either side of the threshold. We show that in the problem dependent regime a binary search with corrections can match this bound, up to a $\log(K)$ term which disappears for large T. The intuition behind this is that as the depth of the tree is only $\log(K)$ the binary search can quickly find the point of interest and spend the majority of its time there. As both the concave and monotone problems can be solved with a binary search they therefore have the same rate.

In the problem independent regime the situation is slightly more nuanced. In terms of lower bounds one is no longer restricted to a narrow class of problems and can consider a number of different problems, all close in terms of distributional distance but nevertheless disagreeing on the classification of certain arms above or below threshold. The cardinality of these sets differs between the monotone and concave setting - being $\log(K)$ and $\text{loglog}(K)$ respectively. This then leads to a difference in the lower bound. Upper bounds naturally must follow suit, while an adaptation of the standard binary search is still optimal in the monotone case in the concave case an algorithm using a binary search on a log scale is required. The above is by no means a rigorous explanation but hopefully gives the reader some intuition behind the differences in rates between the problem dependent and independent regimes, for more detail refer to Cheshire, Ménard, and Carpentier [26].

3.5.4 Related work

Unstructured *TBP* As mentioned in Section 3.3 and demonstrated in Table 3.1 the unstructured problem dependent *TBP* is already well studied in the literature, see [23, 24] for the fixed confidence setting and Chen et al. [23], Locatelli, Gutzeit, and Carpentier [68], and Mukherjee et al. [71], Zhong, Huang, and Liu [87] for the fixed budget. As mentioned in Section 3.1, [68] is most relevant to our setting as they consider the fixed budget case. Their rate for the unstructured case depends upon the distribution of gaps across all the arms, which is of course to be expected. This again highlights the fact that the rate for the monotone setting depends only upon the minimum gap - that is the one adjacent to the threshold.

Monotone constraint *MTBP* The *MTBP* problem was first introduced by Garivier et al. [39] in the context of drug dosing. Their results are in contrast to ours as they consider the *fixed confidence* setting. Furthermore the algorithm proposed is shown to be optimal only in the asymptotic case, i.e when the confidence $1-\delta$ converges to 1. The monotone shape constraint of the *MTBP* implies it is related to a 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. A naive approach to the *MTBP* would be a binary search with $\frac{T}{\log(K)}$ samples at each step of the binary search. However for our setting this is not optimal, even in the problem independent case, see [26]. In [33] this issue is solved by introducing a binary search with corrections. They describe an algorithm structurally similar to `ProbDep-Explore`, using a binary tree with infinite extension however they consider a simpler setting where the probability of correct labeling is fixed as some $\delta > \frac{1}{2}$ and go on to show that there exists an algorithm that will correctly insert an element with probability at least $1-\delta$ in $\mathcal{O}\big(\log\big(\frac{K}{\delta}\big)\big)$ steps. For further literature on the related yet different problem of noisy binary search see, [33], Ben-Or and Hassidim [8], Emamjomeh-Zadeh, Kempe, and Singhal [30], [73]. Again, these papers consider settings with more structural assumptions than our own and are focused on the problem independent, fixed confidence regime. The minimax rate on expected regret for the problem independent `MTB` is presented by Cheshire, Ménard, and Carpentier [26].

For us the adaptation of the algorithm in Cheshire20 to the problem dependent regime is not obvious. An important fact in our problem dependent regime is that the number of arms K stops appearing in the error bound which is of order $\exp(-cT\Delta^2_{\min})$ whenever T is large enough, i.e. larger than $\log(K)/\Delta^2_{\min}$. In Cheshire, Ménard, and Carpentier [26], the number of arms appeared in all bounds and was the main topic of study therein - the bound for the monotone problem was $\sqrt{\log(K)/T}$. A key interesting phenomenon here is that somewhere between the problem independent and problem dependent regime, K stops playing a role. This implies that a very different dynamic is happening in the problem dependent regime, as compared to the problem independent regime.

Precisely in Cheshire, Ménard, and Carpentier [26] they consider a sequence of events ξ_t that depend on K and occur with constant probability - which is the target probability of error in the worst case. Lemma 15 therin then applies Hoeffding's-Azuma to the summation of the indicator functions of said events to achieve a bound on the probability of making too many bad decisions in the tree. In order to achieve a problem dependent bound, we consider events ξ_t which are problem dependent - they depend on $\Delta_{\min}$ - but NOT on K. This event is now problem dependent and the probability of its complement depends on both $\Delta_{\min}$ and $T/\log(K)$ (the number of times we sample each arm), i.e. is of order $\exp(-cT\Delta_{\min}/\log K)$, which, interestingly, is NOT the target probability of error in the problem dependent regime, but is quite larger. Our Lemma 22 is then substantially more than just a problem dependent adaptation of Lemma 15 of Cheshire, Ménard, and Carpentier [26], as we need to leverage the fact that there are many events ξ_t - here $\log K$ - in order to bypass the fact that the probability of each individual ξ_t depends on K in our setting. We use a Chernoff bound to bound the sum of the indicator functions of said events - and then in turn the probability of error - by $\exp(-T\Delta_{\min}^2)$ - which is much smaller than the probability of each individual ξ_t. This phenomenon is not needed in Cheshire, Ménard, and Carpentier [26].

Another point in favour of the PD-MTB is that it is significantly simpler than that of the MTB of Cheshire, Ménard, and Carpentier [26]. We use an infinite extension to the binary tree which allows it to take the final node as output. This means we don't require an additional subroutine to choose from a list of arms the algorithm has collected.

Concave constraint To the best of our knowledge the *CTBP* was first introduced in [26] in the *problem independent regime*. However the related problems of estimating a concave function and optimising a concave function are well studied in the literature. Both problems are considered primarily in the continuous regime which makes comparison to the K-armed bandit setting difficult. The problem of estimating a concave function has been thoroughly studied in the noiseless setting, and also in the noisy setting, see e.g. [79], where 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. [72][Chapter 9], and [82, 1, 66] to name a few, all of them in a continuous setting with dimension d. The focus of this literature is however very different to ours and [26], as the main difficulty under their assumption is to obtain a good dependence in the dimension d, and with this in mind logarithmic factors are not very relevant.

3.6 Potential further work

3.6.1 Algorithms that are problem dependent and minimax-optimal

As described earlier after the related theorems, our algorithm `ProbDep-Explore` is optimal for minimising the probability of error, in a problem dependent sense, and up to universal multiplicative constants in the exponential. A relevant question is on whether it is possible to construct a strategy that is optimal both in this problem dependent sense, but also in a problem independent sense - i.e. global minimax - when it comes to the simple regret.

While designed for the problem independent regime - and reaching in this regime the minimax optimal simple regret of order $\sqrt{\frac{\log K}{T}}$ - we conjecture the `MTB` algorithm, described by Cheshire, Ménard, and Carpentier [26] is optimal also in the problem dependent regime, i.e. that it achieves an upper bound on the probability of error of same order as that of `ProbDep-Explore` in Theorem 28. However note that to prove such an opitmaility, at least for us, would be none trivial, see the above Section 3.5.4.

As with `ProbDep-Explore` the `MTB` algorithm takes a monotone bandit problem mapped to a binary tree - although without the infinite extension, as input. The `MTB` algorithm then consists of two sub algorithm. The first, `Explore` is an exploration phase, identical to our algorithm `ProbDep-Explore`. However, as opposed to simply outputting the end node the history of the random walk is passed to the second algorithm, Choose. The algorithm Choose selects all arms whose sample mean is within a certain tolerance of the threshold - chosen to be as small as possible while still producing a none empty set, and then takes the median of said set. This additional step is required as the MTB algorithm aims to achieve the minimax rate on *expected regret* - that is $\sqrt{\frac{\log(K)}{T}}$, and therefore wishes to output any arm $k : |\mu_k - \tau| \lesssim \sqrt{\frac{\log(K)}{T}}$. The idea being that during the explore phase enough time will be spent on nodes containing such arms.

If we consider the problem dependent regime, and whenever we are not in the trivial regime where $\bar{\Delta}_{\min} \lesssim \sqrt{\frac{\log(K)}{T}}$, we conjecture that the MTB algorithm will spend sufficient time on the unique node $\tilde{v} : \mu_{\tilde{v}(l)} < \tau < \mu_{\tilde{v}(r)}$ with high probability matching the bound of Theorem 28. The algorithm Choose will then output arm $\tilde{v}(r)$. The problem dependent regime allows for a less convoluted approach - indeed ProbDep-Explore is very simple in comparison to MTB. However, it is nevertheless important to note that for the monotone setting there exists an algorithm that is optimal in both problem dependent and problem independent regimes.

In regards to the concave case it is not as immediate that the CTB algorithm by Cheshire, Ménard, and Carpentier [26] will also be optimal in the problem dependent concave setting. The CTB algorithm is significantly more complex than the MTB as it successively applies a noisy binary search on a log scale to find arms increasingly close to threshold at a geometric rate. We however conjecture that it will be the case the CTB is also optimal in the problem dependent regime.

3.6.2 Unimodal constraint

A natural additional shape constraint for the *TBP* is a Unimodal one. Indeed bandit problems with a unimodal constraint are already considered in the literature, for the problem of minimising the cumulative regret or identifying the best arm under unimodal constraints see Yu and Mannor [85], Combes and Proutiere [27], Paladino et al. [74] and Combes and Proutiere [28]. The *TBP* in particular with a unimodal constraint is studied in Cheshire, Ménard, and Carpentier [26] in the *problem independent* regime. With the above work already in hand it is natural to consider a unimodal shape constraint on the *TBP* in the *problem dependent* regime. A possible algorithm would be one which, similar to the ProbDep-CTB, first finds an arm above threshold and then reduces the problem to one with a monotone constraint. We conjecture that if one considers a class of problems with M arms above threshold the regret of the problem will be dominated by that of finding a single arm above threshold and will be of the order $\exp\left(-\frac{MT\bar{\Delta}_{\min}}{K}\right)$ with a matching lower bound. If one wishes to consider a narrower class based on a single vector of gaps, as in the concave or monotone setting one might hope to achieve a rate $\exp\left(-\frac{MT}{K}(\frac{1}{M}\sum_{i=1}^{M}\bar{\Delta}_i)^2\right)$ however this result, for both an upper and lower bound, appears not so straightforward.

3.7 Proofs relating to the Monotone setting

We first state a useful inequality. Let $\mathrm{kl}(p, q)$ be the Kullback-Leibler divergence between two Bernoulli distributions of parameter p and q,

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

It holds

$$\begin{aligned}\mathrm{kl}(p,q) &= p\log\left(\frac{1}{q}\right) + (1-p)\log\left(\frac{1}{1-q}\right) + p\log(p) + (1-p)\log(1-p)\\ &\geq p\log\left(\frac{1}{q}\right) - \log(2)\,. \end{aligned}\tag{3.2}$$

Proof of Theorem 27. We denote by N_k^t the number of times the arm k is pulled until and included time t, i.e. $N_k^t = \sum_{s=1}^t \mathbb{1}_{k_s=k}$. Let $i = \arg\min_{k\in[K]} \bar{\Delta}_k$, that is $\bar{\Delta}_i = \bar{\Delta}_{\min}$. Consider the two bandit problems $\underline{\nu}^+$ and $\underline{\nu}^-$ where

$$\nu_k^+ = \begin{cases}\mathcal{N}(\bar{\Delta}_k,\sigma^2) & \text{if } k\geq i\\ \mathcal{N}(-\bar{\Delta}_k,\sigma^2) & \text{else}\end{cases}, \qquad \nu_k^- = \begin{cases}\mathcal{N}(\bar{\Delta}_k,\sigma^2) & \text{if } k> i\\ \mathcal{N}(-\bar{\Delta}_k,\sigma^2) & \text{else}\end{cases}.$$

Note these bandit problems belong to the class of *MTBP* $\mathcal{B}_m^{\bar{\Delta}}$. In particular we can lower bound the error by the probability to make a mistake in the prediction of the label of arm i

$$e_T^{\underline{\nu}^+} \geq \mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = -1) \qquad e_T^{\underline{\nu}^-} \geq \mathbb{P}_{\underline{\nu}^-}(\widehat{Q}_i = 1)\,.$$

We can assume that $\mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = -1) \leq 1/2$ otherwise the bound is trivially true. Thanks to the chain rule then the contraction of the Kullback-Leibler divergence (e.g. see Garivier, Ménard, and Stoltz [40]) and (3.2), it holds

$$\begin{aligned}T\frac{\bar{\Delta}_{\min}^2}{2\sigma^2} \geq \mathbb{E}_{\underline{\nu}^+}[N_i^T]\frac{\bar{\Delta}_{\min}^2}{2\sigma^2} &= \mathrm{KL}(\mathbb{P}_{\underline{\nu}^+}^{I_T}, \mathbb{P}_{\underline{\nu}^-}^{I_T})\\ &\geq \mathrm{kl}\left(\mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = 1), \mathbb{P}_{\underline{\nu}^-}(\widehat{Q}_i = 1)\right)\\ &\geq \mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = 1)\log\left(\frac{1}{\mathbb{P}_{\underline{\nu}^-}(\widehat{Q}_i = 1)}\right) - \log(2)\,,\end{aligned}$$

where we denote by $\mathbb{P}_{\underline{\nu}}^{I_T}$ the probability distribution of the history I_T under the bandit problem $\underline{\nu}$. Thus, using that $\mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = 1) = 1 - \mathbb{P}_{\underline{\nu}^+}(\widehat{Q}_i = -1) \geq 1/2$ we obtain

$$\mathbb{P}_{\underline{\nu}^-}(\widehat{Q}_i = 1) \geq \frac{1}{4}\exp\left(-\frac{T\bar{\Delta}_{\min}^2}{\sigma^2}\right).$$

Which allows us to conclude that

$$\max(e_T^{\underline{\nu}^+}, e_T^{\underline{\nu}^-}) \geq \frac{1}{4}\exp\left(-\frac{T\bar{\Delta}_{\min}^2}{\sigma^2}\right).$$

□

Proof of Theorem 28. We assume in the proof, without loss of generality, that

$$\Delta_{\min} \geq c_{\min}\sqrt{\frac{\sigma^2\log(K)}{T}}$$

with $c_{\min} = 13$. Indeed, otherwise, the bound of Theorem 28 is trivially true.

The proof of Theorem 28 is structured in the following manner. In our *original* binary tree we know there is a unique leaf v_Δ, such that $\tau \in [\mu_{v_\Delta(l)}, \mu_{v_\Delta(r)}]$. Essentially we want to show that the explore algorithm will terminate in the subtree of this $v_{\bar{\Delta}}$ with high probability - recall that we extend our binary tree by attaching an infinite sub tree to each leaf, the nodes of which are identical to the respective leaf. At time t we say our algorithm makes a favourable decision if all sample means are well concentrated - that is with $\Delta_{\min}$ of their true mean. On such a favourable decision we show that the explore algorithm will make a step towards the subtree of v_Δ, or go deeper if it is already in it. Therefore if overall we can make sufficient proportion of favourable events we are guaranteed to terminate in the subtree of v_Δ. We then show that this favorable event holds with high probability.

Step 1: Initial definitions and lemmas We denote by $ST(v)$ the subtree rooted at node v.

Definition 4. The subtree $ST(v)$ of a node v is defined recursively as follows: $v \in ST(v)$ and

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

We define $Z_{\Delta_{\min}}$, the set of good nodes, as

$$Z_{\Delta_{\min}} := \{v : \exists k \in \{l, m, r\} : |\mu_{v(k)} - \tau| \leq \Delta_{\min}\}\;,$$

Note that $Z_{\Delta_{\min}}$ is simply the leaf v_Δ and it's sub tree attached during the infinite extension of the binary tree. At time t we define w_t as the node of maximum depth whose subtree contains both v_t and $Z_{\Delta_{\min}}$. Formally, for $t \leq T_1$, we let

$$w_t \in \operatorname*{arg\,max}_{\{v:\tau\in[\mu_{v(l)},\mu_{v(r)}]\ \&\ v_t\in ST(v)\}} |v|\;. \tag{3.3}$$

Lemma 11. *The node w_t is unique.*

Proof. At time t consider, a node q_t which also satisfies (3.3). As $v_t \in ST(w_t)$ and $v_t \in ST(q_t)$ we can assume without loss of generality $q_t \in ST(w_t)$ with $|q_t| \geq |w_t|$. This then implies, from (3.3), that $|q_t| = |w_t|$ and as $q_t \in ST(w_t)$, we have $q_t = w_t$. □

For $t \leq T_1$ we define D_t as the relative distance from v_t to v_Δ, it is taken as the length of the path running from v_t up to w_t and then down (or up if $v_t \in Z_{\Delta_{\min}}$) to v_Δ. Formally, we have

$$D_t := |v_t| - |w_t| + |v_\Delta| - |w_t|.$$

Note the following properties of D_t and w_t,

$$ST(v_t) \cap Z_{\Delta_{\min}} \neq \emptyset \Rightarrow v_t = w_t\;, \tag{3.4}$$

$$D_t \leq 0 \Rightarrow v_t = w_t \text{ and } w_t, v_t \in Z_{\Delta_{\min}}\;. \tag{3.5}$$

We define the favorable event where the estimates of the means are close to the true ones for all the arms in v_t, At time t we define the event

$$\xi_t := \{\forall k \in \{l, m, r\}, |\widehat{\mu}_{k,t} - \mu_{v_t(k)}| \leq \Delta_{\min}\}$$

and we denote $\bar{\xi}_t$ as the complement of ξ_t.

Step 2: Actions of the algorithm on all iterations After any execution of algorithm ProbDep-Explore note the following, for $t \leq T_1$, v_t and v_{t+1} are separated by at most one edge, i.e.

$$v_{t+1} \in \{L(v_t), R(v_t), P(v_t)\}\;. \tag{3.6}$$

Lemma 12. *On execution of algorithm* ProbDep-Explore *for all* $t \le T_1$ *we have the following,*

$$D_{t+1} \le D_t + 1$$

Proof. As the algorithm moves at most 1 step per iteration, see (3.6), for $t \le T_1$, it holds

$$|v_t| - |w_t| \ge |v_{t+1}| - |w_t| - 1 \ .$$

We consider two cases. Firstly, assume we are in the event $\{v_{t+1} \neq P(v_t)\} \cup \{w_t \neq v_t\}$. Under this event note that $v_{t+1} \in ST(w_t)$. It follows

$$\begin{aligned} D_t &= |v_t| - |w_t| + |v_\Delta| - |w_t| \\ &\ge |v_{t+1}| - |w_t| + |v_\Delta| - |w_t| - 1 \\ &\ge |v_{t+1}| - |w_{t+1}| + |v_\Delta| - |w_{t+1}| - 1 \\ &= D_{t+1} - 1 \ , \end{aligned}$$

where the third line comes from the definition of w_{t+1}, see (3.3).
In the case where $w_t = v_t$ and $v_{t+1} = P(v_t)$ note that $w_{t+1} = v_{t+1}$ and,

$$D_{t+1} = |v_\Delta| - |w_{t+1}| = |v_\Delta| - |w_t| + 1 = D_t + 1 \ .$$

Therefore in all cases we have $D_{t+1} \le D_t + 1$. □

Step 3: Actions of the algorithm on ξ_t

Lemma 13. *On execution of algorithm* ProbDep-Explore *for all* $t \le T_1$*, on* ξ_t*, we have the following,*

$$D_{t+1} \le D_t - 1 \ .$$

Proof. Note that on the favorable event ξ_t, we have $\forall j \in \{l, m, r\}$,

$$\mu_{v_t(j)} \ge \tau \Rightarrow \hat{\mu}_{j,t} \ge \tau \ , \tag{3.7}$$

$$\mu_{v_t(j)} \le \tau \Rightarrow \hat{\mu}_{j,t} \le \tau \ . \tag{3.8}$$

We consider the following three cases:

- If $\tau \notin \left[\mu_{v_t(l)}, \mu_{v_t(r)}\right]$. From (3.7) and (3.8), under ξ_t, we get $\tau \notin [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$, and therefore $v_{t+1} = P(v_t)$ and $w_t = w_{t+1}$. Thus thanks to Lemma 11, under ξ_t,

  $$D_{t+1} = |v_{t+1}| - |w_{t+1}| + |v_\Delta| - |w_{t+1}| = |v_t| - 1 - |w_t| + |v_\Delta| - |w_t| = D_t - 1 \, .$$

- If $\tau \in \left[\mu_{v_t(l)}, \mu_{v_t(r)}\right]$ and $v_t \notin Z_{\Delta_{\min}}$. 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 v_Δ, id est $w_t = v_t$. Since $v_t \notin Z_{\Delta_{\min}}$, without loss of generality, we can assume for example $\mu_{v_t(m)} > \tau$. From (3.7) and (3.8), under ξ, we then have $\tau \in [\hat{\mu}_{l,t}, \hat{\mu}_{r,t}]$ and $\hat{\mu}_{m,t} \ge \tau$. Hence algorithm ProbDep-Explore goes to the correct subtree, $v_{t+1} = L(v_t)$. In particular we also have for this node

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

  therefore it holds again $w_{t+1} = v_{t+1}$. Thus combining the previous remarks we obtain thanks to Lemma 11, under ξ_t,

  $$D_{t+1} = |v_\Delta| - |w_{t+1}| = |v_\Delta| - |w_t| - 1 = D_t - 1 \ .$$

- If $\tau \in \big[\mu_{v_t(l)}, \mu_{v_t(r)}\big]$ and $v_t \in Z_{\Delta_{\min}}$. Firstly note that $w_t = v_t$. Now, using the same reasoning as in the previous case, as $\tau \in [\mu_{v_t(l)}, \mu_{v_t(r)}]$ we have $v_{t+1} = L(v_t)$ or $v_{t+1} = R(v_t)$. In either case we get $v_{t+1} \in Z_{\Delta_{\min}}$ because of (3.7) and (3.8), thus it holds $w_{t+1} = v_{t+1}$. Therefore we have

$$D_{t+1} = |v_{t+1}| - |w_{t+1}| + |v_\Delta| - |w_{t+1}| = |v_\Delta| - |w_t| - 1 = D_t - 1\,.$$

□

Step 4: Upper bound on D_{T_1+1}

Lemma 14. *For any execution of algorithm* ProbDep-Explore

$$D_{T_1+1} \le 2\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t} - \frac{3T_1}{4}\,.$$

Proof. Combining Lemma 12 and Lemma 13 respectively we have

$$D_{t+1} \le D_t + \mathbb{1}_{\bar{\xi}_t} - \mathbb{1}_{\xi_t}\,.$$

Using this inequality we obtain

$$\begin{aligned} D_{T_1+1} &= D_1 + \sum_{t=1}^{T_1}(D_{t+1} - D_t) \\ &\le D_1 + \sum_{t=1}^{T_1}\left(\mathbb{1}_{\bar{\xi}_t} - \mathbb{1}_{\xi_t}\right) \\ &\le D_1 + 2\sum_{t=1}^{T_1}\mathbb{1}_{\bar{\xi}_t} - T_1 \\ &\le 2\sum_{t=1}^{T_1}\mathbb{1}_{\bar{\xi}_t} - \frac{3T_1}{4}\,, \end{aligned}$$

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

Lemma 15. *For $c_{\text{mon}} = 1/48$ and $c'_{\text{mon}} = 12$ it holds*

$$\mathbb{P}\left(\sum_{t=1}^{T_1}\mathbb{1}_{\bar{\xi}_t} \ge \frac{T_1}{4}\right) \le \exp\left(c_{\text{mon}}\frac{-T\Delta_{\min}^2}{\sigma^2}\right).$$

Proof. Let $\mathcal{F}_t$ be the information available at and including step t of algorithm ProbDep-Explore. Thanks to the Chernoff inequality and the choice of T_2, we have for all $j \in \{l, m, r\}$,

$$\begin{aligned} \mathbb{P}\big(\left|\hat{\mu}_{j,t} - \mu_{v_t(j)}\right| \ge \Delta_{\min}|\mathcal{F}_{t-1}\big) &\le 2\exp\left(-\frac{T_2\Delta_{\min}^2}{2\sigma^2}\right) \\ &\le 2\exp\left(-c_{\min}^2\frac{\log(K)}{36\log(K)+6}\right) \\ &\le \frac{1}{24} \end{aligned}$$

as we assume $\Delta_{\min} > c_{\min}\sqrt{\frac{\sigma^2 \log K}{T}}$ and $c_{\min} \geq 13$. Therefore by a union bound

$$p_t := \mathbb{P}(\bar{\xi}_t | \mathcal{F}_{t-1}) \leq p_0 := 6 \exp\!\left(-\frac{T_2 \Delta_{\min}^2}{2\sigma^2}\right) \leq \frac{1}{8}. \tag{3.9}$$

We will apply the Chernoff inequality to upper bound the sum of indicator function. Thanks to the Markov inequality for $\lambda \geq 0$ we have

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t} \geq \frac{T_1}{4}\right) \leq \mathbb{E}\left[\exp\!\left(\lambda \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t}\right)\right] e^{-\lambda \frac{T_1}{4}}\,. \tag{3.10}$$

Let $\phi_p(\lambda) = \log(1 - p + pe^{\lambda})$ be the log-partition function of a Bernoulli of parameter $p \in [0,1]$. Note that for $\lambda \geq 0$, since $p \mapsto \phi_p(\lambda)$ is non-decreasing and because of (3.9) it holds $\phi_{p_t}(\lambda) \leq \phi_{p_0}(\lambda)$ for all t. Thus by induction we have

$$\begin{aligned}
\mathbb{E}\left[\exp\!\left(\lambda \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t}\right)\right] &= \mathbb{E}\left[\mathbb{E}\big[\exp\!\big(\lambda \mathbb{1}_{\bar{\xi}_t}\big) | \mathcal{F}_{T_1-1}\big] \exp\!\left(\lambda \sum_{t=1}^{T_1-1} \mathbb{1}_{\bar{\xi}_s}\right)\right] \\
&= \mathbb{E}\left[e^{\phi_{p_{T_1}}(\lambda)} \exp\!\left(\lambda \sum_{t=1}^{T_1-1} \mathbb{1}_{\bar{\xi}_t}\right)\right] \leq e^{\phi_{p_0}(\lambda)} \mathbb{E}\left[\exp\!\left(\lambda \sum_{s=1}^{t-1} \mathbb{1}_{\bar{\xi}_t}\right)\right] \\
&\leq \mathbb{E}\left[e^{T_1 \phi_{p_0}(\lambda)}\right].
\end{aligned}$$

Then going back to (3.10) and using that $\sup_{\lambda \geq 0} \lambda q - \phi_p(\lambda) = \mathrm{kl}(q,p)$ when $q \geq p$ we get

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t} \geq \frac{T_1}{4}\right) \leq \exp\!\left(-T_1 \sup_{\lambda \geq 0}\left(\lambda \frac{1}{4} - \phi_{p_0}(\lambda)\right)\right) = e^{-T_1 \mathrm{kl}(1/4, p_0)}\,.$$

It remains to conclude with (3.2)

$$\begin{aligned}
T_1 \mathrm{kl}(1/4, p_0) &\geq T_1 \frac{1}{4} \log(1/p_0) - T_1 \log(2) \\
&\geq \frac{1}{8} \frac{T_1 T_2 \bar{\Delta}_{\min}^2}{\sigma^2} - T_1\left(\log(2) + \frac{1}{4}\log(3)\right) \\
&\geq \frac{1}{48} \frac{T \bar{\Delta}_{\min}^2}{\sigma^2} - T_1\left(\log(2) + \frac{1}{4}\log(3)\right) \\
&\geq c_{\mathrm{mon}} \frac{T \bar{\Delta}_{\min}^2}{\sigma^2} - c'_{\mathrm{mon}} \log(K)
\end{aligned}$$

where $c_{\mathrm{mon}} = 1/48$ and $c'_{\mathrm{mon}} = 12$.

□

By combination of Lemmas 14 and 15 we have that $D_{T_1+1} \leq 0$ with probability greater than $1 - \exp\!\left(-c_{\mathrm{mon}} \frac{T\Delta_{\min}^2}{\sigma^2} + c'_{\mathrm{mon}} \log(K)\right)$. Thus with said probability we output an arm $\hat{a}$ such that $\tau \in [\mu_{\hat{a}}, \mu_{\hat{a}+1}]$.

3.8 Proofs relating to the concave setting

Before proceeding with the proof of Theorem 29 we present the following structural lemma.

Lemma 16. *Let $\Delta \in \Delta\mathcal{B}_c$ and let $(\mu_k)_k$ be an associated concave sequence of means. There exists a sequence of means $(\mu'_k)_k$ - with associated gaps $\Delta' = |\mu' - \tau|$ - such that*

(a) $(\mu'_k)_k$ is concave.

(b) μ and μ' have not all the arms classified in the same way:

$$\exists k \in [K] : \mathrm{sign}(\mu_k - \tau) \neq \mathrm{sign}(\mu_k - \tau)\,.$$

(c) For all $k \in [K]$ it holds that

$$|\mu'_k - \mu_k| \leq 3\Delta_{\min}.$$

(d) For all $k \in [K]$ it holds that

$$\frac{\Delta_k}{10} \leq \Delta'_k \leq 3\Delta_k.$$

Proof. Let $k^* \in \arg\min_{k\in[K]} \Delta_k$. We proceed in two cases: either this arm is up threshold, or it is below threshold. In everything that follows we set $\Delta_{\min} := \min_{k\in[K]} \Delta_k$.

Case 1: Arm below threshold, i.e. $\mu_{k^*} \leq \tau$. Let us write k_L, k_R for the two arms that are 'just' below threshold, i.e. such that $\mu_{k_L} \leq \tau \leq \mu_{k_L+1}$ and $\mu_{k_R} \leq \tau \leq \mu_{k_R-1}$. These two arms can be defined without loss of generality since there is at least one arm above threshold, and since we can always take two virtual means μ_0, μ_{K+1} at $-\infty$ on the boundaries.

In the context where $\mu_{k^*} \leq \tau$ it is clear that we can pick $k^* \in \{k_L, k_R\}$ and so let us assume w.l.g. that $k^* = k_L$.

In this case, we define μ' either:

- if $\Delta_{k_R} \leq 3\Delta_{\min}/2$, for all $k \in [K]$,

$$\mu'_k = \mu_k + 2\Delta_{\min}.$$

- if $\Delta_{k_R} \geq 3\Delta_{\min}/2$, for all $k \in [K]$,

$$\mu'_k = \mu_k + 5\Delta_{\min}/4.$$

(a) holds as we just translated vertically the concave means. Also (b) holds since we switched the sign of arm k^* by construction. (c) holds also since we precisely added at most $2\Delta_{\min}$ to the means. And finally for (d): we have for any $k \in [K]$ that $|\bar{\Delta}_k - \Delta'_k| \leq 2\Delta_{\min}$, so that

$$\Delta'_k \leq 3\Delta_k.$$

Moreover for all arms k above threshold, it is clear that $\Delta'_k \geq \Delta_k$. On the other hand, for any arm k below threshold and that are not next to an arm up threshold - i.e. not k_L or k_R - we have by concavity that

$$\tau - \mu_k \geq 3\Delta_{\min},$$

which implies

$$\tau - \mu'_k \geq \tau - \mu_k - 2\Delta_{\min} \geq \frac{\tau - \mu_k}{3},$$

i.e. $\Delta'_k \geq \Delta_k/3$. Finally for $\{k_L, k_R\}$: it is clear that $\Delta'_{k^*} \geq \Delta_{k^*}/4$ by construction so that $\Delta'_{k_L} \geq \Delta_{k_L}/4$. And also by construction:

- if $\Delta_{k_R} \leq 3\Delta_{\min}/2$, then $\Delta'_{k_R} \geq \Delta_{\min}/2 \geq \Delta_{k_R}/3$.
- if $\Delta_{k_R} \geq 3\Delta_{\min}/2$, then $\Delta'_{k_R} \geq \Delta_{k_R} - 5\Delta_{\min}/4 \geq \Delta_{k_R}/6$.

So that in both situations (d) holds.

Case 2: Arm above threshold, i.e. $\mu_{k^*} \geq \tau$. Note first that if k^* is the only arm above threshold, we simply set for any k

$$\mu'_k = \mu_k - 2\Delta_{\min},$$

and this satisfies the requirements (a)-(d). Assume now that this case does not hold, so that $k^* \in \{k_L+1, k_R-1\}$ and $k_L+1 < k_R-1$.We now again consider several cases. Note that in any case $k^* \in \{k_L+1, k_R-1\}$.
Sub-case 1: μ not too flat around the threshold. Assume first that $\Delta_{k_L+2} \wedge \Delta_{k_R-2} \geq 3\Delta_{\min}/2$. Assume w.l.o.g. that $k^* = k_L+1$. In this sub-case we define μ' either as:

- if $\Delta_{k_R-1} \geq 5\Delta_{\min}/4$ set
$$\mu' = \mu - 9\Delta_{\min}/8,$$
- otherwise if $\Delta_{k_R-1} \leq 5\Delta_{\min}/4$ set
$$\mu' = \mu - 11\Delta_{\min}/8.$$

It is clear that (a) holds (vertical translation of a concave sequence), (b) holds (arm k^* changes sides of threshold) and (c) holds since we translate at most by $11\Delta_{\min}/8$. Now for (d): it is clear in both cases that $\Delta'_k \leq \Delta_k + 11\Delta_{\min}/8 \leq 3\Delta_k$. Moreover:

- if $\Delta_{k_R-1} \geq 5\Delta_{\min}/4$, then for all $k \neq k^*$, we have $\Delta'_k \geq \Delta_k - 9\Delta_{\min}/8 \geq \Delta_k/8$ - and also by definition $\Delta_{k^*} = \Delta_{k^*}/8$. And so (d) holds in this case.
- if $\Delta_{k_R-1} \leq 5\Delta_{\min}/4$ we have for all k such that $\mu_k \leq \tau$ that $\Delta'_k \geq \Delta_k$, and for any $k \in \{k_L+2, ..., k_R-2\}$ that $\Delta'_k \geq \Delta_k - 11\Delta_{\min}/8 \geq \Delta_k/8$ since for such k we have $\Delta_k \geq 3\Delta_{\min}/2$. Also $\Delta'_{k_L+1} \geq \Delta'_{k_R-1} \geq \Delta_{\min}/8 \geq \Delta_{k_R-1}/10 \geq \Delta_{k_L+1}/10$. And so (d) holds in this case.

Sub-case 2: μ quite flat around the threshold. Assume now that $\Delta_{k_L+2} \wedge \Delta_{k_R-2} \leq 3\Delta_{\min}/2$. Assume w.l.o.g. that $\Delta_{k_L+2} \leq 3\Delta_{\min}/2$ and set

$$\mu'_{k_L+1} = \mu_{k_L+1} - 9\Delta_{k_L+1}/8.$$

and for $k \neq k_L+1$

$$\mu'_k = \mu_k - \Delta_{k_L+1}/2.$$

(b) holds since $\mu'_{k_L+1} \leq \tau \leq \mu_{k_L+1}$. Since $\Delta_{k_L+1} \leq \Delta_{k_L+2} \leq 3\Delta_{\min}/2$, we know that (c) and (d) hold. Finally note that

$$\begin{aligned} \mu_{k_L+1} - \mu_{k_L} \geq \mu'_{k_L+1} - \mu'_{k_L} &= 3\Delta_{k_L+1}/8 + \Delta_k \\ \geq \mu'_{k_L+2} - \mu'_{k_L+1} + 5\Delta_{k_L+1}/8 = \mu'_{k_L+2} - \mu'_{k_L+1} &\geq \mu_{k_L+2} - \mu_{k_L+1}, \end{aligned}$$

since $\mu'_{k_L+2} - \mu'_{k_L+1} \leq \Delta_{\min}/2$ - since $\Delta_{k_L+1} \leq \Delta_{k_L+2} \leq 3\Delta_{\min}/2$ - so that $\Delta_k \geq \Delta_{\min} \geq \mu'_{k_L+2} - \mu'_{k_L+1} + \Delta_{k_L+1}/4$. So (a) holds since for any $k \notin \{k_L+1, k_L+2\}$, we have $\mu_k - \mu_{k-1} = \mu'_k - \mu'_{k-1}$.

□

Proof of Theorem 29. Consider $\bar{\Delta} \in \Delta\mathcal{B}_c$ associated with the vector of means $(\mu_k)_{k\in[K]}$. We define $\underline{\nu}$ as the Gaussian bandit problem with these means, that is, $\nu_k = \mathcal{N}(\mu_k, \sigma^2)$ for all $k \in [K]$. Thanks to Lemma 16 there exists a vector of means $(\mu'_k)_{k\in[K]}$ that verifies the conditions of Lemma 16. We denote by $\underline{\nu}'$ the Gaussian bandit problem such that $\nu'_k = \mathcal{N}(\mu_k, \sigma^2)$ for all $k \in [K]$. Thanks to (a) and (d) we know that $\underline{\nu}' \in \mathcal{B}_c$. Thanks to (b) there exists $i \in [K]$ such that, for example, $\mu_i > \tau$ and $\mu'_a < \tau$. In particular we can lower bound the error by the probability to make a mistake in the prediction of the label of arm i

$$e_T^{\underline{\nu}} \geq \mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = -1) \qquad e_T^{\underline{\nu}'} \geq \mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = 1)\,.$$

We then conclude as in the proof of Theorem 27. We can assume that $\mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = -1) \leq 1/2$ otherwise the bound is trivially true. Thanks to (c), the chain rule, the contraction of the Kullback-Leibler divergence and (3.2), it holds

$$\begin{aligned} T\frac{9\bar{\Delta}^2_{\min}}{2\sigma^2} &\geq \mathrm{KL}(\mathbb{P}^{I_T}_{\underline{\nu}}, \mathbb{P}^{I_T}_{\underline{\nu}'}) \\ &\geq \mathrm{kl}\left(\mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = 1), \mathbb{P}_{\underline{\nu}'}(\widehat{Q}_i = 1)\right) \\ &\geq \mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = 1)\log\!\left(\frac{1}{\mathbb{P}_{\underline{\nu}'}(\widehat{Q}_i = 1)}\right) - \log(2)\,, \end{aligned}$$

where we denote by $\mathbb{P}^{I_T}_{\underline{\nu}}$ the probability distribution of the history I_T under the bandit problem $\underline{\nu}$. Thus, using that $\mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = 1) = 1 - \mathbb{P}_{\underline{\nu}}(\widehat{Q}_i = -1) \geq 1/2$ we obtain

$$\mathbb{P}_{\underline{\nu}'}(\widehat{Q}_i = 1) \geq \frac{1}{4}\exp\!\left(-9\frac{T\bar{\Delta}^2_{\min}}{\sigma^2}\right).$$

Which allows us to conclude that

$$\max(e_T^{\underline{\nu}^+}, e_T^{\underline{\nu}^-}) \geq \frac{1}{4}\exp\!\left(-9\frac{T\bar{\Delta}^2_{\min}}{\sigma^2}\right).$$

□

Proof of Theorem 30. We assume in the proof, without loss of generality, that

$$\Delta_{\min} \geq c_{\mathrm{con-min}}\sqrt{\frac{\sigma^2\log(K)}{T}}$$

with $c_{\mathrm{con-min}} = 8064$. Indeed, otherwise, the bound of Theorem 30 is trivially true.

The proof of Theorem 30 is structured in the following manner. In our *original* binary tree we assume there is at least one arm above threshold, the contrary case is dealt with separately, see Lemma 23. We wish to show that with high probability the `Grad-Explore` algorithm will add sufficient arms above threshold to the list S_{T_1} such that when we take it's median we are guaranteed to output an arm above threshold. At time t we say our algorithm makes a favourable decision if all sample means are well concentrated - that it with $\bar{\Delta}_{\min}$ of their true mean. It is important to note that for arms below threshold this also implies the estimated gradients are close to their true values. On such a favourable decision we show that the explore algorithm will make a step towards the subtree of nodes containing an arm above threshold, or remain inside if it is already in it. We also show that upon encountering an arm above threshold, on a good decision said arm is always added to S_{T_1}. Therefore if overall we can make sufficient proportion of favourable events we are guaranteed to have a sufficient number of arms above threshold in S_{T_1}. We then show that this favorable event holds with high probability. Once we have identified an arm above threshold the problem is essentially split into two monotone problems - see Remark 14, where the point the arms cross threshold on either side can be found by applying the `PD-DEC-MTB` and `ProbDep-Explore` algorithms in opposite directions.

Step 1: Initial definitions and lemmas We thus assume first that there is an arm $k*$ such that $\mu_{k^*} > \tau$.

Definition 5. 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 6. A consecutive tree U with root $u_{\texttt{root}}$ is a set of nodes such that $u_{\texttt{root}} \in U$ and

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

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, the set of good nodes with at least an arm with a mean above the threshold,

$$Z := \{v : \exists j \in \{l, m, r\} : \mu_{v(j)} > \tau\}\ .$$

At a given time t note the following property of Z and v_t,

$$ST(v_t) \cap Z \neq \emptyset \Leftrightarrow k^* \in [v_t(l), v_t(r)]\,. \tag{3.11}$$

Proposition 20. *Z is a consecutive tree with root z_{root} the unique element $v \in Z$ such that $P(v) \notin Z$ if there exists at least one, otherwise $z_{\text{root}} = \texttt{root}$.*

Proof. First, if for all $v \in Z$ we have $P(v) \in Z$ then $\texttt{root} \in Z$ and Z is a consecutive tree with root $z_{\text{root}} = \texttt{root}$. Otherwise, consider $v \in Z$, such that $P(v) \notin Z$, there is at least one such node. We first prove that v is unique. As $v \in Z$ we know that

$$\exists j \in \{l, m, r\} : \mu_{v(j)} > \tau\ . \tag{3.12}$$

Now since $v(l), v(r) \in P(v)$ and $P(v) \notin Z$, it follows that, thanks to (3.12),

$$\forall k \in \{l, r\} : \mu_{v(k)} < \tau\ .$$

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 this then implies $\mu_{q(l)} > \tau$ a contradiction. Hence $v = q$, and thus v is unique which implies $\forall q \in Z :\ q \neq v,\, P(q) \in Z$. □

At time t we define w_t as the node of maximum depth whose sub tree contains both v_t and Z. Formally, for $t \leq T_1$,

$$w_t := \underset{\{ST(w) \cap Z \neq \emptyset \ \& \ v_t \in ST(w)\}}{\arg\max} |w| \,. \tag{3.13}$$

Lemma 17. *The node w_t is unique.*

Proof. At time t consider, a node q_t which also satisfies (3.13). As $v_t \in ST(w_t)$ and $v_t \in ST(q_t)$ we can assume without loss of generality $q_t \in ST(w_t)$ with $|q_t| \geq |w_t|$. This then implies, from (3.13), that $|q_t| = |w_t|$ and as $q_t \in ST(w_t)$, we have $q_t = w_t$. □

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

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

Note the following properties of D_t and w_t,

$$ST(v_t) \cap Z \neq \emptyset \Rightarrow v_t = w_t \ , D_t = 0 \Rightarrow v_t = w_t \text{ and } w_t, v_t \in Z \ .$$

Define at time t the counter G_t, tracking the number of good arms in S_t,

$$G_t := \left| \left\{ k \in S_t : \ \mu_k > \tau \right\} \right| . \tag{3.14}$$

At time t we define the following favorable event where the sampled arms at time t a well concentrated around their means,

$$\xi_t := \{\forall j \in \{l, l+1, m, m+1, r, r+1\}, \left|\widehat{\mu}_{j,t} - \mu_{v_t(j)}\right| \leq \Delta_{\min}\}.$$

Step 2: Actions of the algorithm on all iterations After any execution of algorithm Grad-Explore note the following,

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

$$v_{t+1} \in \{L(v_t), R(v_t), P(v_t)\} \,, \tag{3.15}$$

- for $t \leq T_1$,

$$|S_t| \leq |S_{t+1}| \leq |S_t| + 1 \,. \tag{3.16}$$

Lemma 18. *On execution of algorithm* **Grad-Explore** *for all $t \leq T_1$ we have the following,*

$$D_{t+1} \leq D_t + 1, \tag{3.17}$$
$$G_{t+1} \geq G_t \,. \tag{3.18}$$

Proof. As the algorithm moves at most 1 step per iteration, see (3.15), for $t \leq T_1$, it holds

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

Noting that,

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

where the third line comes from the definition of w_{t+1}, see (3.3), we obtain $D_{t+1} \leq D_t + 1$. By (3.16) we have, for $t \leq T_1$,

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

hence $G_{t+1} \geq G_t$. □

Step 3: Actions of the algorithm on ξ_t We first state several properties relating to the event ξ_t. Firstly for all t we have that under event ξ_t,

$$\forall k \in \{l, m, r\},\ \mathrm{sign}(\widehat{\mu}_{k,t} - \tau) = \mathrm{sign}(\mu_k - \tau)\ . \tag{3.19}$$

Since there is at least an arm above the threshold, due to the concave property, note the following,

$$\forall k \in [K]:\ \ \mu_k < \tau,\ |\mu_k - \mu_{k+1}| \geq 2\Delta_{\min}\ , \tag{3.20}$$

thus from (3.20) for all t under event ξ_t, we have that,

$$\forall j \in \{l, m, r\} : \mu_{v_t(j)} < \tau,\ \mathrm{sign}(\widehat{\nabla}_{j,t}) = \mathrm{sign}(\nabla_{v_t(j)})\,. \tag{3.21}$$

Lemma 19. *On execution of algorithm* ***Grad-Explore*** *for all $t \leq T_1$, on ξ_t, we have the following,*

$$D_{t+1} \leq \max(D_t - 1, 0)\ , \tag{3.22}$$

$$G_{t+1} \geq G_t + \mathbb{1}_{\{D_t = 0\}}\,. \tag{3.23}$$

Proof. We first prove (3.23). If $D_t = 0$ then we know $v_t \in Z$. If $v_t \in Z$ then under ξ_t there exists $j \in \{l, m, r\}$ such that $\widehat{\mu}_{j,t} > \tau$, see (3.19), and arm is added to S_{t+1}, thus $G_{t+1} \geq G_t + \mathbb{1}_{\{D_t=0\}}$.

We now prove (3.22). We consider the following three cases:

- If $Z \cap ST(v_t) = \emptyset$. First of all we have that $\forall j \in \{l, m, r\} : \mu_{v_t(j)} \leq \tau$. Therefore from (3.19) the algorithm will not add an arm to S_t. Now, we have that $k^* \notin [v_t(l), v_t(r)]$, see (3.11), therefore via the concave property $\nabla_{v_t(l)} < 0$ or $\nabla_{v_t(r)} > 0$. Via (3.21) this implies that $\widehat{\nabla}_{v_t(l)} < 0$ or $\widehat{\nabla}_{v_t(r)} > 0$ respectively. Thus by action of the algorithm $v_{t+1} = P(v_t)$. Since in this case we are getting closer to the set of good nodes by going up in the tree we know that $w_t = w_{t+1}$. Thus thanks to Lemma 2, under ξ_t,

$$D_{t+1} = |v_{t+1}| - |w_{t+1}| + (|z_{\texttt{root}}| - |w_{t+1}|)^+ = |v_t| - 1 - |w_t| + (|z_{\texttt{root}}| - |w_t|)^+ = D_t - 1\,.$$

- If $k^* \in ST(v_t)$ and $v_t \notin Z$. First of all we have that $\forall j \in \{l, m, r\} : \mu_{v_t(j)} \leq \tau$. Therefore from (3.19) the algorithm will not add an arm to S_t. Now 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 good node, id est $w_t = v_t$. Since $v_t \notin Z$, without loss of generality, we can assume for example $\widehat{\nabla}_{t,m} > 0$. From (3.21), under ξ_t, we then have that $\nabla_{v_t(m)} > 0$ which implies $k^* \in [v_t(l), v_t(m)]$. Hence algorithm Grad-Explore goes to the correct subtree, $v_{t+1} = L(v_t)$. In particular we also have for this node

$$k^* \in [v_{t+1}(l),\, v_t(m)]\,,$$

therefore it holds again $w_{t+1} = v_{t+1}$. Thus combining the previous remarks we obtain thanks to Lemma 2, under ξ_t,

$$D_{t+1} = \left(|w_{t+1}| - |z_{\texttt{root}}|\right)^+ = \left(|w_t| - |z_{\texttt{root}}|\right)^+ - 1 = D_t - 1\,.$$

- If $k^* \in ST(v_t)$ and $v_t \in Z$. In this case there exists $j \in \{l, m, r\}$ such that $\mu_{v_t(j)} > \tau$. From 3.19 we have for said j that, $\widehat{\mu}_{j,t} > \tau$. Hence the algorithm will not move giving $v_t = v_{t+1}$ thus $D_t = D_{t+1} = 0$.

□

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

Lemma 20. *For any execution of algorithm Grad-Explore,*

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

Proof. Combining (3.22) and (3.17) from Lemma 18 and Lemma 19 respectively we have

$$\begin{aligned} D_{t+1} &\leq D_t + \mathbb{1}_{\bar{\xi}_t} - \mathbb{1}_{\xi_t}\mathbb{1}_{\{D_t>0\}} \\ &= D_t + 2\mathbb{1}_{\bar{\xi}_t} - 1 + \mathbb{1}_{\xi_t}\mathbb{1}_{\{D_t=0\}}\,. \end{aligned}$$

Using this inequality with (3.23) we obtain

$$\begin{aligned} G_{T_1+1} &= \sum_{t=1}^{T_1} G_{t+1} - G_t \\ &\geq \sum_{t=1}^{T_1} \mathbb{1}_{\xi_t}\mathbb{1}_{\{D_t=0\}} \\ &\geq \sum_{t=1}^{T_1} \left(D_{t+1} - D_t - 2\mathbb{1}_{\bar{\xi}_t} + 1\right) \\ &\geq T_1 - D_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_{t,}} \\ &\geq \frac{3}{4}T_1 - 2\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_{t,}}\,, \end{aligned}$$

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 . □

Lemma 21. *Upon execution of algorithm* Grad-Explore *with budget* $\frac{T}{3}$ *we have that,*

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t} \leq \frac{T_1}{8}\right) \leq \exp\left(-c_{\text{con}} \frac{T\bar{\Delta}^2_{\min}}{\sigma^2} + c'_{\text{con}} \log(K)\right).$$

where $c_{\text{con}} = \frac{1}{576}$ *and* $c'_{\text{con}} = 12$.

Proof. The proof follows as in the proof of Lemma 15, with altered constants. □

Lemma 22. *Under the assumption* $\exists k : \mu_k > \tau$*, upon execution of algorithm* Grad-Explore *with output* $\hat{a}$ *we have that* $\mu_{\hat{a}} \geq \tau$ *with probability greater than*

$$1 - \exp\left(-c_{\text{con}} \frac{T\bar{\Delta}^2_{\min}}{\sigma^2} + c'_{\text{con}} \log(K)\right).$$

Proof. By combination of Lemmas 5 and 21 we have that $G_{T_1+1} \geq \frac{1}{2}T_1$ with probability greater than $1 - \exp\left(-c_{\text{con}} \frac{T\bar{\Delta}^2_{\min}}{\sigma^2} c'_{\text{con}} \log(K)\right)$. As $|S_{T_1+1}| \leq T_1$ and as the arms G_{T_1+1} form a segment (they are all above threshold) by taking the median of S_{T_1+1} under the circumstance $G_{T_1+1} \geq \frac{1}{2}T_1$ we have that the output of Grad-Explore $\hat{a}$ is such that $\mu_{\hat{a}} > \tau$. This then gives the result. □

With the following lemma we deal with the special case where all arms are below threshold before finally completing the proof of Theorem 30.

Lemma 23. *Under the assumption* $\forall k \in [K] : \mu_k < \tau$*, upon execution of algorithm* Grad-Explore *with output* $\hat{a}$ *we have that* $\forall k \in [K], \widehat{Q}_k = -1$ *with probability greater than*

$$1 - \exp\left(-c_{\text{con}} \frac{T\bar{\Delta}^2_{\min}}{\sigma^2} + c'_{\text{con}} \log(K)\right).$$

where $c_{\text{con}} = \frac{1}{576}$ *and* $c'_{\text{con}} = 12$.

Proof. Under the assumption $\forall k \in [K] : \mu_k < \tau$, for all $t < T_1$, we have that under the event ξ_t, $S_{t+1} = S_t$, see (3.19). Therefore the following holds,

$$|S_{T_1}| \leq \sum_{t=1}^{T_1} \mathbb{1}_{\bar{\xi}_t}.$$

The proof now follows from direct application of Lemma 21. □

We are now ready to prove Theorem 30.

Proof of Theorem 30. In the case where $\mu_k < \tau$, $\forall k \in [K]$ Lemma 23 immediately gives the result. Therefore we consider the case in which $\exists k \in [K] : \mu_k > \tau$. Under this assumption the algorithm Grad-Explore will return an arm $\hat{a} : \mu_{\hat{a}} > \tau$ with probability greater than

$$1 - \exp\left(-\frac{1}{576} \frac{T\bar{\Delta}^2_{\min}}{\sigma^2} + 12 \log(K)\right),$$

see Lemma 22. In this case we have the sets of arms $[1, \hat{a}]$, $[\hat{a}, K]$ which satisfy the assumption described in Remark 14. Therefore via Theorem 28 and a union bound we have that with probability greater than

$$1 - 2\exp\left(-\frac{1}{48}\frac{T\Delta^2}{\sigma^2} + 12\log(K)\right)$$

we will correctly classify arms on both these sets. With an additional union bound we achieve the result. □

3.9 Experiments

We conduct some preliminary experiments to test the performance of both ProbDep-Explore and ProbDep-CTB to illustrate our theoretical understanding. As a bench mark we will use both a Uniform algorithm and also a naive binary search - that is without back tracking, that we will term Naive, for an exact description of both see Appendix. Note that Naive essentially behaves as a uniform sampling algorithm on a bandit problem with $\log(K)$ arms. As our theoretical bounds are likely far to loose in terms of constants we also include a parameter tuned version of the ProbDep-Explore where we tune the constants in the definition of T_1 and T_2, see Equation (3.1).

We would expect the Naive algorithm to have an upper bound of the order $\exp\left(-\frac{T\bar{\Delta}^2_{\min}}{\log(K)}\right)$. This is sub-optimal compared to ProbDep-Explore which removes the $\log(K)$, see Theorem 28. However, ProbDep-Explore must divide it's budget across several arms at each round, while Naive algorithm samples only one. This may out weigh the benefit of backtracking when K is not very large.

In our experiments we consider two thresholding bandit problems. In Setting 1 the gap of one arm is set to Δ, with the remaining gaps very large - i.e. 100, In Setting 2 all gaps are set to Δ, for the CTB we modify this to a concave setting where all arms are Delta apart. The former problem should more favour ProbDep-Explore as it can quickly traverse the binary tree and expend most of it's budget on the leaf in question.

In Figure 3.2 we consider consider the expected error in Setting 1 as a function of the gap Δ and as a function of the number of arms K.The effect of varying Δ follows our intuition. Firstly all algorithms show an increased performance for greater Δ, this should be completely expected. Secondly, in Setting 1 the ProbDep-Explore algorithm decrease in probability of error faster than Naive and much faster than Uniform. This is also unsurprising as in this setting the Uniform, and to a lesser extent Naive, algorithms are forced to waste an unnecessary amount of their budget on arms far from threshold. In the case of varied K, on the right, ProbDep-Explore appears to outperform Naive, showing no obvious dependency on K past a certain point, however there is considerable noise.

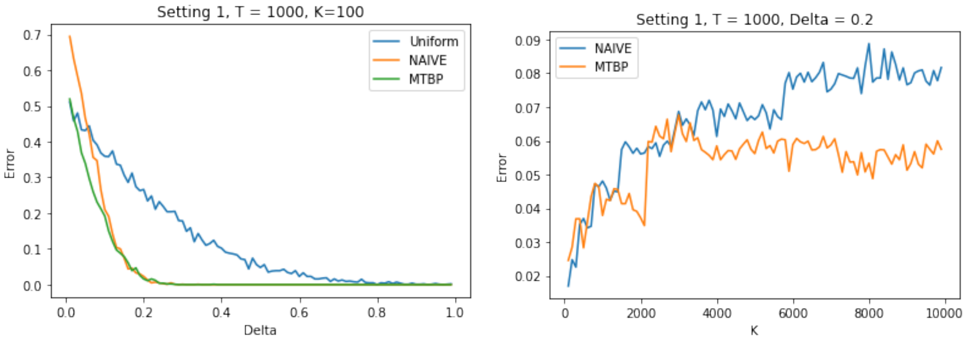

FIGURE 3.2: On the right: expected error as a function of the number of arms $K \in (100 \times i)_{i\in[100]}$ with $T = 1000$ and $\Delta = 0.2$ in Setting 1 averaged over 10000 Monte Carlo simulations. On the left: expected error as a function of the gap $\Delta \in (0.01 \times i)_{i\in[100]}$ with $T = 1000$ and $K = 100$ in Setting 1 averaged over 1000 Monte Carlo simulations.

In figure 3.3 we consider consider the expected error in Setting 2 as a function of the gap Δ and as a function of the number of arms K. In both cases Naive out performs both ProbDep-Explore and it's tuned version, vastly so for larger K. It would appear that here dividing our budget cancels out any gains one receives from reducing dependency on $\log(K)$. It is unfortunate that we were unable to find heuristic evidence of a lack of dependency on $\log K$, although this was perhaps expected. Based on our results, see Theorem 28, to remove such a dependency one would need $T\Delta^2 >> \log(K)$. This would lead to extremely low probabilities of error which are near impossible to detect accurately without huge numbers of Monte Carlo simulations, unfortunately beyond the scope of this paper.

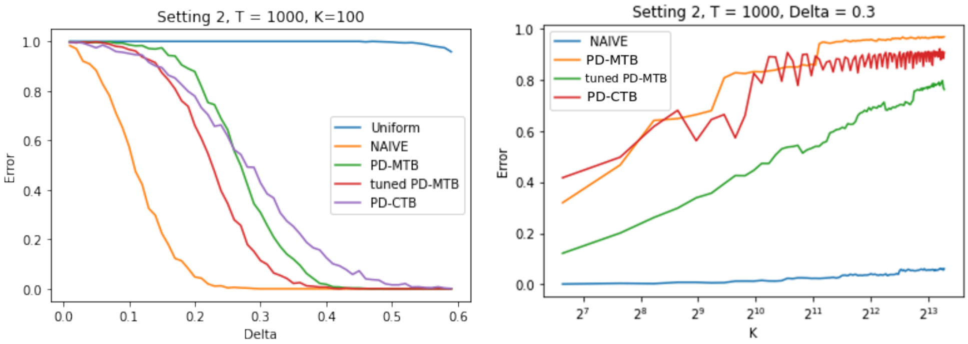

FIGURE 3.3: On the right: Expected error as a function of the number of arms $K \in (100 \times i)_{i\in[100]}$ with $T = 1000$, $\Delta = 0.3$, in Setting 2, plotted on a log scale averaged over 10000 Monte Carlo simulations. On the left: expected error as a function of the gap $\Delta \in (0.01 \times i)_{i\in[60]}$ with $K = 100$, $T = 1000$, in Setting 2, averaged over 1000 Monte Carlo simulations

```
Initialization: $v_1 = \texttt{root}$ for $t = 1 : T_1$ do
    sample $\lfloor \frac{T}{\log(K)} \rfloor$ times each arm in $v_t(m)$
    if $\hat{\mu}_{m,t} \leq \tau$ then
        $v_{t+1} = R(v_t)$
    end
    else
        $v_{t+1} = L(v_t)$
    end
end
Set $\hat{a} = v_{T_1+1}(r)$
Output: $(\hat{a}, \widehat{Q}) : \quad \widehat{Q}_k = 2\mathbb{1}_{\{k \geq \hat{a}\}} - 1$
```

Algorithm 21: `Naive`

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

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

Algorithm 22: Uniform

Chapter 4

Bandits with Many Optimal Arms

In this chapter we present the following work, [45], authored by Rianne de Heide, James Cheshire, Pierre Ménard and Alexandra Carpentier.

Acknowledgements The work of J. Cheshire is supported by the Deutsche Forschungsgemeinschaft (DFG) GRK 2297 MathCoRe. The work of P. Ménard is supported by the SFI Sachsen-Anhalt for the project RE-BCI. 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 FG DFG, by the DFG CRC 1294 'Data Assimilation', Project A03, by the Forschungsgruppe FOR 5381 „Mathematische Statistik im Informationszeitalter – Statistische Effizienz und rechentechnische Durchführbarkeit", Project 02, by the Agence Nationale de la Recherche (ANR) and the DFG on the French-German PRCI ANR ASCAI CA 1488/4-1 "Aktive und Batch-Segmentierung, Clustering und Seriation: Grundlagen der KI" and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the SFI Sachsen-Anhalt for the project RE-BCI.

4.1 Introduction

In the classical stochastic multi-armed bandit model – see [63] for a recent survey – a learner interacts with an environment in several rounds. At each round, the learner chooses an *arm* to play, and receives a random reward from the associated probability distribution. Popular settings are respectively the fixed budget *cumulative regret setting* [76], and *best-arm identification setting* [32, 12, 2]. In the first setting, the learner is interested in maximizing the sum of rewards gathered – or minimizing the cumulative regret – and in the best-arm identification setting, the learner is asked at the end of the game to output a guess for the arm with the largest mean reward, and is interested in the quality of this guess – typically measured by the probability of error in the guess.

In most of the papers that concern this topic, it is assumed (i) that there is a single optimal arm, i.e. arm with highest mean, and (ii) that the number of arms is bounded and small when compared to the time horizon, i.e. the number of rounds where the player is allowed to choose an arm. However in many realistic applications, it is not the case, for example in image classification, mining of resources, personalized medicine, or hyperparameter tuning (see [9] for more examples). And while it is clear that in all generality, the task of the learner becomes unsolvable if the number of arms is too large, it intuitively makes sense that if the proportion of optimal arms is also large, this should help the learner.

In this paper, we lift both assumptions summarised in (i) and (ii) and study both the cumulative regret and best-arm identification setting. See Section 4.1.3 for literature related to this that we will discuss later. We will focus on the *problem* *dependent setting* and will aim at characterising optimal learning rates depending on the proportion of optimal arms, and on the minimal gap between the mean of an optimal arm and the mean of a sub-optimal arm.

4.1.1 Setting

We consider a setting with a (potentially infinite) set of arms $\mathcal{A}$, which we call the *reservoir*. Each arm $a \in \mathcal{A}$ is associated with a probability distribution ν_a, which we assume to be supported on $[0,1]$, and we denote its mean by μ_a. Write $\mu^* = \max_{a\in\mathcal{A}} \mu_a$ for the highest mean[1], $\mu_{sub} = \sup_{a\in\mathcal{A}:\mu_a\neq\mu^*} \mu_a$ for the second highest mean, and $\Delta = \mu^* - \mu_{sub}$ for the associated minimal gap. We will focus throughout this paper on the case where $\Delta > 0$.

We further assume that there exists a partition $\mathcal{A} = \mathcal{A}^* \cup \mathcal{A}_{sub}$ such that each arm $a \in \mathcal{A}^*$ is optimal, i.e. $\mu_a = \mu^*$, and each arm $a \in \mathcal{A}_{sub}$ is sub-optimal, i.e. $\mu_a \leq \mu_{sub}$. We assume that the agent can pick arms uniformly at random from the reservoir $\mathcal{A}$[2], and this arm belongs either to the set $\mathcal{A}^*$ with probability $p^\star$, i.e. there is a proportion $p^\star$ of optimal arms in the reservoir; or it belongs to the set $\mathcal{A}_{sub}$ with probability $1-p^\star$, i.e. there is a proportion $1-p^\star$ of sub-optimal arms in the reservoir.

The learner interacts with the environment in several rounds $t = 1, 2, \ldots, T$, where we fix the time horizon T. At each round $t \leq T$, the learner chooses an arm a_t by either picking a new arm from the reservoir $\mathcal{A}$ or playing a past arm, and gets a reward $Y_t \sim \nu_{a(t)}$. The arm choice depends only on the past observations, the past arm choices, and possibly some exogenous randomness. The rewards for each arm a are i.i.d. random variables with mean μ_a unknown to the learner.

Cumulative regret setting. The first setting we study is that of minimizing the *cumulative regret*. This setting enforces the *exploration-exploitation trade-off*: the learner needs to balance exploratory actions to get a better estimate of the reward distributions, and exploitative actions to maximize the total return – and minimise the associated cumulative regret. The cumulative regret is the difference between the sum of expected rewards the learner would have obtained by only choosing the arm with the highest mean reward, and the sum of expected rewards she actually collected:

$$R(T) = \sum_{t=1}^{T} \mu^\star - \mu_{a(t)} \,.$$

Best-arm identification setting In the second setting we study, we are interested in identifying an arm with the highest mean reward. At the end of T rounds, the agents selects an arm $\hat{a}_T$ and aims at minimising the probability of outputting an arm with sub-optimal mean:

$$\mathrm{e}(T) = \mathbb{P}(\hat{a}_T \notin \mathcal{A}^*).$$

A closely related popular measure of error is the *simple regret*, which is not discussed in this paper.

Equivalent settings Firstly, our setting is directly applicable to the problem of competing against the j-th best arm, where we assume w.l.o.g. the arms to be ordered according to their means. Indeed our setting translates to this if we replace $p^\star$ by j/K

[1] We assume that it is attained for some arm(s).

[2] In case of infinite $\mathcal{A}$, one can obviously not sample from a uniform distribution. Our analysis extends to general distributions on $\mathcal{A}$.

and Δ by the gap between the $j/2$-th and the $j+1$-th best arm, i.e. $\Delta = |\mu_{j/2} - \mu_{j+1}|$. Secondly, our setting is directly applicable to that of identifying an ε good arm, and thirdly, our setting is directly applicable to finding any arm in the reservoir with a mean larger than the quantile of a known order – see the discussion in Section 4.1.3.

4.1.2 Contributions

We characterise the optimal learning rates both for the cumulative regret setting, and for best-arm identification, for our problem described above. We characterise the optimal learning rates in terms of the problem parameters $T, p^\star$, and Δ.

In order to describe our results, let us write for $\bar{\Delta} > 0$, $\bar{p^\star} \in [0, 1)$: $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$, for the set of bandit problems whose reservoir distribution is such that $p^\star \geq \bar{p^\star}$ and such that $|\bar{\mu}^* - \mu_{sub}| \geq \bar{\Delta}$.

Cumulative regret We provide an algorithm, *that takes $p^\star$ as a parameter*, that is such that (see Theorem 31)

$$\mathbb{E}R(T) \leq O\left(\frac{\log T \log(1/\Delta)}{p^\star \Delta}\right).$$

Conversely, we prove in Theorem 32 that for $\bar{p^\star} \leq 1/4$ and $\bar{\Delta} \leq 1/4$, and for any algorithm, there exists a problem in $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$ such that

$$\mathbb{E}R(T) \geq \Omega\left(\frac{\log T}{\bar{p^\star}\bar{\Delta}}\right).$$

These two bounds match up to a multiplicative factor of order $\log(1/\Delta)$. They highlight the intuitive fact that we should pay the number of arms in the rate only relative to the number of optimal arms – i.e. only through $p^\star$. Indeed, the probability of picking an optimal arm in the reservoir when sampling uniformly at random being $p^\star$, if we sample about $1/p^\star$ arms at random from the reservoir, we will have sampled one optimal arm with constant probability – so that $1/p^\star$ plays the same role as the number of arms.

Having said that, there is a main conceptual difficulty in order to get a rate that is tight in terms of its dependence in T. If we sample only $1/p^\star$ arms from the reservoir, the probability of having no optimal arms in the chosen set of arms is also a constant – so that the regret is linear in T. It is therefore essential to sample *more* arms. In order to have a logarithmic regret in T, we need to sample at least about $\log T/p^\star$ arms from the reservoir – in which case at least one of them is optimal with probability polynomially decaying with T. But if we do this, we get a regret of order $\frac{(\log T)^2}{p^\star \Delta}$, as there are about $\log T/p^\star$ sub-optimal arms whenever $p^\star$ is not too close to 1. This is much larger than the bound that we have, where the dependence on T is only $\log T$. In order to achieve this bound, we need to take into account the fact that when sampling $\log T/p^\star$ arms from the reservoir, there is typically not just 1, but $\log T$ optimal arms with high probability – and leverage this fact both in our algorithm and in the associated proof. We describe this in more detail in Section 4.2.1.

Best-arm identification We provide an algorithm *that does not take $p^\star$ as a parameter*, such that,

$$\mathrm{e}(T) \leq O\left(\log(T) \exp\left(-c\frac{T\Delta^2 p^\star}{\log(T)}\right)\right),$$

where c is some universal constant. Conversely, we prove that for $p^\star \leq 1/4$ and $\Delta \leq 1/4$, and for any algorithm, there exists a problem in $\mathfrak{B}_{\bar{\Delta},\bar{p^\star}}$ such that $\mathrm{e}(T) \geq \Omega\big(\exp\big(-cT\Delta^2 p^\star\big)\big)$, where $c > 0$ is some universal constant. These two bounds match in order up to a factor of order $\log(T)$ in the exponential, it is an open question here whether this term is necessary or not.

These bounds highlight the intuitive fact that we should pay the number of arms in the rate only relative to the number of optimal arms – i.e. only through $p^\star$. As in the cumulative regret setting, if we sample about $1/p^\star$ arms at random from the reservoir, we will have sampled one optimal arm with constant probability – so that $1/p^\star$ plays the same role as the number of arms.

As in the cumulative regret setting, there is again a main conceptual difficulty in order to get a rate that is tight in terms of its dependence in T. If we sample only $1/p^\star$ arms from the reservoir, the probability of having no optimal arms in the chosen arms is also a constant – which is way smaller than the targeted best-arm identification probability. In order to have at least one optimal arm in the set of arms picked from the reservoir with a probability that decays exponentially with $p^\star T\Delta^2$, the number of arms that have to be sampled should be larger than $T\Delta^2$. But if we do this, we get an upper bound on the probability of error that is of constant order – which is much larger than the bound that we have. In order to obtain our upper bound, we need to take into account the fact that when sampling $T\Delta^2$ arms from the reservoir, there is typically not just 1, but $p^\star T\Delta^2$ optimal arms with high probability – and leverage this fact both in our algorithm and in the associated proof. We describe this in more detail in Section 4.3.1.

Adaptation to $p^\star$: diverging pictures for cumulative regret and best-arm identification The algorithm for cumulative regret takes (a lower bound on) $p^\star$ as parameter, but the algorithm for best-arm identification does not take anything related to $p^\star$ or Δ as a parameter. And so, while our algorithm for best-arm identification is adaptive to $p^\star$ and Δ, our cumulative regret algorithm is adaptive to Δ but not $p^\star$. In Section 4.2.3 we prove that it is not just a weakness of our analysis, but that it is *impossible to adapt to $p^\star$ when it comes to the cumulative regret*. The phenomenon of adaptation to the problem hyper-parameters being possible for best-arm identification but not for cumulative regret, was observed earlier: In the $\mathcal{X}$-armed bandit setting [67] show it is impossible to adapt to smoothness and [44] further classifies the cost of adaptation in this case. [89] explore the cost of adaptation to $p^\star$ for the problem independent case where the number of arms is large.

4.1.3 Related work

Finite and small number of arms. The regret-minimization setting, introduced by [76], has been well-studied for *finite*-armed bandit models. Algorithms for this problem fall into several categories: algorithms based on upper-confidence bounds (UCB) for the unknown arm means [54, 4, 6, 16], algorithms that exploit a posterior distribution on the means, such as Thompson Sampling [80, 60], and many more such as explore-then-commit [38] and phased-elimination [31]. Logarithmic instance-dependent lower bounds have already been obtained in the seminal paper by [62], and were generalized later, e.g. by [15], see [40] for an overview and simple proofs. In the setting where the number of arms $|\mathcal{A}|$ is finite and not too large – much smaller than T – a classical problem dependent upper bound on the expected cumulative regret is[3]

$$\sum_{a\in\mathcal{A}\setminus\mathcal{A}^*}\left(\frac{8\log T}{\mu^*-\mu_a}+2\right)\leq|\mathcal{A}_{sub}|\frac{\log T}{\Delta}+2|\mathcal{A}_{sub}|. \tag{4.1}$$

The bound in the RHS is tight if all sub-optimal arms have the same gap Δ. Moreover, this regret bound asymptotically matches the lower bound by [15] up to a multiplicative constant. In the case where there are infinitely many sub-optimal arms, on the other hand, this upper bound is infinite, *even when the proportion of optimal arms $p^\star$ is large and where one would hope for better performances.*

The fixed-budget best-arm identification setting was introduced by [12, 2] and has been widely studied. It is well-known that algorithms that are optimal for cumulative-regret minimization cannot yield optimal performance for best-arm identification [11, 58]. Write[3] $H=\sum_{a\in\mathcal{A}\setminus\mathcal{A}^*}\frac{1}{(\mu^*-\mu_a)^2}\leq\frac{|\mathcal{A}_{sub}|}{\Delta^2}$. The bound in the RHS is tight if all sub-optimal arms have gap Δ. It is proven by [2] that given H, there exists an algorithm such that the probability of misidentifying an optimal arm is of order $\exp(-cT/H)$, where $c>0$ is some universal constant. In the case where there is *a single optimal arm* this bound is provably optimal [17] when H is known. However, in the case where there are infinitely many sub-optimal arms this upper bound is larger than 1 and thus vacuous, *even when the proportion of optimal arms $p^\star$ is large and where one would hope for better performances.*

Importantly, our results in both settings extend to finite bandits. Furthermore we do not need infinite $\mathcal{A}$ for our results to be near optimal. In the finite setting with K arms and $p^\star K$ optimal arms the problem is strictly harder than one with $\frac{1}{p^\star}$ arms and a single optimal arm. Indeed, the latter problem would correspond to one where the learner receives, as additional information, a partition of the set of K arms in $\frac{1}{p^\star}$ groups, where one of the groups contains all optimal arms, and the others are only composed of sub-optimal arms. One can then see that we match the classical UB and LB for the finite bandit problem, up to $\log(1/\Delta)$ terms.

Large to infinite number of arms. The setting with an infinite number of arms – and sometimes also many optimal arms – has been studied in different settings.

A setting that is very related to ours is the infinitely many-armed setting where a distribution is assumed on the reservoir – called the reservoir distribution. At each round, the learner can pull a previously queried arm, or a new arm that is sampled according to the reservoir distribution. A classical assumption on the reservoir is that the proportion of $\bar{\Delta}$-near optimal arms is of larger order than $\bar{\Delta}^{-\alpha}$ for any $\bar{\Delta}$. This setting been studied for both cumulative regret minimization [9, 83, 10, 29] and for best-arm identification [18, 7, 20]. A classical strategy is to select a subset of arms from the reservoir, large enough so that it contains a near optimal arm with high probability, and to use classical bandit strategies on these arms. The minimax order of magnitude of the cumulative regret is then $\sqrt{T}\vee T^{\alpha/(\alpha+1)}$ and for the simple regret it is $T^{-1/2}\vee T^{-1/\alpha}$.

Related results have also be obtained in the setting where the number of arms is finite, but large – i.e. $K>T$ – and under related assumptions on the frequency of near-optimal arms [89]. While our setting is extremely related to this setting, the assumption about the frequency of near-optimal arms differs in the above literature from the assumption we make in this paper. Their bounds are not dependent upon Δ – they assume $\forall k\in[K],\mu_k\in[0,1]$, and instead focus on achieving semi adaptivity

[3] In the case where $\mathcal{A}$ is finite otherwise the quantity below is infinite.

in regards to an unknown α^*, where $\alpha^* := \inf\{\alpha : K/|S_*| < T^\alpha\}$. In the context of our setting T^α would act as a upper bound on $1/p^\star$. They propose an algorithm with user defined parameter β that has no guarantees on regret for $\beta < \alpha$. And while our assumption is more restrictive, we also expect to obtain much smaller optimal rates. Our results differ from this stream of literature in the same way that, in the classical MAB, *problem dependent results differ from problem independent results.*

Another setting takes a regularity assumption on the reservoir distribution around μ^* – that is, the proportion of arms in the reservoir whose gap is of order greater than $\bar{\Delta}$ is bounded above by a function of $\bar{\Delta}$, typically $\bar{\Delta}^\alpha$, where α is the regularity coefficient. For best-arm identification adaptivity is possible without knowledge of α and [18] provide algorithms for the simple regret with LB matching up to $\log(T)$ terms. In the case of cumulative regret [83] and [10] again provide near optimal results but in the case of *known* α. While the above literature considers a weaker assumption on the reservoir distribution, their results are also considerably weaker than our own. For best-arm identification they identify a sub optimal arm whose distance to the optimal arm is bounded polynomially with T. For cumulative regret the regret is bounded polynomially with T. These bounds are in both cases much larger than our bounds – which essentially reflects that their assumption are weaker.

Closer to our setting are the works [20] and [7], where they try to find any arm in the reservoir with a mean larger than the quantile of a known order (with respect to the reservoir distribution) with high probability. This can be seen as the fixed confidence version of our setting for best-arm identification where the order of the quantiles is our known proportion of optimal arms $p^\star$ and the gap Δ is the difference between the first and the second quantile of order $p^\star$. Precisely, [7] provide an algorithm that can find an arm above the quantile of order $p^\star$ with probability at least $1 - \delta$ in less than $H_{\Delta,p^\star} \log(1/\delta)^2$ samples on average, where $H_{\Delta,p^\star} \approx 1/(p^\star\Delta^2)$ is the problem dependent constant. The fixed confidence result of [7] translates, in the fixed budget setting, into an upper bound on the probability of error $\mathrm{e}(T)$ of order $\exp(-c\sqrt{Tp^\star\Delta^2})$ where $c > 0$ is some universal constant – which is much larger than our bound for large T. Similarly, [21] consider the regret with respect to a fixed quantile of order $p^\star$ of the distribution of the means in the reservoir which is again quite related to the regret in our setting. They obtain an algorithm with a bound on cumulative regret of order $R(T) \leq O\big(1/p^\star + \sqrt{(T/p^\star)\log(p^\star T)}\big)$, for any $\Delta > 0$ – in this sense, this analysis is problem independent.

Also closely related is the paper [56] which deals with identifying an ε good arm – in the case where there are many such ε good arms, with high probability. Again this can be seen as a fixed confidence version of our setting, with the proportion of ε good arms being equivalent to our $p^\star$. However, the focus of their results differs considerably to our own. Specifically, in our setting, Theorem 2 of [56] provides an upper bound on the expectation of a stopping time for epsilon good arm identification, of the order $\bar{\mathcal{H}} \log(\bar{\mathcal{H}})$ where $\bar{\mathcal{H}} \approx 1/(p^\star\Delta^2)\log(1/\delta)$ but this bound does not hold in high probability, which would be necessary if one wished to directly compare their results to ours. Indeed for the stopping time of their algorithm to be bounded in high probability one would need to pay a $\log(1/\delta)^2$ term, corresponding to $\exp(-\sqrt{\Delta^2 p^* T})$ in our setting, see Remark 4 in [56] and page 15 in the appendix of the full version [55]. The focus of [56] is instead to get more complete gap dependent bounds, considering also the gaps within the epsilon good arms but as mentioned their results cannot be applied directly to our setting and, as they point out, extending their approach to include high probability guarantees would be strictly sub optimal compared to our results.

We can also view the *most-biased coin problem* studied by [19] and [49] as a particular instance of our setting where all optimal arms are distributed according to a Bernoulli distribution $\mathcal{B}\mathrm{er}(\mu^\star)$ and any sub-optimal arm is distributed according to the *same* Bernoulli distribution $\mathcal{B}\mathrm{er}(\mu^-)$. The goal is then to identify an optimal arm with high probability with as few samples as possible. Precisely, [49] prove that they can find an optimal arm with probability at least $1-\delta$ with $\log\big(1/(p^\star\Delta^2)\big)\frac{\log(1/\delta)}{p^\star\Delta^2}$ samples in expectation when $\mu^\star, \mu^-$ and $p^\star$ are unknown to the agent and with $\frac{\log(1/\delta)}{p^\star\Delta^2}$ samples if $p^\star$ is known. It is also worth mentioning the problem of $p^\star$ estimation for the biased coin problem. For unknown $p^\star$ and Δ, [65] describe, in the fixed confidence setting, the optimal learning rate for estimating $p^\star$, up to an additive error ε, of the order $\frac{p^\star}{\varepsilon^2\Delta^2}\log(1/\delta)$.

The translation of the result from [49] to the fixed budget setting is much closer to our result, as it would provide a bound of order $\exp\big(-cTp^\star\Delta^2/\log(1/(p^\star\Delta^2))\big)$ where $c > 0$ is some universal constant. This is very similar to our bound, but there is a main difference: we do not assume that there are just two possible distribution for the arms as [49] – the set $\mathcal{A}_{sub}$ of sub-optimal arms might contain arms of diverse means, all being at a gap more than Δ from μ^*. This makes the problem *significantly more difficult* – in particular regarding the adaptation to $p^\star$ – since in our setting, it is impossible to estimate the minimal gap Δ, see Section 4.5. In fact, extending to a more general reservoir is an open question of interest left at the end of the above paper.

Otherwise, there are some other formulations of the infinitely-many armed bandit problem that are quite popular, but very different from our setting, and that we mention here for completeness. Many works are devoted to the setting where there is some topological relation between the index of the arms, and the mean of the arms [61, 13, 43]. This setting is often referred to as the $\mathcal{X}-$armed bandit setting, and not related to our work as we do not make such topological assumptions. Finally, a paper in which the setting is close to ours, but where the goal is very different, is the one by [50]. The authors consider a partition of the (infinite) space Ω of K-armed bandit models $\nu = (\nu_1, \ldots, \nu_K)$, and want to identify for a given bandit model $\mu \in \Omega$ the correct partition component it belongs to.

Fixed confidence to fixed budget setting In the fixed confidence setting for best-arm identification, given some $\delta > 0$, one aims to bound the expected number of samples one needs to correctly identify an optimal arm with probability greater than $1-\delta$. With our best-arm identification upper bound (Theorem 34) in mind, we can essentially translate our result to the fixed confidence setting by considering $\delta = \exp\left(-\frac{Tp^\star\Delta^2}{\log(1/\Delta)}\right)$, and solving for T. This leads to a upper bound on the number of samples `Elimination` needs to be δ-approximately correct of: $\frac{\log(\frac{1}{\delta})\log(\frac{1}{\Delta})}{p^\star\Delta^2}$. The papers [49] and [7] both deal with settings very related to our own but from the fixed confidence perspective. [7] deals with quantile estimation and as highlighted above their results can be applied to our setting but with a significantly worse bound on probability of error of order $\exp(\sqrt{Tp^\star\Delta})$. In [49] the problem of best-arm identification is tackled directly but with strong restriction on the reservoir distribution, they consider the case were all sub optimal arms are identically distributed.

Pair matching An additional setting that can be seen in the context of our problem is that of pair matching. Here the learner is presented with a finite graph of nodes, N. The set of nodes N is partitioned into 2 or more communities. The general idea is that nodes in the same community are more likely to be connected by an edge than those in separate communities. A simple and well studied situation is where the graph is generated according to a stochastic block model (SBM), see [46]. In this setting the probability of an edge forming between two nodes of the same community is p and the probability of an edge forming between two nodes of differing communities is q, with $p > q$. Much of the literature is then concerned with identifying communities given complete access to the graph, see [64] and references therein. Of more relation to our specific setting is the paper [42]. Here the learner does not immediately observe the complete graph but is instead able to sequentially query whether two nodes are connected up to a budget T. Their objective is then to minimise their sampling regret, the number of times they query and edge between 2 nodes of differing communities. The problem can be viewed as a bandit problem where each pair of vertices represents an arm following a bernoulli distribution of mean p or q. In our setting the minimal proportion p^* would then be the proportion of pairs which belong to the same community and the gap as $\Delta = p - q$. The fundamental difference is that each arm can only be pulled once, making the problem significantly harder, however, the learner can exploit the SBM structure to their advantage. Assuming the case with exactly two equally sized communities with $T \leq |N|^2$, in [42] they show it is possible to attain a sub linear regret of the order $T\Delta \wedge \frac{(p+q)T}{\Delta}$. The significant worsening of their rate, in comparison to our own, is due to the fact one cannot sample an arm more than once, which significantly changes the flavour of their algorithms.

4.2 Cumulative regret

We first present an algorithm and prove an upper bound on its cumulative regret, and then we present a problem-dependent lower bound that shows we match the regret bound up to poly-log terms in Δ. Lastly, we provide a theorem to the effect that adaptation to the proportion of optimal arms $p^\star$ is not possible in this setting.

4.2.1 Upper bound

We present Sampling-UCB for cumulative regret minimization. This algorithm is an Upper Confidence Bound (UCB) type algorithm [63]. We first sample a set $\mathcal{L}$ of arms large enough such that with high probability (of order $1 - 1/T$) there is a proportion of order $p^\star$ optimal arms. Then we build an upper confidence bound on the empirical mean of each sampled arm, see (4.2), where $\widehat{\mu}_a^t$ is the empirical mean of arm a at time t and N_a^t the number of times arm a was pulled until time t. At time t we pull the arm $a \in \mathcal{L}$ with the highest upper confidence bound U_a^t. The complete procedure is detailed in Algorithm 23. Notably, we do not tune the upper confidence bounds such that they are exceeded with probability less than $1/T$, as for finite-armed bandits. In that setting, a common choice is to have bonuses of the form $\widehat{\mu}_a^t + \sqrt{2\log(T)/N_a^t}$, see [63]. Instead we use an exploration function that does not depend on T, such that the upper confidence bounds are exceeded with probability smaller than a fixed constant, see (4.2). Thus we only pay a constant regret of order $\log(1/\Delta)$ on the set of sampled arms $\mathcal{L}$. This is made possible by leveraging the fact that we know that there is a proportion of order $p^\star$ optimal arms.

We prove the following regret bound for Sampling-UCB in Appendix 4.6.

Theorem 31. *For $T \geq 2$, $\gamma \in (0,1)$ and $L = \lceil 4\log(T)/(p^\star\gamma^2)\rceil$, the expected cumulative regret of Sampling-UCB is upper bounded as follows:*

$$\mathbb{E}R(T) \leq O\left(\frac{\log(T)\log(1/\Delta)}{p^\star\Delta}\right),$$

```
Input: $\gamma \in (0,1)$, $L \geq 1$
Initialize: Pick $\mathcal{L}$, with $|\mathcal{L}| = L$, arms from the reservoir $\mathcal{A}$. Sample each arm once.
for $t = L+1$ to $T$ do
    Compute for each arm $a \in \mathcal{L}$ the quantity
        $U_a^t = \widehat{\mu}_a^t + \sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(N_a^t)}{2N_a^t}},$        (4.2)
    Play $a_t = \arg\max_{a \in \mathcal{L}} U_a^t$.
end
```

Algorithm 23: Sampling UCB

see the end of the proof for a precise bound, i.e. (4.3)*.*

Note that this bound matches the lower bound of Theorem 32 of Section 4.2.2, for T large enough and up to a $\log(1/\Delta)$ multiplicative factor. Also, L can be calibrated with a lower bound on $p^\star$ instead of $p^\star$, but this lower bound will appear in the rate instead of $p^\star$.

Remark 15. Algorithm `Sampling-UCB` samples L arms uniformly at random from the reservoir. What we mean by this is that each arm is pulled at random from $\mathcal{A}$ *independently from the other pulled arms*. In other words, by doing this, we potentially artificially create several independent copies of the same arm – which might seem counter-intuitive, but is formally not a problem.
What this anyway implies is that the case $|\mathcal{A}| \leq L$ is not a problem – with this idea of independent copies, we can pull more arms from the reservoir than the number $|\mathcal{A}|$ of arms.

Remark 16. Our algorithm is reminiscent of that of [47], which, as our own, uses a UCB which does not depend on the time horizon, but only on the number of times an arm has been pulled. However, they do so for different reasons, namely to adapt to the infinite time horizon of the fixed confidence setting.

4.2.2 Lower bound

We can prove an equivalent of the [62] lower bound for finite-armed bandits for our setting. The following theorem is proved in Appendix 4.6.

Theorem 32. *Consider $\Delta \in (0, 1/4)$ and $p^\star \in (0, 1/4]$. For any bandit algorithm, there exists a bandit problem in $\mathfrak{B}_{\Delta,p^\star}$ such that*

$$\mathbb{E}R(T) \geq \min\left(\frac{1}{60}\frac{\max\big\{\log(\Delta^2 T/16), 0\big\}}{p^\star\Delta}, \sqrt{T}\right)$$

Note that if we consider the gap Δ and the proportion of optimal arms $p^\star$ as fixed and T large in comparison, i.e. $\Delta \gg \sqrt{1/T}$, then our lower bound is of order $\log(T)/(p^\star\Delta)$. This is the problem-dependent regime that we consider in this paper. On the contrary, if $\Delta \approx \sqrt{1/T}$ then our lower bound is of order $\sqrt{T}$. This is rather the problem-independent regime studied by [21]. We can make a parallel between the lower bound in our setting and the one for finite-armed bandits. Indeed, if we consider that the proxy for the number of arms is $|\mathcal{A}| \sim 1/p^\star$ which implies that there is $p^\star|\mathcal{A}| \sim 1$ optimal arm, then we recover the problem-dependent lower bound of order $|\mathcal{A}|\log(T)/\Delta$, if there are $|\mathcal{A}|-1$ sub-optimal arms with gap Δ.

4.2.3 Impossibility of adapting to $p^\star$

The following theorem shows that in the setting of minimizing the cumulative regret, it is impossible to adapt to the proportion of optimal arms $p^\star$. The theorem is proved in Appendix 4.6.

Theorem 33. *Let $p^\star \le \frac{1}{4}$ and $c > 0$ such that $T \ge 4\Big(\frac{c\log(T)}{p^\star\Delta^2}\Big)^2$. For any bandit algorithm $\mathfrak{A}$ such that for all bandit problems in $\mathfrak{B}_{\Delta,p^\star}$, we have,*

$$\mathbb{E}R(T) \le \frac{c\log(T)}{p^\star\Delta}$$

one has that $\forall q^\star \le \frac{4p^\star}{c}$ there exists a problem in $\mathfrak{B}_{\Delta,q^\star}$ such that

$$\mathbb{E}R(T) \ge \frac{\sqrt{T}\Delta}{4}\,.$$

Remark 17. The Sampling-UCB algorithm takes a user defined parameter γ (which can be taken as a universal constant) and L, which should be calibrated depending on (a lower bound on) $p^\star$. While this is necessary, it is important to not that none of the parameters requires knowledge of Δ.

4.3 Best-arm identification

We present our Elimination algorithm for best-arm identification, together with an upper bound on the probability of outputting a sub-optimal arm; next we prove a lower bound, which is matched by our upper bound up to a $1/\log(T)$ factor in the exponential.

4.3.1 Upper bound

As its name suggests, the Elimination algorithm (summarized in Algorithm 24) works by successive elimination of arms – through the update at round i of a set $\mathcal{A}_i$ – although with a twist. We begin by sampling approximately T arms at the first round. Namely, we first select a set $\mathcal{A}_1$ of $|\mathcal{A}_1| = \lfloor \bar{c}T/\log T\rfloor$ arms taken at random from the reservoir, for some constant $\bar{c} > 0$. Then at each round we use a $T/\log T$ fraction of our budget to sample the arms in our set. And so at round i we sample each arm in the set $\mathcal{A}_i$ a number of $t_i = \lfloor \bar{c}T/(|\mathcal{A}_i|\log T)\rfloor$. We then eliminate half of the arms based on the arms' empirical means – namely, we just keep the $\big\lfloor |\mathcal{A}_i|/2\big\rfloor \vee 1$ arms in $\mathcal{A}_i$ that have highest empirical means – and introduce an additional number of arms sampled from the reservoir distribution – namely $\big\lfloor |\mathcal{A}_i|/4\big\rfloor$ – such that the final size of our arm set is reduced by $\frac{3}{4}$. At the end of the budget, we have one arm remaining – due to the choices of $\bar{c}$ – which is the arm that we return. Note that Remark 15 applies here too so that it is not a problem if $|\mathcal{A}|$ is smaller than the number of arms required by the algorithm. Theorem 34 is proved in Appendix 4.7.

Theorem 34. *Set $\bar{c} = \log(4/3)$. Elimination satisfies*

$$\mathbb{P}(\hat{a}_T \in \mathcal{A}^\star) \ge 1 - 2\log(T)\exp\left(-c\frac{\Delta^2 p^\star T}{\log T}\right),$$

```
Input: $\bar{c}$
set $i \leftarrow 1$
while $i < \log T/\bar{c}$ do
    Sample each arm in $\mathcal{A}_i$ a number $t_i = \lfloor \bar{c}T/(|\mathcal{A}_i| \log T) \rfloor$ of times and
      compute their empirical means $(\hat{\mu}_i(a))_{a \in \mathcal{A}_i}$
    Put in $\mathcal{A}_{i+1}$ the $1 \vee \lfloor |\mathcal{A}_i|/2 \rfloor$ arms that have highest empirical means
      $(\hat{\mu}_i(a))_{a \in \mathcal{A}_i}$, and add on top of that $\lfloor |\mathcal{A}_i|/4 \rfloor$ new arms taken at random
      from the reservoir
    $i \leftarrow i + 1$
end
Return any $\hat{a}_T$ in $\mathcal{A}_i$
```

Algorithm 24: Elimination

where $c = \bar{c}/19200$

Remark 18. Elimination works by discarding many sub-optimal arms and few optimal arms in each round, so that at the end, when just one arm remains, it is optimal with high probability. A key element is that Elimination adds *fresh arms* from the reservoir at each round. This is to ensure that our algorithm is adaptive to $p^\star, \Delta$, as ensured by Theorem 34. Whenever the arms in $\mathcal{A}_i$ are pulled less than about Δ^{-2} times, there is no guarantee on what happens when half of the arms are eliminated. Therefore, we have to make sure that when the algorithm arrives at a round i such that $t_i \gtrsim \Delta^{-2}$, the proportion of optimal arms is of larger order than $p^\star$ with high enough probability. This is ensured by adding the fresh arms added from the reservoir. Note that for some arm distributions, we do not need to add fresh arms and the algorithm would function also by just halving at each step the number of arms. Indeed, in the case where all arms follow a Bernoulli distribution, in terms of preserving the proportion of optimal arms, one can prove that halving the set of arms according to the empirical means is no worse than random halving of the set. Thus, in this case, with high probability we increase the proportion of optimal arms at each step, without diminishing it. This is however specific to the case of Bernoulli distributions and some other parametric families, and it is an open question whether this would be true in general.

Remark 19. The successive halving strategy our algorithm for best-arm identification is based on was first introduced by [52], however, without the trick of adding fresh arms, as they didn't need to be adaptive to $p^\star$.

4.3.2 Lower bound

The following Theorem provides a lower bound on the probability of error for best arm identification in our setting. The proof of Theorem 35 can be found in Appendix 4.7.

Theorem 35. *Consider* $\Delta \in (0, 1/4)$ *and* $p^\star \in [0, 1/4]$. *For any bandit algorithm, there exists a bandit problem in* $\mathfrak{B}_{\Delta,p^\star}$ *such that*

$$\mathrm{e}(T) \geq \frac{1}{4} \exp\left(-Tp^\star \frac{\Delta^2}{32}\right).$$

In proving the above theorem we essentially show that an agent cannot accurately distinguish between two cases: $\mu^* = \frac{1}{2}$ and $\mu^* = \frac{1}{2} + \Delta$. That is, we consider two reservoirs $\mathbf{R}_0$ and $\mathbf{R}_1$ where $\mu_0^* = \frac{1}{2}$ and $\mu_1^* = \frac{1}{2} + \Delta$. Using a coupling argument we bound the KL divergence between the distribution of samples collected on $\mathbf{R}_0$ and $\mathbf{R}_1$. The results then follows by application of Bretagnolle-Huber's inequality.

4.4 Experiments

We conduct a preliminary set of experiments to test the performance of our algorithms. Specifically, for cumulative regret we compare our `Sampling-UCB` to the QRM1 algorithm by [21] and the SR algorithm by Zhu and Nowak [89]. For simple regret we compare our `Elimination` to the BUCB algorithm by [56]. In both cases our performance appears comparable to the literature. See Appendix 3.9 for details.

4.5 Conclusion and open questions

Classifying optimal learning rates on the continuous armed bandit problems with a proportion of optimal arms and general reservoir distribution has been a question of interest in the literature for some time, see [49]. Recent papers – [7] and [89], while focused on a slightly different setting, have considerably weaker results when applied to our setting. Therefore, we believe our results mark a significant improvement in the state of the art. An extension of our results would be to remove the $\log(1/\Delta)$ discrepancy between UB and LB for cumulative regret. However, this appears non-trivial and in particular we struggle to see how a UCB based strategy would achieve this tighter bound in the case of the cumulative regret. Another possibility for further work is an expansion of our setting. Consider the arm reservoir $\mathcal{A}$ partitioned into K possible distributions, each with associated probability p_k. Let $k^* = \arg\max_{[K]} \mu_k$ and take gaps $(\Delta_k)_{[K]} = (\mu_{k^*} - \mu_k)_{[K]}$. One could then consider more detailed bounds, dependent on the sequence $((p_k, \Delta_k))_{[K]}$ as opposed to just $p^\star$ and the smallest gap. The main difficulty here would be to deal with the case where some p_k are much smaller than the proportion $p^\star$ corresponding to the optimal arm.

4.6 Cumulative regret proofs

4.6.1 Upper Bound

Proof of Theorem 31. We denote by $\mathcal{L}$ the set of arms sampled from the reservoir such that $|\mathcal{L}| = L$. We also denote by $\mathcal{L}^\star = \{a \in \mathcal{L} : \; a \in \mathcal{A}^\star\}$ the set of optimal arms in $\mathcal{L}$ and by $L^\star = |\mathcal{L}^\star|$ its cardinality. Note that these quantities are all random.

Because of the choice of $L = \lceil 4\log(T)/(p^\star\gamma^2)\rceil$, we know that with high probability there is at least a proportion of $\gamma p^\star$ optimal arms in $\mathcal{L}$. Precisely, if we denote this favorable event by $\mathcal{E} = \{L^\star/L \geq (1-\gamma)p^\star\}$ then by Chernoff's inequality (see Lemma 26), we have

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\big(L^\star/L < (1-\gamma)p^\star\big) \leq e^{-\frac{\gamma^2}{4}Lp^\star} \leq \frac{1}{T}\,.$$

We can decompose the regret given this event and its complement:

$$\begin{aligned}\mathbb{E}[R(T)] &= \mathbb{E}\left[\sum_{a\in\mathcal{L}}(\mu^\star - \mu_a)\mathbb{E}[N_a^T|\mathcal{L}]\mathbb{1}_{\mathcal{E}}\right] + T\mathbb{P}(\mathcal{E}^c)\\ &\leq \mathbb{E}\left[\sum_{a\in\mathcal{L}/\mathcal{L}^\star}\Delta_a\mathbb{E}[N_a^T|\mathcal{L}]\mathbb{1}_{\mathcal{E}}\right] + 1\,.\end{aligned}$$

We now follow the classical proof of UCB-type strategies to upper-bound the number of times a sub-optimal is pulled. From now on, we fix a set of sampled arms $\mathcal{L}$. Fix an $a \in \mathcal{L}\setminus\mathcal{L}^\star$. We have

$$\mathbb{E}[N_a^T|\mathcal{L}] \leq 1 + \sum_{t=L+1}^{T}\mathbb{P}(\forall b\in\mathcal{L}^\star,\, U_{t-1}^b \leq \mu^\star|\mathcal{L}) + \mathbb{P}(a_t = a,\, U_{t-1}^a \geq \mu^\star|\mathcal{L})\,.$$

For the first term in the summation we use the fact that there are many optimal arms. Precisely, using Hoeffding's inequality, we have

$$\begin{aligned}\mathbb{P}(\forall b\in\mathcal{L}^\star,\, U_{t-1}^b \leq \mu^\star|\mathcal{L}) \leq \mathbb{P}\Bigg(&\forall b\in\mathcal{L}^\star, \exists n\in[T]:\; \widehat{\mu}_{b,n}\\ &+\sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(n)}{2n}} \leq \mu^\star\Bigg|\mathcal{L}\Bigg)\\ \leq{}& \prod_{b\in\mathcal{L}^\star}\left(\sum_{n=1}^{T}\frac{1}{n^2}e^{-\gamma^2(1-\gamma)^{-1}/4-\log(\pi^2/6)}\right)\\ ={}& e^{-\frac{\gamma^2}{4}(1-\gamma)^{-1}L^\star}\,.\end{aligned}$$

For the second term we proceed as usual. Let

$$n_0 = \inf\left\{n\in\mathbb{N} : \sqrt{\frac{\gamma^2(1-\gamma)^{-1}/4 + \log(\pi^2/6) + 2\log(n)}{2n}} \leq \Delta/2\right\}$$

be such that pulling any arm $a \in \mathcal{A}_{sub}$ more than n_0 times is a small probability event. Note that thanks to Lemma 27

$$n_0 \le 4\frac{(1-\gamma)^{-1} + \log\big(24(1-\gamma)^{-1}/\Delta^2\big)}{\Delta^2} + 1\,.$$

Then, using again Hoeffding's inequality for an arm $a \in \mathcal{L} \setminus \mathcal{L}^\star$, we obtain

$$\begin{aligned}\sum_{t=L+1}^{T} \mathbb{P}(a_t = a,\, U_{t-1}^a \ge \mu^\star | \mathcal{L}) &\le \sum_{n=n_a+1}^{T} \mathbb{P}(\widehat{\mu}_{a,n} - \mu \ge \Delta/2) + n_0 \\ &\le \sum_{n\ge 1} e^{-n\Delta^2/2} + n_0 \le n_0 + \frac{2}{\Delta^2}\,.\end{aligned}$$

Collecting the previous inequalities we can conclude for $T \ge 2$

$$\begin{aligned}\mathbb{E}[R(T)] &\le \mathbb{E}\left[\sum_{a\in\mathcal{L}/\mathcal{L}^\star} Te^{-\gamma^2(1-\gamma)^{-1}L^\star/4}\mathbb{1}_{\mathcal{E}} + 1 + \Delta n_0 + \frac{2}{\Delta}\right] + 1 \\ &\le \mathbb{E}\left[\sum_{a\in\mathcal{L}\setminus\mathcal{L}^\star} Te^{-\gamma^2 L/4}\mathbb{1}_{\mathcal{E}} + 1 + \Delta n_0 + \frac{2}{\Delta}\right] + 1 \\ &\le L\left(2 + \Delta n_0 + \frac{2}{\Delta}\right) + 1 \\ &\le \frac{8\log(T)}{p^\star\Delta\gamma^2}\Big(10(1-\gamma)^{-1} + 4\log\big(24(1-\gamma)^{-1}/\Delta^4\big)\Big) + 1\,. \qquad (4.3)\end{aligned}$$

□

4.6.2 Lower Bound

We denote by $\mathcal{B}\mathrm{er}(p)$ the Bernoulli distribution of parameter p. The Kullback-Leibler (KL) divergence between probability distributions P and Q is denoted by $\mathrm{KL}(P,Q)$. In particular, the KL divergence between two Bernoulli distributions $\mathcal{B}\mathrm{er}(p)$ and $\mathcal{B}\mathrm{er}(q)$ is

$$\mathrm{kl}(p,q) = \mathrm{KL}\big(\mathcal{B}\mathrm{er}(p), \mathcal{B}\mathrm{er}(q)\big) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

Proof of Theorem 32. We fix a partition of the reservoir $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$ and set $p^\star$ the probability to sample an arm in $\mathcal{A}_1$, $\mathcal{A}_2$ and $1-2p^\star$ the probability to sample an arm in $\mathcal{A}_3$. We define two bandits problems associated with this reservoir. The bandit problem ν where the arms in $\mathcal{A}_1$ have probability distribution $\mathcal{B}\mathrm{er}(1/2)$, the arm in $\mathcal{A}_2$ and $\mathcal{A}_3$ have probability distribution $\mathcal{B}\mathrm{er}(1/2-\Delta)$. The second bandit problem ν' is such that the arms in $\mathcal{A}_1$ have probability distribution $\mathcal{B}\mathrm{er}(1/2)$, the arms in $\mathcal{A}_2$ have probability distribution $\mathcal{B}\mathrm{er}(1/2+\Delta)$ and the arms in $\mathcal{A}_3$ have probability distribution $\mathcal{B}\mathrm{er}(1/2-\Delta)$. We denote by $\mathbb{E}_\nu$ respectively $\mathbb{E}_{\nu'}$ the expectation under the bandit problem ν respectively ν'.

Let $N_{\mathcal{A}_i}^T = \sum_{t=1}^T \mathbb{1}_{\{a_t\in\mathcal{A}_i\}}$ be the number of times an arm in $\mathcal{A}_i$ is pulled. Note that since the arms in $\mathcal{A}_2$ and $\mathcal{A}_3$ are indistinguishable for the agent in the problem ν, it holds

$$\mathbb{E}_\nu[N_{\mathcal{A}_2}^T] = \frac{p^\star}{1-p^\star}\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T]\,.$$

Let I^t be the information available by the agent at time t, i.e. the collection of collected rewards and arms pulled. We denote by $\mathbb{P}_\nu^{I^t}$ respectively $\mathbb{P}_{\nu'}^{I^t}$ the distribution of this random variable under the bandit problem ν respectively ν'. Thanks to the chain rule and the above remark we can upper bound the Kullback-Leibler divergence between these two probability distributions

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_\nu^{I^T}, \mathbb{P}_{\nu'}^{I^T}) &= \mathrm{kl}(1/2-\Delta, 1/2+\Delta)\mathbb{E}_\nu[N_{\mathcal{A}_2}^T] \\ &= \mathrm{kl}(1/2-\Delta, 1/2+\Delta)\frac{p^\star}{1-p^\star}\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T] \\ &\le 22p^\star\Delta^2\mathbb{E}_\nu[N_{\mathcal{A}_2}^T + N_{\mathcal{A}_3}^T] = 22p^\star\Delta\mathbb{E}_\nu\big[R(T)\big]\,, \end{aligned} \tag{4.4}$$

where in the last inequality we used that $p^\star \le 1/4$ and

$$\mathrm{kl}(1/2-\Delta, 1/2+\Delta) = 2\Delta\log\left(1+\frac{2\Delta}{1/2-\Delta}\right) \le \frac{4\Delta^2}{1/2-\Delta} \le 16\Delta^2.$$

We assume that

$$\mathbb{E}_\nu\big[R(T)\big] = \Delta\big(T-\mathbb{E}_\nu[N_{\mathcal{A}_1}^T]\big) \le \sqrt{T}, \qquad \mathbb{E}_{\nu'}\big[R(T)\big] = \Delta\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]+2\Delta\mathbb{E}_{\nu'}[N_{\mathcal{A}_3}^T] \le \sqrt{T},$$

otherwise the result is trivially true. In particular this implies that

$$1-\sqrt{\frac{1}{\Delta^2 T}} \le \frac{\mathbb{E}_\nu[N_{\mathcal{A}_1}^T]}{T} \qquad \frac{\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]}{T} \le \sqrt{\frac{1}{\Delta^2 T}}\,. \tag{4.5}$$

Using the contraction of the entropy (see Garivier, Ménard, and Stoltz [40]), the inequality $\mathrm{kl}(x,y) \ge x\log(1/y) - \log(2)$ then (4.5), we obtain

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_\nu^{I^T}, \mathbb{P}_{\nu'}^{I^T})) &\ge \mathrm{kl}\big(\mathbb{E}_\nu[N_{\mathcal{A}_1}^T]/T, \mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]/T\big) \\ &\ge \frac{\mathbb{E}_\nu[N_{\mathcal{A}_1}^T]}{T}\log\left(\frac{T}{\mathbb{E}_{\nu'}[N_{\mathcal{A}_1}^T]}\right) - \log(2) \\ &\ge \frac{1}{2}\left(1-\sqrt{\frac{1}{\Delta^2 T}}\right)\log(\Delta^2 T) - \log(2)\,. \end{aligned}$$

The previous inequality with the fact that the Kullback-Leibler divergence is positive yields

$$\mathrm{KL}(\mathbb{P}_\nu^{I^T}, \mathbb{P}_{\nu'}^{I^T}) \ge \frac{3}{8}\log(\Delta^2 T/16)^+\,. \tag{4.6}$$

Indeed if $\Delta^2 T/16 \le 1$ then (4.6) is trivially true. In the other case we have

$$\begin{aligned}\frac{1}{2}\left(1-\sqrt{\frac{1}{\Delta^2 T}}\right)\log(\Delta^2 T) - \log(2) &\ge \frac{3}{8}\log(\Delta^2 T) - \frac{1}{4}\log(16) \\ &\ge \frac{3}{8}\log(\Delta^2 T/16)\,. \end{aligned}$$

Combining (4.4) and (4.6) allows us to conclude

$$\mathbb{E}_\nu\big[R(T)\big] \ge \frac{1}{60}\frac{\log(\Delta^2 T/16)^+}{p^\star\Delta}.$$

Remark 20. During this proof we have fixed a partition of the reservoir $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$. We remind the reader that, the learner does not have access to the categories $\mathcal{A}_1, \mathcal{A}_2$ or $\mathcal{A}_3$ directly, and may only draw an arm from the reservoir. Furthermore, we remind the reader that, upon drawing an arm a from the reservoir, the learner is *not given* the information as to which category a belongs. As a result, a trivial algorithm such as "only pull arms from $\mathcal{A}_1$" is not possible. Therefore, when considering the class algorithms for which our lower bound will hold, we do not need a so called "invariance to labelling assumption", as seen in, for example, Giraud et al. [42][Section 2.2]. This remains true for all lower bounds considered in this paper.

□

4.6.3 Impossibility of adaptation to $p^\star$

Sketch of proof of Theorem 33 For the proof of Theorem 33 we construct two reservoir distributions:

- The reservoir distribution $\mathbf{R}_0$ characterised by $p_1 = p^\star$ and $p_2 = 1 - p^\star$ and $\nu_1 = \mathcal{B}(1/2)$ and $\nu_2 = \mathcal{B}(1/2 - \Delta)$.
- The reservoir distribution $\mathbf{R}_1$ characterised by $p_1 = q^\star$, $p_2 = p^\star$ and $p_3 = 1 - q^\star - p^\star$ and $\nu_1 = \mathcal{B}(1/2 + \Delta)$ and $\nu_2 = \mathcal{B}(1/2)$ and $\nu_3 = \mathcal{B}(1/2 - \Delta)$.

where we assume that $p^\star, q^\star$ follow the assumptions of Theorem 33. For the reservoir $\mathbf{R}_0$, optimal arms make up a proportion $p^\star$ of the reservoir, and following the assumption of Theorem 33, the learner will suffer a low regret on $\mathbf{R}_0$. Our goal is to then show that as a result of this assumption, the learner must suffer a high regret on $\mathbf{R}_1$.

Note that, on reservoir $\mathbf{R}_0$ arms with mean $1/2$ are optimal, thus for the learner to suffer a small regret on $\mathbf{R}_0$ they must pull arms with mean $1/2$ *many times*. On reservoir $\mathbf{R}_1$, however, arms with mean $1/2$ are sub optimal and for the learner to suffer a small regret on $\mathbf{R}_1$ the must pull arms with mean $1/2$ *few times*. To prove Theorem 33, we will show that the reservoirs $\mathbf{R}_0$ and $\mathbf{R}_1$ are hard for the learner to distinguish between, and thus, if with high probability, on $\mathbf{R}_0$ the learner will pull arms with mean $1/2$ many times, then they will also, with high probability pull arms with mean $1/2$ many times on $\mathbf{R}_1$. Thus, if the leaner suffers a small regret on $\mathbf{R}_0$, as a consequence they must then suffer a large regret on $\mathbf{R}_1$.

The question remains as to how we will show $\mathbf{R}_0$ and $\mathbf{R}_1$ are hard for the learner to distinguish between. To do this in practice, we will upper bound the divergence between distributions on the samples the learner is able to collect from each of the reservoirs, $\mathbf{R}_0$ and $\mathbf{R}_1$. Consider the samples the learner collects on $\mathbf{R}_0$. As the learner suffers low regret on $\mathbf{R}_0$, we can expect a large number of these samples to be on arms with mean $1/2$. Essentially the distribution on these samples will not differ to the case where the learner draws from $\mathbf{R}_1$, as both $\mathbf{R}_0$, $\mathbf{R}_1$ have an equal proportion of arms $a : \mu_a = 1/2$. This is already a significant step in upper bounding the divergence between the two sampling distributions on $\mathbf{R}_0$ and $\mathbf{R}_1$.

A second step is to instead consider the samples the learner draws from arms $a : \mu_a \neq 1/2$. Conditional on $\mu_a \neq 1/2$, under reservoir $\mathbf{R}_0$ the probability $\mu_a = 1/2 - \Delta$ is 1, under reservoir $\mathbf{R}_1$ the probability $\mu_a \neq 1/2 - \Delta$ is $\frac{q^\star}{1-p^\star}$. For this reason we will be able to recover an additional multiplicative $\frac{q^\star}{1-p^\star}$ term in our upper bound on the divergence between the two sampling distributions on $\mathbf{R}_1$ and $\mathbf{R}_0$, see our coupling argument and Equation (4.7) for details.

It remains to consider the event where the learner pulls arms with mean $1/2$ more than $T/2$ times. The probability of this event on reservoir $\mathbf{R}_0$ must be large, following the assumptions of Theorem 33. Then, using a standard information theoretic inequality, the Bretagnolle-Huber inequlaity and our previous upper bound on the divergence between the two sampling distributions on $\mathbf{R}_1$ and $\mathbf{R}_0$, we can show that the probability of said event on $\mathbf{R}_1$, must also be large, providing the result.

Proof of Theorem 33. Consider $\Delta \in (0, 1/4)$ and the following two definitions of two reservoir distributions:

- The reservoir distribution $\mathbf{R}_0$ characterised by $p_1 = p^\star$ and $p_2 = 1 - p^\star$ and $\nu_1 = \mathcal{B}(1/2)$ and $\nu_2 = \mathcal{B}(1/2 - \Delta)$.
- The reservoir distribution $\mathbf{R}_1$ characterised by $p_1 = q^\star$, $p_2 = p^\star$ and $p_3 = 1 - q^\star - p^\star$ and $\nu_1 = \mathcal{B}(1/2 + \Delta)$ and $\nu_2 = \mathcal{B}(1/2)$ and $\nu_3 = \mathcal{B}(1/2 - \Delta)$.

Note that the Bernoulli distribution is completely characterised by its mean and so we can use the mean to characterise the distribution. Let $\tilde{\mu} = (\tilde{\mu}_j)_{j \le T}$ be T i.i.d. means corresponding to T i.i.d. distributions sampled according to the reservoir distribution $\mathbf{R}_1$. Note that $\tilde{\mu}_j \in \{1/2 - \Delta, 1/2, 1/2 + \Delta\}$. Write also $\tilde{\mu}' = (\tilde{\mu}'_j)_{j \le T}$ for the vector of means such that $\tilde{\mu}'_j = \tilde{\mu}_j$ if $\tilde{\mu}'_j \in \{1/2 - \Delta, 1/2\}$, and $\tilde{\mu}'_j = 1/2 - \Delta$ otherwise. Note that then, we have that $(\tilde{\mu}'_j)_{j \le T}$ are T i.i.d. means corresponding to T i.i.d. distributions sampled according to the reservoir distribution $\mathbf{R}_0$, by definition of $\mathbf{R}_0$. Write $\mathbb{E}_{\mathbf{R}_1}$ for the expectation according to the distribution of $\tilde{\mu}$, i.e. according to $\mathbf{R}_1^{\otimes T}$, and $\mathbb{E}_{\mathbf{R}_0}$ for the expectation according to the distribution of $\tilde{\mu}'$, i.e. according to $\mathbf{R}_0^{\otimes T}$.

Consider an algorithm $\mathfrak{A}$ and a bandit problem involving Bernoulli distributions characterised by a vector of means $m = (m_j)_{j \le T}$. Write $\mathbb{P}_m^{\mathfrak{A}}$ for the distribution of the samples obtained by the algorithm run on this problem, and $\mathbb{E}_m^{\mathfrak{A}}$ the associated expectation. Consider now another Bernoulli bandit problem characterised by the means $m' = (m'_j)_{j \le T}$. We have because of the chain rule

$$\mathrm{KL}(\mathbb{P}_{m'}^{\mathfrak{A}}, \mathbb{P}_m^{\mathfrak{A}}) = \sum_{j \le T} \mathbb{E}_{m'}^{\mathfrak{A}}[T_j]\, \mathrm{kl}(m'_j, m_j) ,$$

where $\mathbb{E}_{m'}^{\mathfrak{A}}$ is the expectation according to problem m' on which algorithm $\mathfrak{A}$ is used, and where T_j is the number of times arm j is sampled at time T.

From our assumption on $\mathfrak{A}$ we have that $\mathbb{E}_{\mathbf{R}_0}[R(T)] \le \frac{c \log(T)}{p^\star \Delta}$. Now, we can obtain

$$\begin{aligned}
\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}_{\tilde{\mu}'}^{\mathfrak{A}}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}_{\tilde{\mu}}^{\mathfrak{A}}) &= \mathrm{KL}(\mathbb{E}_{\mathbf{R}_1}\mathbb{P}_{\tilde{\mu}'}^{\mathfrak{A}}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}_{\tilde{\mu}}^{\mathfrak{A}}) \\
&\le \mathbb{E}_{\mathbf{R}_1}\left[\mathrm{KL}(\mathbb{P}_{\tilde{\mu}'}^{\mathfrak{A}}, \mathbb{P}_{\tilde{\mu}}^{\mathfrak{A}})\right] = \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \le T} \mathbb{E}_{\tilde{\mu}'}^{\mathfrak{A}}[T_j]\, \mathrm{kl}(\tilde{\mu}'_j, \tilde{\mu}_j)\right] \\
&\le \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \le T} \mathbb{E}_{\tilde{\mu}'}^{\mathfrak{A}}[T_j] \frac{\Delta^2}{16} \mathbf{1}\{\tilde{\mu}_j = 1/2 + \Delta\}\right] \\
&= \mathbb{E}_{\mathbf{R}_0}\left[\sum_{j \le T} \mathbb{E}_{\tilde{\mu}'}^{\mathfrak{A}}[T_j] \frac{\Delta^2}{16} \mathbf{1}\{\tilde{\mu}'_j = 1/2 - \Delta\} \frac{q^\star}{1 - p^\star}\right] \\
&= \frac{q^\star \Delta}{8} \mathbb{E}_{\mathbf{R}_0}[R(T)] \le \frac{c q^\star}{8 p^\star} \log(T) \le \frac{1}{2} \log(T) , \qquad (4.7)
\end{aligned}$$

where the penultimate equality follows since by definition of $\mathbf{R}_0, \mathbf{R}_1$, conditionally on $\tilde{\mu}'_j = 1/2 - \Delta$, the probability that $\tilde{\mu}_j = 1/2 + \Delta$ is $\frac{q^\star}{1 - p^\star} \le 2q^\star$, and otherwise it is 0. And where the final inequality comes from our assumption $p^\star > \frac{c q^\star}{4}$.

Consider the event,

$$E := \left\{ \sum_{j \le T} T_j \mathbf{1}\{\tilde{\mu}'_j = 1/2\} > T/2 \right\} .$$

Note that on $\mathbf{R}_0$, we have $\mu^* = \frac{1}{2}$. Thus, on $\mathbf{R}_0$ the event E^C will signify a regret greater than $\frac{T\Delta}{2}$, similarly on $\mathbf{R}_1$ the event E signifies a regret greater than $\frac{T\Delta}{2}$. Thus,

$$\mathbb{E}_{\mathbf{R}_0} R(T) \ge \mathbb{E}_{\mathbf{R}_0} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) \frac{T\Delta}{2} , \qquad \mathbb{E}_{\mathbf{R}_1} R(T) \ge \mathbb{E}_{\mathbf{R}_1} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) \frac{T\Delta}{2} \tag{4.8}$$

Now from our assumption upon $\mathfrak{A}$ we have that $\mathbb{E}_{\mathbf{R}_0} R(T) \le \frac{c\log(T)}{p^\star \Delta}$, therefore Equation (4.8) leads to,

$$\mathbb{E}_{\mathbf{R}_0} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) \le \frac{c\log(T)}{p^\star \Delta} \times \frac{2}{T\Delta} . \tag{4.9}$$

Now, using the Bretagnolle-Huber's inequality (see Theorem 14.2 by Lattimore and Szepesvári [63]) in combination with (4.7) we obtain

$$\begin{aligned} \mathbb{E}_{\mathbf{R}_0} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) + \mathbb{E}_{\mathbf{R}_1} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) &\ge \frac{1}{2} \exp\Big(- \mathrm{KL}(\mathbb{E}_{\mathbf{R}_0} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \Big) \\ &\ge \frac{1}{2\sqrt{T}} . \end{aligned}$$

This result in combination with Equation (4.9) gives the following,

$$\mathbb{E}_{\mathbf{R}_1} \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) \ge \frac{1}{2\sqrt{T}} - \frac{2c\log(T)}{p^\star T \Delta^2} \ge \frac{1}{4\sqrt{T}} , \tag{4.10}$$

where the final inequality comes from our assumption $T \ge 4\Big(\frac{c\log(T)}{p^\star \Delta^2}\Big)^2$. Finally our result follows from combination of Equation (4.8) and Equation (4.10).

□

4.7 Best-arm identification proofs

4.7.1 Upper Bound

Proof of Theorem 34. **Proof-specific notations and preliminary considerations.** At round i, write $K_i = |\mathcal{A}_i|$ and write p_i for the proportion of optimal arms in $\mathcal{A}_i$, namely

$$p_i = |\mathcal{A}_i \cap \mathcal{A}^*| / |\mathcal{A}_i| .$$

We also write M_i for the number of optimal arms in $\mathcal{A}_i$ such that $\hat{\mu}_i(a) \ge \mu^* - \Delta/2$, namely

$$M_i = \big|\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \ge \mu^* - \Delta/2\}\big| ,$$

and N_i for the number of sub-optimal arms in $\mathcal{A}_i$ such that $\hat{\mu}_i(a) \ge \mu^* - \Delta/2$, namely

$$N_i = \big|\{a \in \mathcal{A}_i \cap \mathcal{A}_{sub} : \hat{\mu}_i(a) \ge \mu^* - \Delta/2\}\big| .$$

Note that by definition

$$K_{i+1} = \left(1 \vee \left\lfloor \frac{K_i}{2} \right\rfloor\right) + \left\lfloor \frac{K_i}{4} \right\rfloor .$$

Therefore the following bounds holds

$$\left(\left(\frac{3}{4}\right)^i K_1\right) \vee 1 \geq K_i \geq \left(\frac{1}{2}\right)^i K_1 - 4\,. \tag{4.11}$$

We write I for the smallest index i such that $K_i = 1$ and will not investigate what happens at rounds $i > I$. By the upper bound (4.11) on K_i it holds $I \leq \log_{4/3}(K_1) \leq \log_{4/3}(T)$. Note that since $\log_{4/3}(T) = \bar{c}\log T$, the algorithm terminates with a set containing just one arm.

Step 1: Introduction of high-probability events of interest. We define the constant

$$c = \frac{\bar{c}}{10}.$$

We define j^* as the largest j smaller than or equal to I such that

$$K_j \geq cT\Delta^2/(2\log T).$$

Note that such j^* exists since $K_1 \geq \bar{c}T/(2\log T)$, and since $K_I = 1$. We prove below the following upper bound on j^*. Take any round i. Note that for any k, conditionally on $\mathcal{A}_i$, by Hoeffding's inequality, for any $a \in \mathcal{A}_i$

$$\mathbb{P}\Big(\big|\hat{\mu}_i(a) - \mu_i(a)\big| \geq \Delta/2 \Big| \mathcal{A}_i\Big) \leq 2\exp(-\Delta^2 t_i/2) = q_i, \tag{4.12}$$

where $\mu_i(a)$ is the true mean associated with arm a. We now state the following technical lemma proved below.

Lemma 24. *Assume that $p^\star \leq 1/2$, and consider $I \geq i \geq j^*$. Under the assumptions of the theorem, we have*

$$q_i^{-1/2} \geq 200 \geq e^2 - 1\,, \tag{4.13}$$

$$\Delta^2 t_i/4 \geq \log 2\,. \tag{4.14}$$

We define for $i \geq j^*$ and $\bar{p}_i := \left(\frac{p^\star}{6}(5/4)^{i-j^*} \wedge (1/2)\right)$, the event

$$\xi_i = \{p_i > \bar{p}_i\}.$$

Consider from now on $i \geq j^*$.

Step 2: Lower bound on M_i conditional to ξ_i. We have by definition of M_i:

$$M_i = \sum_{a \in \mathcal{A}_i \cap \mathcal{A}^*} \mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\},$$

where by Equation (4.12), and conditionally on $\mathcal{A}_i$, the $\mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\}$ are independent and dominate stochastically $\mathcal{B}(1 - q_i)$, for any $a \in \mathcal{A}_i \cap \mathcal{A}^*$. And so conditionally on $\mathcal{A}_i$, we have that M_i stochastically dominates $\mathcal{B}(K_i p_i, 1 - q_i)$. And so by Chernoff's inequality, for any $x \geq \sqrt{q_i}$:

$$\begin{aligned}\mathbb{P}(M_i - p_iK_i(1-q_i) \leq -xp_iK_i|\mathcal{A}_i) &\leq \left[\frac{e^{x/q_i}}{(1+x/q_i)^{1+x/q_i}}\right]^{K_ip_iq_i}\\ &\leq \exp\Big[xK_ip_i - \log(1+x/q_i)(K_ip_iq_i + xK_ip_i)\Big]\\ &\leq (1+x/q_i)^{-xK_ip_i/2}.\end{aligned}$$

as for $i > j^*$ we have $\log(1+x/q_i) > 2$, see Lemma 24.

So that for $x \geq \sqrt{q_i}$

$$\mathbb{P}(M_i \leq K_ip_i(1-2x)|\mathcal{A}_i) \leq \exp\Big(-x\Delta^2 t_iK_ip_i/16\Big),$$

since $\log(q_i^{-1}) = \Delta^2 t_i/2 - \log 2 \geq \Delta^2 t_i/4$ for $I \geq i \geq j^*$ - see Lemma 24.

And so since $p_i \geq \frac{p^\star}{6}$ on ξ_i

$$\mathbb{P}(M_i \leq p_iK_i(1-2x)|\xi_i) \leq \exp\big(-\bar{c}' xp^\star\Delta^2 T/\log T\big) := u. \tag{4.15}$$

where $\bar{c}' = \bar{c}/96$ and recalling $t_i = \lfloor \bar{c}T/\big(K_i\log(T)\big)\rfloor$.

Step 3: Upper bound on N_i conditional to ξ_i. We have by definition of N_i:

$$N_i = \sum_{a\in\mathcal{A}_i\cap\mathcal{A}_{sub}} \mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \Delta/2\},$$

where by Equation (4.12), and conditionally on $\mathcal{A}_i$, the $\mathbf{1}\{\hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are independent and are stochastically dominated by $\mathcal{B}(q_i)$, for any $a \in \mathcal{A}_i \cap \mathcal{A}_{sub}$. And so conditionally on $\mathcal{A}_i$, we have that N_i is stochastically dominated by $\mathcal{B}(K_i, q_i)$. And so by Chernoff's inequality for any $x \geq 2$:

$$\mathbb{P}(N_i - K_iq_i \geq xK_i|\xi_i) \leq \left[\frac{e^{x/q_i}}{(1+x/q_i)^{1+x/q_i}}\right]^{K_iq_i} \leq (1+x/q_i)^{-xK_i/2},$$

similar to Step 2.

So that for $x \geq \sqrt{q_i}$

$$\mathbb{P}(N_i \geq 2K_ix|\mathcal{A}_i) \leq \exp\Big(-x\Delta^2 t_iK_i/16\Big),$$

as in Step 2.

And so similar to in Step 2:

$$\mathbb{P}(N_i \geq 2xK_i|\xi_i) \leq \exp\big(-\bar{c}' x\Delta^2 T/\log T\big) \leq u. \tag{4.16}$$

Step 4: Bound on the probability of ξ_i and conclusion. First we have – since we add $K_{j^*-1}/4 = K_{j^*}/3$ fresh arms to the set $\mathcal{A}_{j^*}$ - that

$$\left\{\left|\sum_{a\in\mathcal{A}_{j^*}} \mathbf{1}\{a \in \mathcal{A}^*\} - \frac{1}{3}p^\star K_{j^*}\right| \leq \frac{1}{6}p^\star K_{j^*}\right\} \subset \xi_{j^*},$$

where it holds that $\mathbf{1}\{a \in \mathcal{A}^*\} \sim \mathcal{B}(p^*)$ for the fresh arms and $|\mathcal{A}_{j^*}| = K_{j^*}$. And so by Chernoff's inequality:

$$\mathbb{P}(\xi_{j^*}) \geq 1 - 2\exp(-p^\star K_{j^*}/10) \geq 1 - 2\exp\left(-c\frac{p^\star T\Delta^2}{20\log T}\right) =: 1 - v, \tag{4.17}$$

by definition of j^*.

Now consider $i > j^*$, let,

$$\xi_i' = \left\{p_{i+1} \geq \frac{5}{4}p_i \wedge \frac{1}{2}\right\}.$$

Lemma 25. *Assume that $2x \leq 1/100$. We have for $I \geq i > j^*$:*

$$\xi_i'' := \{M_i > p_i K_i(1-2x)\} \cap \{N_i < 2xK_i\} \subset \xi_i'.$$

Note also that

$$\mathbb{P}(\xi_i''|\xi_i) \geq 1 - 2u,$$

by Equations (4.15) and (4.16), so that by Lemma 25

$$\mathbb{P}(\xi_i'|\xi_i) \geq 1 - 2u. \tag{4.18}$$

By induction it holds that for any $1 \leq m \leq I - j^*$

$$\xi_{j^*} \cap \bigcap_{j^* < i \leq j^*+m} \xi_i' \subset \bigcap_{j^* \leq i \leq j^*+m} \xi_i,$$

so that by Equations (4.17) and (4.18)

$$\mathbb{P}\left(\bigcap_{j^* \leq i \leq j^*+m} \xi_i\right) \geq (1-v)(1-2u)^m \geq 1 - v - 2um.$$

In particular using the previous inequality for $m = I - j^*$ and since $I \leq \log T$ it holds

$$\mathbb{P}\left(\bigcap_{j^* \leq i \leq I} \xi_i\right) \geq 1 - v - 2u\log T.$$

Since $K_I = 1$, and since by definition of the ξ_i we know that on ξ_I we have that the only arm in $\mathcal{A}_I$ is optimal, this concludes the proof - taking $x = 1/200$, which is compatible with $x \geq \sqrt{q_i}$ as $q_i \leq 1/200^2$ by Lemma 24.

□

We prove now successively, Lemma 24, Lemma 25 used in the proof of Theorem 34.

Proof of Lemma 24. Note first that for $I \geq i \geq j^*$ we have

$$K_{i+1} = \lfloor K_i/2 \rfloor \vee 1 + \lfloor K_i/4 \rfloor \leq \frac{3K_i}{4} \vee 1.$$

So that for any $0 \leq m < I - j^*$ we have by definition of I as the first index such that $K_I = 1$

$$K_i \leq K_{j^*}(3/4)^m. \tag{4.19}$$

Also for any i such that $K_i \geq 4$

$$K_{i+1} \geq K_i/2,$$

and for any i such that $K_i < 4$, we have

$$K_{i+1} = 1,$$

so that for any $0 \leq m < I - j^*$ we have

$$K_i \geq K_{j^*}(i/2)^m.$$

Inequality (4.13): We therefore have for $I > i \geq j^*$ and by Equation (4.19)

$$q_i^{-1/2} = 2^{-1/2}\exp(\Delta^2 t_i/4) \geq 2^{-1/2}\exp\left(\bar{c}\frac{\Delta^2 T}{2K_{j^*}\log(T)}\right),$$
$$\geq 2^{-1/2}\exp(10) \geq 200 \geq e^2 - 1$$

Inequality (4.14): We have,

$$q_i = \exp(-\Delta^2 t_i/2),$$

thus by inequality (4.13) we have

$$\exp(\Delta^2 t_i/4) \geq \sqrt{2}(e^2 - 1),$$

so that

$$\Delta^2 t_i/4 \geq \log 2.$$

□

Proof of Lemma 25. Let i such that $I \geq i > j^*$. Note that on ξ_i'', we have $M_i > 0$ so that $p_i > 0$.

First case: $0 < p_i \leq 2/5$. Assume first that $p_i \leq 2/5$. On ξ_i'' we have that

$$M_i > p_i K_i(1 - 2x),$$

and

$$N_i < 2K_i x,$$

so that

$$M_i + N_i < p_i K_i + 2K_i x \leq (2/5)K_i + K_i/100 \leq K_i/2.$$

since $2x \leq 1/100$ for $i \geq j^*$ - see Lemma 24. And so all M_i arms of $\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are going to be in $\mathcal{A}_{i+1}$. This implies – as in this case $K_i \geq 2$ otherwise we cannot have $0 < p_i \leq 2/5$ – that

$$p_{i+1} \geq \frac{M_i}{K_{i+1}} = \frac{M_i}{1 \vee \lfloor K_i/2 \rfloor + \lfloor K_i/4 \rfloor} \geq \frac{4}{3}(1 - 2x)p_i > \frac{5}{4}p_i,$$

as $2x \leq 1/100$.

Second case: $p_i > 2/5$. Assume now that $p_i > 2/5$. On ξ_i'' we have that

$$M_i > p_i K_i(1-2x) \geq \frac{198}{500} K_i,$$

and

$$N_i < 2K_i x \leq K_i/100,$$

since $2x \leq 1/100$ for $I \geq i > j^*$ – see Lemma 24. Since $198/500 + 1/100 = 203/500 < 1/2$ this implies that at least $\frac{199}{500}K_i$ from the arms in $\{a \in \mathcal{A}_i \cap \mathcal{A}^* : \hat{\mu}_i(a) \geq \mu^* - \bar{\Delta}/2\}$ are going to be in $\mathcal{A}_{i+1}$. So that

$$p_{i+1} \geq \frac{M_i}{K_{i+1}} = \frac{M_i}{1 \vee \lfloor K_i/2 \rfloor + \lfloor K_i/4 \rfloor} \geq \frac{4}{3} \times \frac{198}{500} = \frac{66}{125} > 1/2.$$

This concludes the proof. □

4.7.2 Lower Bound

Proof of Theorem 35. We consider a similar setting to that in the proof of Theorem 33 although with a slightly different construction of $\mathbf{R}_0, \mathbf{R}_1$.

Consider the following two reservoir distributions:

- The reservoir distribution $\mathbf{R}_0$ characterised by $p_1 = p^\star$ and $p_2 = 1 - p^\star$ and $\nu_1 = \mathcal{B}(1/2)$ and $\nu_2 = \mathcal{B}(1/2 - \Delta)$.
- The reservoir distribution $\mathbf{R}_1$ characterised by $p_1 = p^\star$ and $p_2 = p^\star$ and $p_3 = 1 - 2p^\star$ and $\nu_1 = \mathcal{B}(1/2 + \Delta)$ and $\nu_2 = \mathcal{B}(1/2)$ and $\nu_3 = \mathcal{B}(1/2 - \Delta)$.

We define $\tilde{\mu}, \tilde{\mu}'$, and associated expectations and probabilities as in the proof of Theorem 33. Consider also any algorithm $\mathfrak{A}$. We have by similar calculations as Equation (4.7) the following upper bound on the KL divergence

$$\begin{aligned}
\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) &= \mathrm{KL}(\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\mathrm{KL}(\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}})\right] = \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j]\,\mathrm{kl}(\tilde{\mu}'_j, \tilde{\mu}_j)\right] \\
&\leq \mathbb{E}_{\mathbf{R}_1}\left[\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j] \frac{\Delta^2}{16} \mathbf{1}\{\tilde{\mu}_j = 1/2 + \Delta\}\right] \\
&= \mathbb{E}_{\mathbf{R}_0}\left[\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j] \frac{\Delta^2}{16} \mathbf{1}\{\tilde{\mu}'_j = 1/2 - \Delta\} \frac{p^\star}{1 - p^\star}\right],
\end{aligned} \tag{4.20}$$

since by definition of $\mathbf{R}_0, \mathbf{R}_1$, conditionally on $\tilde{\mu}'_j = 1/2 - \Delta$, the probability that $\tilde{\mu}_j = 1/2 + \Delta$ is $\frac{p^\star}{1-p^\star}$, and otherwise it is 0.

By Equation (4.20), and since $\sum_{j \leq T} \mathbb{E}^{\mathfrak{A}}_{\tilde{\mu}'}[T_j] = T$, we have

$$\mathrm{KL}(\mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}, \mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}) \leq T \frac{\Delta^2}{16} \frac{p^\star}{1 - p^\star}\,. \tag{4.21}$$

Let us write $\hat{a}_T$ for the arm that the algorithm $\mathfrak{A}$ recommends. Set

$$E = \{\tilde{\mu}_{\hat{a}_T} = 1/2\}\,.$$

Note that on E, we make a mistake in prediction for $\tilde{\mu}$, and that on E^C, we make a mistake in prediction for $\tilde{\mu}'$. Thus, via Bretagnolle-Huber's inequality (see Theorem 14.2 by Lattimore and Szepesvári [63]), and Equation (4.21), we have

$$\mathbb{E}_{\mathbf{R}_1}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}}(E) + \mathbb{E}_{\mathbf{R}_0}\mathbb{P}^{\mathfrak{A}}_{\tilde{\mu}'}(E^C) \geq \frac{1}{2}\exp\left(-T\frac{\Delta^2}{16}\frac{p^\star}{1-p^\star}\right).$$

This concludes the proof by definition of E. □

4.8 Technical lemmas

Lemma 26. *(Chernoff bound) Let $X_1, \ldots, \mathcal{X}_n \sim \mathcal{B}\mathrm{er}(p)$ be n samples from a Bernoulli distribution and $S_n = \sum_{k=1}^n X_n$ their sum. Then for all $\gamma \in [0,1]$ it holds*

$$\mathbb{P}\left(\frac{S_n}{n} \leq (1-\gamma)p\right) \leq e^{-\frac{\gamma^2}{4}np},$$
$$\mathbb{P}\left(\frac{S_n}{n} \geq (1+\gamma)p\right) \leq e^{-\frac{\gamma^2}{4}np}.$$

Proof. We prove the first inequality; the second one is similar. If $(1-\gamma)p < 0$ or $\gamma = 0$ the inequality is trivially true. Else, because of Chernoff's inequality, we have

$$\mathbb{P}\left(\frac{S_n}{n} \leq (1-\gamma)p\right) \leq e^{-n\,\mathrm{kl}\left((1-\gamma)p,p\right)}.$$

It remains to remark to conclude that

$$\mathrm{kl}\big((1-\gamma)p, p\big) \geq \frac{\gamma^2}{2}p,$$

where we used the refined Pinsker inequality from Garivier, Ménard, and Stoltz [40], for $0 \leq x < y \leq 1$,

$$\mathrm{kl}(y,x) \geq \frac{1}{2\max_{x \leq q \leq y} q(1-q)}(x-y)^2 \geq \frac{1}{2y}(x-y)^2.$$

For the second inequality we use

$$\mathrm{kl}\big((1+\gamma)p, p\big) \geq \frac{1}{2(1+\gamma)p}\gamma^2 p^2 \geq \frac{\gamma^2}{4}p.$$

□

Lemma 27. *Let $A, B, C \geq 0$ be constants such that $A \geq C$, then for $n_0 = \inf\{n \geq 1 : A + B\log(n) \leq nC\}$ we have*

$$n \leq \frac{A + B\log\big((2(B^2+AC)/C^2\big)}{C} + 1.$$

Proof. First let $x_0 \geq 1$ be such that $A + B\log(x_0) = Cx_0$. It exists since $A + B\log(x)/x \to 0$ if $x \to \infty$ and since $A \geq C$. In particular, because of the definition of n_0 we have $x_0 \leq n_0 \leq x_0 + 1$. Then note that $A + B\sqrt{x_0} \leq Cx_0$. Thus $\sqrt{x_0}$ is smaller than the largest roots of the polynomial $Cy^2 - By - A$. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and

$(a+b)^2 \leq 2(a^2+b^2)$ we obtain

$$\begin{aligned} x_0 &\leq \left(\frac{B+\sqrt{B^2+4AC}}{2C}\right)^2 \\ &\leq 2\frac{B^2+AC}{C^2}\,. \end{aligned}$$

Inserting the previous inequality in the definition of x_0 and using $n_0 \leq x_0 + 1$ allows us to conclude

$$n_0 \leq \frac{A + B\log\big(2(B^2+AC)/C^2\big)}{C} + 1\,.$$

□

4.9 Experiments

In this section we conduct preliminary experiments for the cumulative regret and best-arm identification setting.

Cumulative regret For the cumulative regret we compare Sampling-UCB (with $\gamma = 0.5$) with the QRM1 algorithm by [21] and SR algorithm by [89]. We arbitrarily[4] choose the following reservoir: the arms are distributed according to a Bernoulli distribution with possible means [0.5, 0.8] sampled with probabilities [0.8, 0.2]. We remark that the SR algorithm and Sampling-UCB are very similar, they both sample approximately $\log(T)/p^\star$ arms and run a regret minimizer algorithm on this set of arms. The only difference is that the SR algorithm relies on the MOSS algorithm. Whereas the QRM1 algorithm proceeds by progressively adding new arms. In particular this algorithm is anytime. In Figure 4.1 we compare the cumulative regret of the different algorithms for a fixed horizon $T = 20000$. We observe that Sampling-UCB behaves similarly to SR and that QRM1 performs slightly worst (maybe because of the adaptation to T). We also check that all algorithms have a regret that is logarithmic with the horizon as expected. To this aim, in Figure 4.2, we plot the cumulative regret (for the same reservoir) for all horizons $T \in \{100, 200, \ldots, 10000\}$.

Best-arm identification For best arm identification we compare our algorithm with the BUCB algorithm by [56]. In Figure 4.3 we compare the performance of the algorithms across varying Δ for a fixed $T = 1000$. That is, we consider reservoirs of the form $[0.2, \Delta, 1]$ for $\Delta \in (0.01 \times i)_{i\in[79]}$ with probabilities $[0.29, 0.69, 0.02]$. The BUCB algorithm presents an issue as it is designed for the fixed confidence regime the algorithm takes δ as a parameter. We set δ equal to an arbitrarily low constant. The BUCB algorithm works by opening successively large brackets of arms, however as they do not provide results in high probability, only in expectation, they can draw significantly less arms from the reservoir. The performance of Elimination seems favourable compared to BUCB, however, one may be able to improve the performance of BUCB with parameter tuning.

[4] Which is not very important, since we evaluate the algorithms from a problem-dependent point of view

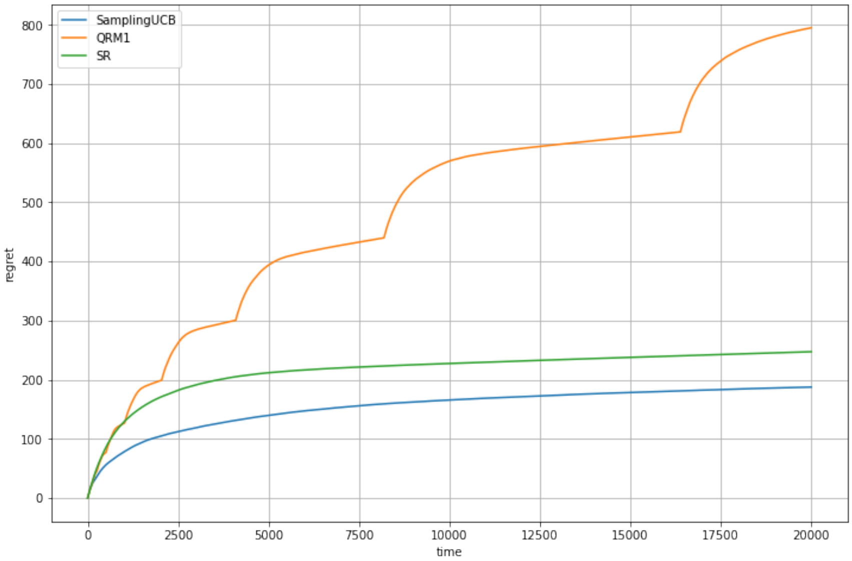

FIGURE 4.1: Cumulative regret in function of the time estimated by 100 Monte-Carlo simulations.

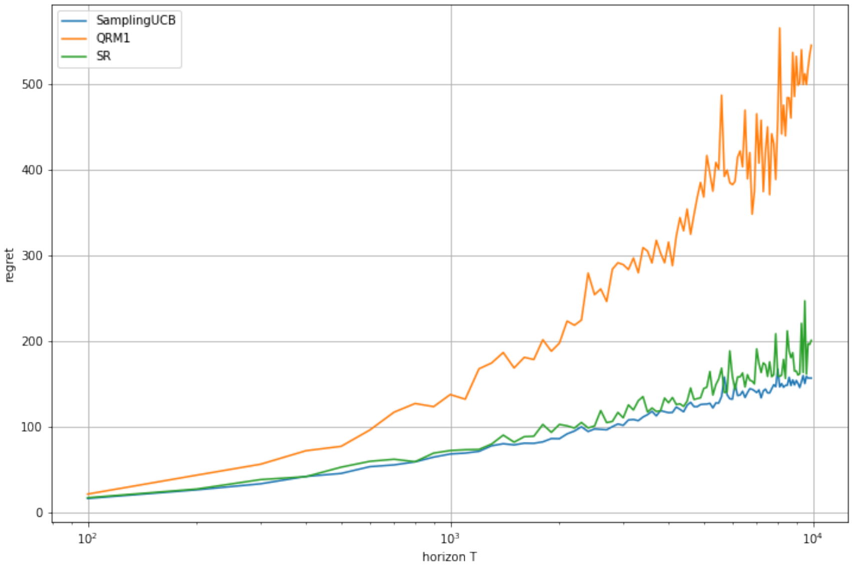

FIGURE 4.2: Cumulative regret in function of the horizon $T \in \{100, 200, \ldots, 10000\}$ estimated by 100 Monte-Carlo simulations.

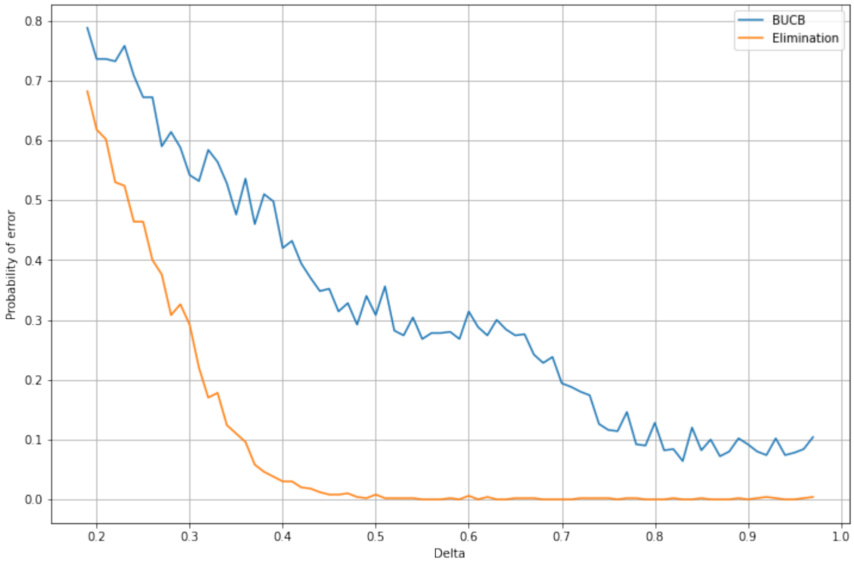

FIGURE 4.3: Probability of error for best arm identification across varying Δ using 500 Monte-Carlo simulations.

Bibliography

[1] 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*. 2011, pp. 1035–1043.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. “Best arm identification in multi-armed bandits.” In: *COLT*. Citeseer. 2010, pp. 41–53.

[3] Jean-Yves Audibert, Sébastien Bubeck, et al. “Minimax Policies for Adversarial and Stochastic Bandits.” In: *COLT*. Vol. 7. 2009, pp. 1–122.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine learning* 47.2 (2002), pp. 235–256.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. “The nonstochastic multiarmed bandit problem”. In: *SIAM journal on computing* 32.1 (2002), pp. 48–77.

[6] Peter Auer and Ronald Ortner. “UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem”. In: *Periodica Mathematica Hungarica* 61.1-2 (2010), pp. 55–65.

[7] Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. “Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence”. In: *Algorithmic Learning Theory*. 2018, pp. 3–24.

[8] 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*. IEEE. 2008, pp. 221–230.

[9] Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. “Bandit problems with infinitely many arms”. In: *The Annals of Statistics* (1997), pp. 2103–2116.

[10] Thomas Bonald and Alexandre Proutiere. “Two-target algorithms for infinite-armed bandits with bernoulli rewards”. In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 2184–2192.

[11] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. “Pure exploration in finitely-armed and continuous-armed bandits”. In: *Theoretical Computer Science* 412.19 (2011), pp. 1832–1852.

[12] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. “Pure exploration in multi-armed bandits problems”. In: *International conference on Algorithmic learning theory*. Springer. 2009, pp. 23–37.

[13] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. “X-Armed Bandits.” In: *Journal of Machine Learning Research* 12.5 (2011).

[14] Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. “Multiple identifications in multi-armed bandits”. In: *International Conference on Machine Learning*. PMLR. 2013, pp. 258–265.

[15] Apostolos N Burnetas and Michael N Katehakis. "Optimal adaptive policies for sequential allocation problems". In: *Advances in Applied Mathematics* 17.2 (1996), pp. 122–142.

[16] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. "Kullback–leibler upper confidence bounds for optimal sequential allocation". In: *The Annals of Statistics* 41.3 (2013), pp. 1516–1541.

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

[18] Alexandra Carpentier and Michal Valko. "Simple regret for infinitely many armed bandits". In: *International Conference on Machine Learning*. 2015, pp. 1133–1141.

[19] Karthekeyan Chandrasekaran and Richard Karp. "Finding a most biased coin with fewest flips". In: *Conference on Learning Theory*. 2014, pp. 394–407.

[20] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. "PAC identification of a bandit arm relative to a reward quantile". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[21] Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. "Quantile-Regret Minimisation in Infinitely Many-Armed Bandits." In: *UAI*. 2018, pp. 425–434.

[22] Lijie Chen and Jian Li. "On the optimal sample complexity for best arm identification". In: *arXiv preprint arXiv:1511.03774* (2015).

[23] 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* 27 (2014), pp. 379–387.

[24] 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*. 2016, pp. 1659–1667.

[25] James Cheshire, Pierre Menard, and Alexandra Carpentier. "Problem Dependent View on Structured Thresholding Bandit Problems". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1846–1854.

[26] James Cheshire, Pierre Ménard, and Alexandra Carpentier. "The Influence of Shape Constraints on the Thresholding Bandit Problem". In: *arXiv preprint arXiv:2006.10006* (2020).

[27] Richard Combes and Alexandre Proutiere. "Unimodal bandits: Regret lower bounds and optimal algorithms". In: *International Conference on Machine Learning*. 2014, pp. 521–529.

[28] Richard Combes and Alexandre Proutiere. "Unimodal bandits without smoothness". In: *arXiv preprint arXiv:1406.7447* (2014).

[29] Yahel David and Nahum Shimkin. "Infinitely many-armed bandits with unknown value distribution". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2014, pp. 307–322.

[30] 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*. ACM. 2016, pp. 519–532.

[31] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. “Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems”. In: *Journal of machine learning research* 7.Jun (2006), pp. 1079–1105.

[32] 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*. Springer. 2002, pp. 255–270.

[33] Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. “Computing with noisy information”. In: *SIAM Journal on Computing* 23.5 (1994), pp. 1001–1018.

[34] 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*. 2012, pp. 3212–3220.

[35] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. “Multi-bandit best arm identification”. In: (2011).

[36] Aurélien Garivier and Olivier Cappé. “The KL-UCB algorithm for bounded stochastic bandits and beyond”. In: *Proceedings of the 24th annual conference on learning theory*. JMLR Workshop and Conference Proceedings. 2011, pp. 359–376.

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

[38] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. “On explore-then-commit strategies”. In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 784–792.

[39] Aurélien Garivier, Pierre Ménard, Laurent Rossi, and Pierre Menard. “Thresholding bandit for dose-ranging: The impact of monotonicity”. In: *arXiv preprint arXiv:1711.04454* (2017).

[40] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. “Explore first, exploit next: The true shape of regret in bandit problems”. In: *Mathematics of Operations Research* 44.2 (2019), pp. 377–399.

[41] Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. “Fano’s inequality for random variables”. In: *arXiv preprint arXiv:1702.05985* (2017).

[42] Christophe Giraud, Yann Issartel, Luc Lehéricy, and Matthieu Lerasle. “Pair matching: When bandits meet stochastic block model”. In: *stat* 1050 (2019), p. 17.

[43] Jean-Bastien Grill, Michal Valko, and Rémi Munos. “Black-box optimization of noisy functions with unknown smoothness”. In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 667–675.

[44] Hédi Hadiji. “Polynomial cost of adaptation for x-armed bandits”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 1029–1038.

[45] Rianne de Heide, James Cheshire, Pierre Menard, and Alexandra Carpentier. “Bandits with many optimal arms”. In: *Advances in Neural Information Processing Systems*. 2021.

[46] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. “Stochastic blockmodels: First steps”. In: *Social networks* 5.2 (1983), pp. 109–137.

[47] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. “lil’ucb: An optimal exploration algorithm for multi-armed bandits”. In: *Conference on Learning Theory*. PMLR. 2014, pp. 423–439.

[48] Kevin Jamieson and Ameet Talwalkar. "Non-stochastic best arm identification and hyperparameter optimization". In: *Artificial intelligence and statistics*. PMLR. 2016, pp. 240–248.

[49] Kevin G Jamieson, Daniel Haas, and Benjamin Recht. "The power of adaptivity in identifying statistical alternatives". In: *Advances in Neural Information Processing Systems*. 2016, pp. 775–783.

[50] Sandeep Juneja and Subhashini Krishnasamy. "Sample complexity of partition identification using multi-armed bandits". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1824–1852.

[51] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. "PAC subset selection in stochastic multi-armed bandits." In: *ICML*. Vol. 12. 2012, pp. 655–662.

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

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

[54] Michael N Katehakis and Herbert Robbins. "Sequential choice from several populations." In: *Proceedings of the National Academy of Sciences of the United States of America* 92.19 (1995), p. 8584.

[55] Julian Katz-Samuels and Kevin Jamieson. "The True Sample Complexity of Identifying Good Arms". In: *arXiv preprint arXiv:1906.06594* (2019).

[56] Julian Katz-Samuels and Kevin Jamieson. "The True Sample Complexity of Identifying Good Arms". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1781–1791.

[57] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On the complexity of A/B testing". In: *Conference on Learning Theory*. PMLR. 2014, pp. 461–481.

[58] Emilie Kaufmann and Aurélien Garivier. "Learning the distribution with largest mean: two bandit frameworks". In: *ESAIM: Proceedings and surveys* 60 (2017), pp. 114–131.

[59] Emilie Kaufmann and Shivaram Kalyanakrishnan. "Information complexity in bandit subset selection". In: *Conference on Learning Theory*. PMLR. 2013, pp. 228–251.

[60] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. "Thompson sampling: An asymptotically optimal finite-time analysis". In: *International conference on algorithmic learning theory*. Springer. 2012, pp. 199–213.

[61] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. "Multi-armed bandits in metric spaces". In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 2008, pp. 681–690.

[62] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

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

[64] Can M Le, Elizaveta Levina, and Roman Vershynin. “Concentration of random graphs and application to community detection”. In: *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*. World Scientific. 2018, pp. 2925–2943.

[65] Jasper C.H. Lee and Paul Valiant. “Uncertainty about Uncertainty: Optimal Adaptive Algorithms for Estimating Mixtures of Unknown Coins*”. In: *ACM-SIAM* (2021).

[66] Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. “On zeroth-order stochastic convex optimization via random walks”. In: *arXiv preprint arXiv:1402.2667* (2014).

[67] Andrea Locatelli and Alexandra Carpentier. “Adaptivity to Smoothness in X-armed bandits”. In: *31st Annual Conference on Learning Theory* 75 (2018), pp. 1–30.

[68] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. “An optimal algorithm for the thresholding bandit problem”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1690–1698.

[69] Shie Mannor and John N Tsitsiklis. “The sample complexity of exploration in the multi-armed bandit problem”. In: *Journal of Machine Learning Research* 5.Jun (2004), pp. 623–648.

[70] Subhojyoti Mukherjee, Naveen Kolar Purushothama, Nandan Sudarsanam, and Balaraman Ravindran. “Thresholding Bandits with Augmented UCB”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (2017). DOI: 10.24963/ijcai.2017/350. URL: http://dx.doi.org/10.24963/ijcai.2017/350.

[71] 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*. AAAI Press. 2017, pp. 2515–2521.

[72] A. Nemirovski and D. Yudin. “Problem Complexity and Method Efficiency in Optimization”. In: *Wiley, New York* (1983).

[73] Robert Nowak. “The Geometry of Generalized Binary Search”. In: *arXic preprint arXiv:0910.4397* (2009).

[74] 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.

[75] Edward Paulson. “A Sequential Procedure for Selecting the Population with the Largest Mean from k Normal Populations”. In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 174–180. ISSN: 00034851. URL: http://www.jstor.org/stable/2238027.

[76] Herbert Robbins. “Some aspects of the sequential design of experiments”. In: *Bulletin of the American Mathematical Society* 58.5 (1952), pp. 527–535.

[77] Karl H. Schlag. “ELEVEN - Tests needed for a Recommendation”. In: 2006.

[78] Max Simchowitz, Kevin Jamieson, and Benjamin Recht. “The simulator: Understanding adaptive sampling in the moderate-confidence regime”. In: *arXiv preprint arXiv:1702.05186* (2017).

[79] Max Simchowitz, Kevin Jamieson, Jordan W Suchow, and Thomas L Griffiths. "Adaptive Sampling for Convex Regression". In: *arXiv preprint arXiv:1808.04523* (2018).

[80] William R Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294.

[81] William R. Thompson. "W. Thompson. On the likelihood that one unknown probability ex- ceeds another in view of the evidence of two samples." In: (1933).

[82] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. "Stochastic zeroth-order optimization in high dimensions". In: *arXiv preprint arXiv:1710.10551* (2017).

[83] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. "Algorithms for infinitely many-armed bandits". In: *Advances in Neural Information Processing Systems* 21 (2008), pp. 1729–1736.

[84] Yichong Xu, Xi Chen, Aarti Singh, and Artur Dubrawski. "Thresholding Bandit Problem with Both Duels and Pulls". In: *arXic preprint arXiv:1910.06368v1* (2019).

[85] Jia Yuan Yu and Shie Mannor. "Unimodal bandits". In: (2011).

[86] Jie Zhong, Yijun Huang, and Ji Liu. "Asynchronous parallel empirical variance guided algorithms for the thresholding bandit problem". In: *arXiv preprint arXiv:1704.04567* (2017).

[87] Jie Zhong, Yijun Huang, and Ji Liu. "Asynchronous Parallel Empirical Variance Guided Algorithms for the Thresholding Bandit Problem". In: *arXic preprint arXiv:1704.04567* (2017).

[88] Yuan Zhou, Xi Chen, and Jian Li. "Optimal PAC multiple arm identification with applications to crowdsourcing". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 217–225.

[89] Yinglun Zhu and Robert Nowak. "On Regret with Multiple Best Arms". In: *Advances in Neural Information Processing Systems*. 2020.