Unimodal Bandits with Continuous Arms: Order-optimal Regret without Smoothness

Richard Combes, Alexandre Proutière, Alexandre Fauquette

Unimodal Bandits without Smoothness

Richard Combes * and Alexandre Proutière †

March 9, 2015

Abstract

We consider stochastic bandit problems with a continuous set of arms and where the expected reward is a continuous and unimodal function of the arm. No further assumption is made regarding the smoothness and the structure of the expected reward function. For these problems, we propose the Stochastic Pentachotomy (SP) algorithm, and derive finite-time upper bounds on its regret and optimization error. In particular, we show that, for any expected reward function μ that behaves as $\mu(x) = \mu(x^\star) - C|x - x^\star|^\xi$ locally around its maximizer $x^\star$ for some $\xi, C > 0$, the SP algorithm is order-optimal. Namely its regret and optimization error scale as $O(\sqrt{T\log(T)})$ and $O(\sqrt{\log(T)/T})$, respectively, when the time horizon T grows large. These scalings are achieved without the knowledge of ξ and C. Our algorithm is based on asymptotically optimal sequential statistical tests used to successively trim an interval that contains the best arm with high probability. To our knowledge, the SP algorithm constitutes the first sequential arm selection rule that achieves a regret and optimization error scaling as $O(\sqrt{T})$ and $O(1/\sqrt{T})$, respectively, up to a logarithmic factor for non-smooth expected reward functions, as well as for smooth functions with unknown smoothness.

1 Introduction

This paper considers the problem of stochastic unimodal optimization with bandit feedback which is a generalization of the classical multi-armed bandit problem solved by Lai and Robbins [19]. The problem is defined by a continuous and unimodal expected reward function μ defined on the interval $[0, 1]$. For this problem, we consider algorithms that repeatedly select an arm $x \in [0, 1]$, and get a noisy reward of mean $\mu(x)$. The performance of an algorithm is characterized by its regret and its optimization error up to time horizon T (the number of observed noisy rewards). The regret is the difference between the average cumulative reward one would obtain if the function μ was known, i.e., $T\sup_{x\in[0,1]}\mu(x)$, and the actual average cumulative reward achieved under the algorithm. The optimization error is the difference between $\sup_{x\in[0,1]}\mu(x)$ and the expected reward of the arm selected at time T. Known lower bounds for the regret and optimization error scale as $\Omega(\sqrt{T})$ (for linear reward functions) and $\Omega(1/\sqrt{T})$ (for quadratic reward functions), respectively. Our objective is to devise an algorithm whose regret and optimization error scale as $O(\sqrt{T})$ and $O(1/\sqrt{T})$ up to a logarithmic factor for a large class of unimodal and continuous reward functions. Such an algorithm would hence be order-optimal. Importantly we merely make any assumption on the smoothness of the reward function – the latter can even be non-differentiable. This contrasts with all existing work investigating similar continuum-armed bandit problems, and where strong assumptions are made on the structure and smoothness of the reward function. These structure and smoothness are known to the decision maker, and are explicitly used in the design of efficient algorithms.

*Supelec, France, mail: richard.combes@supelec.fr
†KTH, Sweden, mail: alepro@kth.se

We propose Stochastic Pentachotomy (SP), an algorithm for which we derive finite-time upper bounds on regret and optimization error. In particular, we show that its regret and optimization error scale as $O(\sqrt{T\log(T)})$ and $O(\sqrt{\log(T)/T})$ for any unimodal and continuous reward function μ that behaves as $\mu(x) = \mu(x^\star) - C|x - x^\star|^\xi$ locally around its maximizer $x^\star$ for some $\xi, C > 0$. These scalings are achieved without the knowledge of ξ or C, i.e., without the knowledge of the smoothness of μ. The SP algorithm consists in successively narrowing an interval in $[0, 1]$ while ensuring that the arm with the highest mean reward remains in this interval with high probability. The narrowing subroutine is a sequential test that takes as input an interval and samples a few arms in the interior of this interval until it gathers enough information to actually reduce the interval. We investigate a general class of such sequential tests. In particular, we provide a (finite time) lower bound of their expected sampling complexity given some guaranteed minimax risk, and design a sequential test that matches this lower bound. This optimal test is used in the SP algorithm. Interestingly, we show that to be efficient, a sequential test needs to sample at least three arms in the interior of the interval to reduce. This implies that a stochastic version of the celebrated Golden section search algorithm cannot achieve a reasonably low regret or optimization error over a large class of reward functions. Indeed such an algorithm would sample only two arms in the interval to reduce. We illustrate the performance of our algorithms using numerical experiments and compare its regret to that of existing algorithms that leverage the smoothness and structure of the reward function.

To our knowledge, SP is the first algorithm for continuous unimodal bandit problems that is order-optimal for a large class of expected reward functions: Its regret and optimization error scale as $O(\sqrt{T\log(T)})$ and $O(\sqrt{\log(T)/T})$ for non-smooth reward functions, as well as for smooth functions with unknown smoothness.

Related work. Stochastic bandit problems with a continuous set of arms have recently received a lot of attention. Various kinds of structured reward functions have been explored, i.e., linear [10], Lipschitz [2], [17], [5], and convex [1], [22]. In these papers, the knowledge of the structure greatly helps the design of efficient algorithms (e.g. for Lipschitz bandits, except in [6], the Lipschitz constant is assumed to be known). More importantly, the smoothness or regularity of the reward function near its maximizer is also assumed to be known and leveraged in the algorithms. Indeed, most existing algorithms use a discretization of the set of arms that depends on this smoothness, and this is crucial to guarantee a regret scaling as $O(\sqrt{T})$. As discussed in [5], [4], without the knowledge of the smoothness, these algorithms would yield a much higher regret (e.g. scaling as $O(T^{2/3})$ for the algorithm proposed in [4]).

Unimodal bandits with a continuous set of arms have been addressed in [9], [24]. In [9], the author shows that Kiefer-Wolfowitz (KW) stochastic approximation algorithm achieves a regret of the order of $O(\sqrt{T})$ under some strong regularity assumptions on the reward function (strong convexity). LSE, the algorithm proposed in [24], has a regret that scales as $O(\sqrt{T}\log(T))$, but requires the knowledge of the smoothness of the reward function. LSE is a stochastic version of the Golden section search algorithm, and iteratively eliminates subsets of arms based on PAC-bounds derived after appropriate sampling. By design, under LSE, the sequence of parameters used for the PAC bounds is pre-defined, and in particular does not depend of the observed rewards. As a consequence, LSE may explore too much sub-optimal parts of the set of arms. Our algorithm exploits more adaptive sequential statistical tests to remove subsets of arms, and yields a lower regret even without the knowledge of the smoothness of the reward function. A naive way to address continuous-armed bandit problems consists in discretizing the set of arms, and in applying efficient discrete bandit algorithms. This method was introduced in [18], and revisited in [7] in the case of unimodal rewards. To get a regret scaling as $O(\sqrt{T}\log(T))$ using this method, the reward function needs to be smooth and the discretization should depend on the smoothness of the function near its maximizer.

Our problem is related to stochastic derivative-free optimization problems where the goal is to get close to the maximizer of the reward function as quickly as possible, see e.g. [23], [14], and references therein. However, as explained in [1], minimizing regret and optimization error constitute different ob-

jectives. Finally, it is worth mentioning papers investigating the design of sampling strategies to identify the best arm in multi-armed bandit problems, see e.g. [21], [11], [3], [15], [13]. These strategies apply to finite sets of arms, but resemble our sequential statistical tests used to reduce the interval containing the best arm. We believe that our analysis (e.g. we derive finite-time lower bounds for the expected sampling complexity of a set of tests), and our proof techniques are novel.

2 Problem Formulation and Notation

We consider continuous bandit problems where the set of arms is the interval $[0,1]$, and where the expected reward μ is a continuous and unimodal function of the arm. More precisely, there exists $x^\star$ such that $x \mapsto \mu(x)$ is strictly increasing (resp. decreasing) in $[0, x^\star]$ (resp. in $[x^\star, 1]$). We denote by $\mathcal{U}$ the set of such functions. Define $\mu^\star = \mu(x^\star)$.

Time proceeds in rounds indexed by $n = 1, 2, \ldots$. When arm x is selected in round n, the observed reward $X_n(x)$ is a random variable whose expectation is $\mu(x)$ and whose distribution is $\nu(\mu(x))$, where ν refers to an exponential family of distributions with one parameter (e.g Bernoulli, exponential, Gaussian, ...). We assume that the rewards $(X_n(x), n \geq 1)$ are i.i.d., and are independent across arms. At each round, a decision rule or algorithm selects an arm depending on the arms chosen in earlier rounds, and the corresponding observed rewards. Let $x^\pi(n)$ denote the arm selected in round n under the algorithm π. The set Π of all possible algorithms consists of sequential decision rules π such that for any $n \geq 2$, $x^\pi(n)$ is $\mathcal{F}^\pi_{n-1}$-measurable where $\mathcal{F}^\pi_n$ is the σ-algebra generated by $(x^\pi(s), X_s(x^\pi(s)), s = 1, \ldots, n)$. The performance of an algorithm $\pi \in \Pi$ with time horizon T is characterized by its regret $R^\pi(T)$ and optimization error $E^\pi(T)$ defined as $R^\pi(T) = T\mu^\star - \sum_{n=1}^T \mathbb{E}[\mu(x^\pi(n))]$ and $E^\pi(T) = \mu^\star - \mathbb{E}[\mu(x^\pi(T))]$. Our objective is to devise an algorithm minimizing these performance metrics. Importantly, the only information available to the decision maker about the reward function μ is that $\mu \in \mathcal{U}$. In particular, the smoothness of μ around $x^\star$ remains unknown – actually μ could well not be differentiable, e.g. $\mu(x) = \mu^\star - |x - x^\star|^\xi$ for $\xi \in (0,1)$.

Notation. In what follows, for any α, β, we denote by KL (α, β) the Kullback-Leibler divergence between distributions $\nu(\alpha)$ and $\nu(\beta)$. When $\alpha, \beta \in [0,1]$, and when $\nu(\cdot)$ is the family of Bernoulli distributions, this KL divergence is denoted by $\mathrm{KL}_2(\alpha, \beta) = \text{KL}\ (\alpha, \beta) = \alpha \log(\frac{\alpha}{\beta}) + (1-\alpha)\log(\frac{1-\alpha}{1-\beta})$.

3 Stochastic Polychotomy Algorithms

We present here a family of sequential arm selection rules, referred to as Stochastic Polychotomy (SP). These algorithms consist in successively narrowing an interval in $[0,1]$ while ensuring that the best arm $x^\star$ remains in this interval with high probability. Under the SP algorithms, the set of rounds is divided into *phases*, where each phase consists in running a subroutine narrowing the interval containing the best arm. The narrowing subroutine used the SP algorithms, and referred to as IT_K (Interval Trimming with K sampled arms), starts with an interval $I = [\underline{x}, \overline{x}]$ and K arms $x_1, \ldots, x_K$ with $\underline{x} \leq x_1 < \ldots < x_K \leq \overline{x}$. It samples these K arms until a decision is taken to reduce the interval I and to output interval I' equal to either $I_1 = [\underline{x}, \max\{x_k : x_k < \overline{x}\}]$ or $I_2 = [\min\{x_k : x_k > \underline{x}\}, \overline{x}]$. The subroutine IT_K is described in details in the next subsection, and its outcome is illustrated in Figure 1.

The pseudo-code of the Stochastic Pentachotomy algorithm, an example of SP algorithm, is presented in Algorithm 1. It uses the narrowing subroutine IT_3 exploiting samples from three arms in the interior of the input interval. IT_3 splits the input interval into five parts (hence the name "Pentachotomy"), and outputs a trimmed interval (referred to as I' in the pseudo-code) and its running time (expressed in number of rounds, and referred to as ℓ in the pseudo-code). The subroutine IT_3 takes as input an interval, a time horizon (equal to the remaining number of rounds in the bandit problem), as well as a parameter controlling its risk, defined as the probability that the subroutine outputs an interval that does not contain

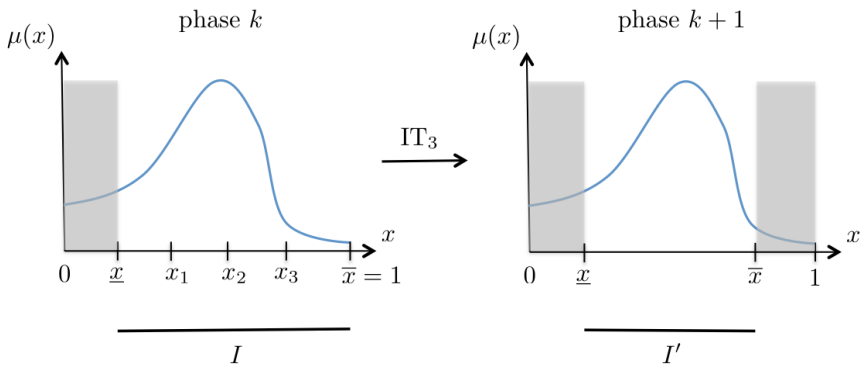

Figure 1: A phase in the SP algorithm. In this phase, we applied the interval narrowing subroutine IT_3 ($K = 3$). The shaded area corresponds to parts of the interval $[0, 1]$ that do not contain the best arm with high probability.

Algorithm 1 The Stochastic Pentachotomy algorithm

Input parameters: time horizon T and confidence parameter $\gamma > 1/2$.
Initialization: $I \leftarrow [0, 1]$ and $s \leftarrow T$.
While $s > 0$:
    Run $\mathrm{IT}_3(I, s, T^{-\gamma})$ and let (I', ℓ) be its output,
    $I \leftarrow I'$,
    $s \leftarrow s - \ell$.

the arm with the highest reward. In the Stochastic Pentachotomy algorithm, the risk parameter in IT_3 is always taken equal to $T^{-\gamma}$ where $\gamma > 1/2$. This choice will ensure that the regret of the algorithm has an optimal scaling in T. Note that LSE, the stochastic version of Golden section search algorithm, belongs to the family of SP algorithms (for LSE, $K = 4$, $x_1 = \underline{x}$, and $x_4 = \overline{x}$).

3.1 IT_K: Asymptotically Optimal Sequential Tests for Interval Trimming

The narrowing subroutines used in each phase of SP algorithms can be interpreted as sequential tests whose final decision is to trim a specific part of the input interval. The IT_K subroutine belongs to the following generic family $\mathcal{T}$ of sequential tests.

Sequential Tests for Interval Trimming. A sequential test $\chi \in \mathcal{T}$ takes as inputs (i) an interval $I = [\underline{x}, \overline{x}] \subset [0, 1]$, and K arms to sample from $x_1, \ldots, x_K$ with $\underline{x} \leq x_1 < \ldots < x_K \leq \overline{x}$, and (ii) a time horizon s that represents the maximum number of samples the test can gather. In round $n \leq s$, the sequential test decides either to terminate and to output a reduced interval $I_1 = [\underline{x}, \max\{x_k : x_k < \overline{x}\}]$ or $I_2 = [\min\{x_k : x_k > \underline{x}\}, \overline{x}]$, or to acquire a new sample from one of the arms $x_1, \ldots, x_K$. The successive decisions taken under sequential test χ are represented by $S^\chi(n) \in \{0, 1, 2\}$. For $n \leq s$, if $S^\chi(n) = 1$, the sequential test terminates and outputs the interval I_1. Similarly if $S^\chi(n) = 2$, χ terminates and outputs I_2. When $S^\chi(n) = 0$ and $n < s$, the sequential test further samples an arm $x^\chi(n)$ in $\{x_1, \ldots, x_K\}$. Finally, if $S^\chi(s) = 0$, we say that the test does not terminate, and it outputs the initial interval I. The sequential test is adapted in the sense that $x^\chi(n)$ and $S^\chi(n)$ are $\mathcal{F}^\chi_{n-1}$-measurable. We denote by $S^\chi \in \{0, 1, 2\}$ the final outcome of the test χ. The length of a sequential test χ is defined as $L^\chi = \inf\{n \leq s : S^\chi(n) \neq 0\}$ if the test terminates and $L^\chi = s$ otherwise. χ also outputs its length.

IT$_K$ Subroutine. To specify our sequential test $\chi =$IT$_K$, we introduce the following notation. Define the sets of functions $B_u = \{\mu \in \mathcal{U} : x^\star \notin I_u\}$, $u \in \{1,2\}$. We also introduce for any $u \in \{1,2\}$, the function $i_u : \mathbb{R}_+^K \to \mathbb{R}$ with

$$i_u(\mu_1, \ldots, \mu_K) = \inf_{\lambda \in B_u} \sum_{k=1}^{K} \text{KL}\,(\mu_k, \lambda(x_k)).$$

We further denote by $t_k^\chi(n) = \sum_{n'=1}^{n \vee L^\chi} \mathbf{1}\{x^\chi(n') = x_k\}$ the number of times arm x_k is sampled up to time n and before the test χ terminates. Finally, we define the empirical average reward of arm x_k up to round $n \leq L_\chi$ as:

$$\hat{\mu}_k(n) = \frac{1}{t_k^\chi(n)} \sum_{n'=1}^{n} X_{n'}(x_k) \mathbf{1}\{x^\chi(n) = x_k\},$$

if $t_k^\chi(n) > 0$ and $\hat{\mu}_k(n) = 0$ otherwise. Let $\hat{\mu}(n) = (\hat{\mu}_1(n), \ldots, \hat{\mu}_K(n))$ and $\bar{t}^\chi(n) = \min_{1 \leq k \leq K} t_k^\chi(n)$. $\chi =$IT$_K$ samples K arms in the interior of $I = [\underline{x}, \overline{x}]$, i.e., $\underline{x} < x_1 < \ldots < x_K < \overline{x}$. To simplify the presentation, we assume that for $k = 1, \ldots, K$, $x_k = \underline{x} + k(\overline{x} - \underline{x})/(K+1)$. This assumption is not crucial, and our analysis remains valid for any choice of arms provided that they lie in the interior of I.

The sequential test $\chi =$IT$_K$ has inputs I and s, as any other test in $\mathcal{T}$. However χ takes an additional input $\zeta > 0$, used to control its risk. Now IT$_K(I, s, \zeta)$ is defined as follows.

Define

$$F(f, s, K) = e^{K+1-f} (f \lceil f \log(s) \rceil / K)^K,$$

and let $f(s, \zeta) \geq K + 1$ be such that $F(f(s, \zeta), s, K) \leq \zeta$ (the precise choice of $f(s, \zeta)$ is free). The test proceeds as follows: For any $n \leq s$:

(i) If there exists $u \in \{1, 2\}$ such that $\bar{t}^\chi(n) i_u(\hat{\mu}(n)) \geq f(s, \zeta)$, then $S^\chi(n) = u$, i.e., χ terminates and its final output is $S^\chi = u$ (ties are broken arbitrarily if both conditions $\bar{t}^\chi(n) i_u(\hat{\mu}(n)) \geq f(s, \zeta)$ for $u = 1, 2$ hold).

(ii) Otherwise $S^\chi(n) = 0$, and χ samples arm $x^\chi(n) = x_{1+(n \bmod K)}$.

The sequential test χ outputs the interval I_{S^χ} where $I_0 = [\underline{x}, \overline{x}]$, $I_1 = [\underline{x}, x_K]$ and $I_2 = [x_1, \overline{x}]$, and its length L^χ.

The performance (i.e. the minimax risk and length) of IT$_K$ will be analysed in Section 4. In view of the results derived in Sections 4 and 5, IT$_K$ is asymptotically optimal among the sequential tests in $\mathcal{T}$. The design of IT$_K$ (e.g. the use of functions i_u, $u \in \{1, 2\}$) is actually motivated by the fundamental performance limits of tests in $\mathcal{T}$ derived in Section 5.

Remark 1 *In the following sections, we will mainly consider the case where the risk $\zeta = s^{-\gamma}$ with $\gamma > 0$. In this case, one may choose $f(s, s^{-\gamma}) = \overline{f}(s) := \gamma \log(s) + 3K \log(\log(s)) + C$, where $C > 0$ is independent of s and γ.*

3.2 IT$'_3$: A Computationally Efficient Sequential Test

Next we present IT$'_3$, a sequential test which is computationally simpler than IT$_3$. IT$'_3$ is not asymptotically optimal, but its implementation is much simpler than that of IT$_3$. Its rationale involves calculating an explicit lower bound of functions i_u, $u \in \{1, 2\}$, and hence IT$'_3$ does not require us to compute i_u. For $\epsilon \geq 0$, we define the function KL$^{\star,\epsilon} : \mathbb{R}^2 \to \mathbb{R}$ as:

$$\text{KL}^{\star,\epsilon}(\mu_1, \mu_2) = \mathbf{1}\{\mu_1 < \mu_2\} \left[\text{KL}\,\left(\mu_1 + \epsilon, \frac{\mu_1 + \mu_2}{2} - \epsilon\right) + \text{KL}\,\left(\mu_2 - \epsilon, \frac{\mu_1 + \mu_2}{2} + \epsilon\right) \right].$$

and $\mathrm{KL}^\star(\mu_1,\mu_2) = \mathrm{KL}^{\star,0}(\mu_1,\mu_2)$. The sequential test $\chi' = \mathrm{IT}'_3$ with inputs I, s and ζ is defined by: for any $n \leq s$,

(i) If $\bar{t}^{\chi'}(n)\mathrm{KL}^\star(\hat{\mu}_1(n),\hat{\mu}_2(n)) \geq f(s,\zeta)$, then $S^{\chi'}(n) = 1$, i.e., χ' terminates and its final output is $S^{\chi'} = 1$. Similarly if $\bar{t}^{\chi'}(n)\mathrm{KL}^\star(\hat{\mu}_3(n),\hat{\mu}_2(n)) \geq f(s,\zeta)$, then $S^{\chi'}(n) = 2$.

(ii) Otherwise $S^{\chi'}(n) = 0$, and χ' samples arm $x^{\chi'}(n) = x_{1+(n \bmod 3)}$.

4 Performance Analysis of the Stochastic Pentachotomy Algorithm

In this section, we analyze the performance of the Stochastic Pentachotomy algorithm. To this aim, we first study how the interval trimming subroutines IT_K (for $K \geq 3$) and IT'_3 perform.

4.1 Minimax Risk and Length of IT_K

Let $\chi \in \mathcal{T}$ be a sequential test for interval trimming. For any $\mu \in \mathcal{U}$, the risk $\alpha^\chi(\mu)$ of χ is the probability that χ outputs an interval that does not contain the optimal arm, i.e, $\alpha^\chi(\mu) = \sum_{u=1}^2 \mathbf{1}\{\mu \in B_u\}\mathbb{P}_\mu[S^\chi = u]$. The minimax risk of χ is then defined as $\alpha^\chi = \sup_{\mu\in\mathcal{U}} \alpha^\chi(\mu)$. Observe that a test that does not terminate (almost surely) has a risk equal to 0, but then its length would be maximal. The analysis of the performance of a test hence consists in characterizing the trade-off between its risk and its length. The next theorem provides upper bounds of the minimax risk of IT_K, as well as of the number of times arms are sampled before the test terminates.

Theorem 4.1 *Let $K \geq 3$ and $I \subset [0,1]$.*
(i) For any $s \geq 1$, the minimax risk of $IT_K(I,s,\zeta)$ is smaller than ζ.
(ii) Let $\gamma > 0$, $u \in \{1,2\}$ and $k \in 1,\ldots,K$. For all $\mu \in \mathcal{U} \setminus B_u$, the test $\chi = IT_K(I,s,s^{-\gamma})$ satisfies:

$$\limsup_{s\to\infty} \frac{\mathbb{E}_\mu[t_k^\chi(s)]}{\log(s)} \leq \frac{\gamma}{i_u(\mu(x_1),\ldots,\mu(x_K))}.$$

The above theorem provides asymptotic guarantees on the length of IT_K. Next, we provide a finite-time analysis of the length of IT_3 and IT'_3, and we also derive an upper bound of the minimax risk of IT'_3.

Finite-time analysis of IT_3 and IT'_3. The next theorem provides explicit upper bounds on the expected length of IT_3 and IT'_3. A high-probability upper-bound on the test length is also provided. This result relies on an explicit lower bound of $i_u(\mu_1,\mu_2,\mu_3)$. Theorem 4.2 will be instrumental in the regret analysis of the Stochastic Pentachotomy algorithm. We restrict the analysis to Bernoulli rewards. This is mainly for simplicity, and the proof techniques can be extended to sub-Gaussian rewards with straightforward modifications.

Theorem 4.2 *Consider $I \subset [0,1]$, $\gamma > 0$ and tests $\chi \in \{IT_3(I,s,s^{-\gamma}), IT'_3(I,s,s^{-\gamma})\}$.*
(i) χ has minimax risk less than $s^{-\gamma}$.
(ii) Define $m = 1$ if $x^\star \in [x_2,\overline{x}]$ and $m = 3$ otherwise. Define $\delta = (\mu(x_2) - \mu(x_m))/2$. Then, we have that for all $0 < \epsilon < \delta/2$, for all $k = 1,2,3$ and all $s \geq 1$:

$$\mathbb{E}_\mu[t_k^\chi(s)] \leq \frac{\overline{f}(s)}{KL^{\star,\epsilon}(\mu(x_m),\mu(x_2))} + 2\epsilon^{-2}.$$

(iii) We have the following inequalities:

$$\text{(a) } \mathbb{P}_\mu[t_k^\chi(s) \geq 8\overline{f}(s)\delta^{-2}] \leq 2e^{-\overline{f}(s)}$$

$$\text{(b) } \mathbb{E}_\mu[t_k^\chi(s)] \leq \frac{32+\overline{f}(s)}{\delta^2}$$

$$\text{(c) } \limsup_{s\to\infty} \frac{\mathbb{E}_\mu[t_k^\chi(s)]}{\log(s)} \leq \frac{\gamma}{KL^\star(\mu(x_m),\mu(x_2))}.$$

Recall that $\overline{f}(s) := \gamma\log(s) + 9\log(\log(s)) + C$, see Remark 1.

4.2 Regret Upper Bounds of the SP algorithm

Next, we analyze the regret of the Stochastic Pentachotomy algorithm. We refer to as SP' the algorithm using the narrowing subroutines IT_3' (instead of IT_3 for SP). Recall that the successive narrowing subroutines IT_3, the risk is always chosen equal to $T^{-\gamma}$, as specified in Algorithm 1. We first derive an upper bound valid for all $\mu \in \mathcal{U}$ and all time horizon T. We then specify the bound when μ behaves as $\mu(x) = \mu(x^\star) - C|x-x^\star|^\xi$ locally around its maximizer $x^\star$ for some $\xi, C > 0$. To simplify the presentation, our bounds are stated and proved for Bernoulli rewards, but the analysis can be extended to other exponential families of distributions.

Let $\mu \in \mathcal{U}$. For any $\Delta > 0$, define the following functions, which will be used to state our regret upper bound:

$$g_\mu(\Delta) = \mu^\star - \max(\mu(x^\star-\Delta), \mu(x^\star+\Delta))$$

$$h_\mu(\Delta) = \min\left\{ \min_{x\in[x^\star, x^\star+\Delta/4]} (\mu(x) - \mu(x+\Delta/4)), \min_{x\in[x^\star-\Delta/4, x^\star]} (\mu(x) - \mu(x-\Delta/4)) \right\}$$

Theorem 4.3 *Let $\psi = 3/4$. Under Algorithm π = SP or π = SP', for all $\mu \in \mathcal{U}$, all $T \geq 1$, and all $N \geq 1$, the regret satisfies:*

$$R^\pi(T) \leq \mu^\star N T^{1-\gamma} + T g_\mu(\psi^N) + 3(\overline{f}(T)+32)\sum_{N'=0}^{N-1} g_\mu(\psi^{N'}) h_\mu(\psi^{N'})^{-2}.$$

We now make the regret upper bound of Theorem 4.3 explicit by considering a particular class of unimodal functions.

Definition 4.4 *For given $0 < C_1 \leq C_2 < \infty$, we define $\mathcal{U}(C_1, C_2)$ the set of all unimodal functions $\mu \in \mathcal{U}$ for which there exists $\xi > 0$ such that:*

(P1) $\mu(x) - \mu(y) \geq C_1(|x^\star - y|^\xi - |x^\star - x|^\xi)$ for all $0 \leq y \leq x \leq x^\star$ and $x^\star \leq x \leq y \leq 1$.

(P2) $|\mu^\star - \mu(x)| \leq C_2|x^\star - x|^\xi$ for all $x \in [0,1]$.

Note that for any $\mu \in \mathcal{U}$ such that $|\mu^\star - \mu(x)| \sim_{x\to x^\star} C|x^\star - x|^\xi$ with $C > 0$, there exists $C_1 > 0$ suitably small and $C_2 < \infty$ suitably large such that $\mu \in \mathcal{U}(C_1, C_2)$. Also note that if $\mu \in \mathcal{U}$ is differentiable on $[0,1] \setminus \{x^\star\}$, with $C_1|x^\star - x|^{\xi-1} \leq |\mu'(x)| \leq C_2|x^\star - x|^{\xi-1}$, then $\mu \in \mathcal{U}(C_1, C_2)$.

Theorem 4.5 *Assume that the algorithm π = SP or π = SP'is parametrized by $\gamma > 1/2$. For all $\mu \in \mathcal{U}(C_1, C_2)$, the regret satisfies:*

$$R^\pi(T) \leq \frac{2\psi^{-3\xi/2}C_2}{C_1 a_\xi}\sqrt{\frac{3T(\overline{f}(T)+32)}{\psi^{-\xi}-1}} + \mu^\star T^{1-\gamma}\frac{\log(TC_1\psi^{-\xi})}{\xi\log(1/\psi)} = O(\sqrt{T\log(T)}).$$

where $a_\xi = 4^{-\xi}\min(1, 2^\xi - 1)$, and where ξ is the parameter associated with μ in Definition 4.4.

Theorem 4.5 states that SP and SP' are order-optimal for all reward functions in $\mathcal{U}(C_1, C_2)$ (with arbitrary C_1 and C_2). They achieve a regret scaling as $O(\sqrt{T\log(T)})$ without the knowledge of the behaviour of the reward function around its maximizer. Although the regret upper bound of Theorem 4.5 is stated for reward functions in class $\mathcal{U}(C_1, C_2)$, we emphasize again that C_1, C_2 and ξ are not input parameters of the algorithms.

4.3 Optimization error of the SP algorithm

We conclude this section by deriving an upper bound on the optimization error of algorithms SP and SP'.

Theorem 4.6 *Let $\psi = 3/4$. Assume that the algorithm $\pi =$ SP or $\pi =$ SP'is parametrized by $\gamma > 1/2$. For all $\mu \in \mathcal{U}(C_1, C_2)$, the optimization error under π satisfies:*

$$E^{\pi}(T) \leq \frac{C_2}{C_1 a_\xi}\sqrt{\frac{24\overline{f}(T)}{T(\psi^{-2\xi}-1)}} + \frac{3T^{-\gamma}\mu^{\star}\log(TC_1\psi^{-\xi})}{\xi\log(1/\psi)} = O(\sqrt{\log(T)/T}),$$

with $a_\xi = 4^{-\xi}\min(1, 2^{\xi}-1)$, and where ξ is the parameter associated with μ in Definition 4.4.

5 Fundamental Performance Limits for Interval Trimming Subroutines

The next theorem provides a lower bound on the expected number of times each arm $x_k, k = 1, \ldots, K$ must be sampled under *any* sequential test with given minimax risk. The lower bound is valid for any time horizon s, which contrasts with the asymptotic lower bounds usually derived in the bandit literature (see e.g. [19]). The proof of this lower bound relies on an elegant information-theoretic argument that exploits the log-sum inequality to derive lower bounds of KL divergence numbers.

Theorem 5.1 *Let $\chi \in \mathcal{T}$ be a sequential test for interval trimming with minimax risk α. Let $\mu \in \mathcal{U}$, and $u \in \{1, 2\}$. Let $\beta = \mathbb{P}_\mu[S^\chi = u]$. If $\alpha \leq \beta$, then*

$$\inf_{\lambda \in B_u} \sum_{k=1}^{K} \mathbb{E}_\mu[t_k^\chi(s)] KL\left(\mu(x_k), \lambda(x_k)\right) \geq KL_2(\beta, \alpha).$$

From the above result, we deduce Corollary 5.2 stating that any sequential test with time horizon s and with minimax risk $s^{-\gamma}$, for $\gamma \in (0, 1]$, has a length that scales at least as $\gamma\log(s)$ as s grows large. Note that the sequential tests IT_K match these lower bound and are hence asymptotically optimal.

Corollary 5.2 *Let $\gamma \in (0, 1]$, $u \in \{1, 2\}$, and $\mu \in \mathcal{U}$. Consider a sequence (indexed by s) of sequential tests χ_s with time horizon s and minimax risk $\alpha^{\chi_s} = s^{-\gamma}$, such that $\lim_{s\to\infty} \mathbb{P}_\mu[S^{\chi_s} = u] = \beta > 0$. Then: $\liminf_{s\to\infty} \inf_{\lambda\in B_u} \sum_{k=1}^{K} \frac{\mathbb{E}_\mu[t_k^{\chi_s}(s)]}{\log(s)} KL\left(\mu(x_k), \lambda(x_k)\right) \geq \gamma\beta$.*

Another consequence of Theorem 5.1 is presented in Corollary 5.3. The latter states that it is impossible to construct a sequential test that samples at most two arms in the interior of I, that terminates before the time horizon s with probability larger than $1/2$ and that has a minimax risk strictly less than $1/4$. Note that if a test terminates before s with probability less than $1/2$, its expected length is at least $s/2$. Such a test would be useless in bandit problems since running it with time horizon $s = T$ would incur a regret linearly growing with T.

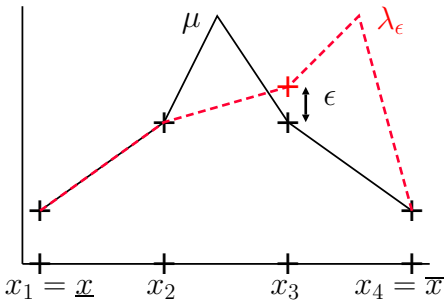

Figure 2: Illustration of Corollary 5.3

Corollary 5.3 *Consider the family of sequential tests running on the interval $I = [\underline{x}, \overline{x}]$, and arms $\underline{x} = x_1 < x_2 < x_3 < x_4 = \overline{x}$. There exists $\mu \in \mathcal{U}$, such that for any sequential test χ of this family with arbitrary finite time horizon s and minimax risk $\alpha < 1/4$, we have $\mathbb{P}_\mu[S^\chi \neq 0] \leq 1/2$ (i.e., the test does not terminate before s with probability $1/2$).*

Recall that Kiefer's Golden section search algorithm [16] uses two points in the interior of the interval to reduce. Hence, the above corollary implies that it is impossible to construct a stochastic version of this algorithm that performs well without additional assumptions on the smoothness and structure of the reward function. Actually, LSE, proposed in [24], is a stochastic version of the Golden section search algorithm, but to analyze its regret, additional assumptions on the structure of the reward function are made (its minimal slope and smoothness).

Corollary 5.3 is a direct consequence of Theorem 5.1: the choice of the reward function μ used in Corollary 5.3 is illustrated in Figure 2, and the result is obtained by considering a sequence (indexed by $\epsilon > 0$) of unimodal functions $\lambda_\epsilon \in B_1$. An efficient test must distinguish between μ and λ_ϵ based on the reward samples at x_1, x_2, x_3, x_4. By letting $\epsilon \to 0$, we see that under such a test, the number of samples from x_3 must be arbitrary large.

6 Numerical Experiments

In this section, we briefly explore the performance of SP$'$ (using parameter $\gamma = 0.6$), and compare it to that of two other algorithms, namely KL-UCB(δ) and KW. KL-UCB(δ) consists in applying the KL-UCB algorithm [12] to the discrete set of arms $\{0, \delta, 2\delta, \ldots, 1\}$. KW is the algorithm proposed in [9]. The performance of LSE [24] is not reported here, since it is generally outperformed by KL-UCB(δ), as shown in [7].

We consider two reward functions satisfying our assumptions with $\xi = 1/2$ and $\xi = 2$, respectively. More precisely, $\mu(x) = 1 - (2|1/2 - x|)^\xi$ for $x \in [0, 1]$. The first function is not differentiable at its maximizer, whereas the second function is just quadratic. Note that KW should then perform well for the quadratic rewards (there the regret scales as $O(\sqrt{T})$ [9]), but there is not guarantee that it would do well for the non-differentiable reward functions. For KL-UCB(δ), the optimal discretization step δ depends on the smoothness of the reward function, and is set to $(\log(T)/\sqrt{T})^{1/\xi}$.

In Figure 3, we present the regret of the various algorithms (averaged over 10 independent runs). Observe that without the knowledge of the smoothness of the function, SP$'$ is able to significantly outperform the two other algorithms. As expected, KW does not perform well when $\xi = 1/2$, but outperforms KL-UCB(δ) for $\xi = 2$.

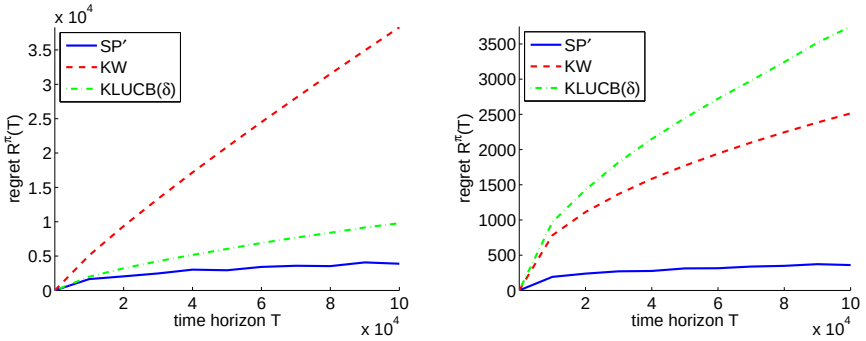

Figure 3: Regret of various algorithms for $\mu(x) = 1 - (2|1/2 - x|)^{\xi}$, $\xi = 0.5$ (left), and $\xi = 2$ (right).

Figure 4 presents a graphical illustration of a typical run of SP′ with reward function $\mu(x) = 1 - (2|1/2 - x|)^{\xi}$, $\xi = 0.5$ (left), and $\xi = 2$ (right), time horizon $T = 10^6$ and $\gamma = 0.6$. We represent the shape of μ and the successive intervals returned by IT'_3, starting at the bottom of the y-axis. The thickness of the segments is an increasing function of the length of IT'_3. In both cases, we observe that the successive intervals contain the optimal arm $x^\star$. When the search interval gets narrower (we are closer to the peak), the intervals get thicker since the duration of the test increases when the separation between arms $\{x_1, x_2, x_3\}$ decreases. Also remark that when the expected reward function is flatter (here $\xi = 2$), the algorithm tends to spend more time on each given interval. Additional numerical experiments are presented in Appendix.

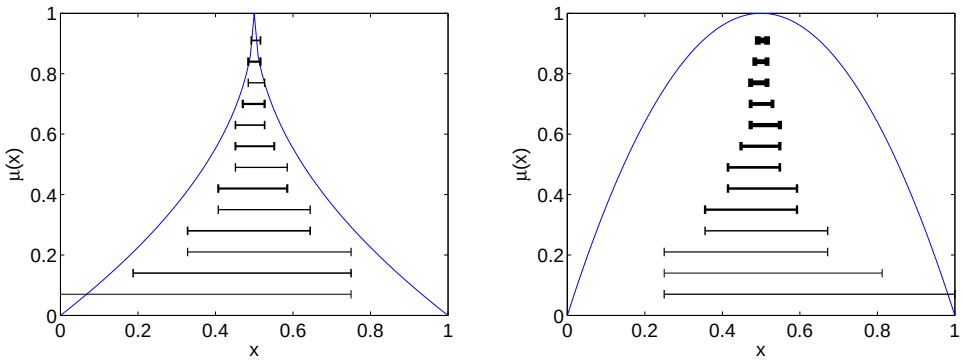

Figure 4: Illustration of a run of SP′ with reward function $\mu(x) = 1 - (2|1/2 - x|)^{\xi}$, $\xi = 0.5$ (left), and $\xi = 2$ (right) and time horizon $T = 10^6$.

7 Conclusion

In this paper, we have presented the first order-optimal algorithms for one-dimensional continuous unimodal bandit problems that do not explicitly take into account the structure or the smoothness of the expected reward function. In some sense, the proposed algorithm learns and adapts its sequential decisions to the smoothness of the function. Future work will be devoted to applying the techniques used to

devise our algorithms to other structured bandits with continuum set of arms (i.e., Lipschitz or convex bandits). We also would like to extend our analysis to the case where the set of arms lies in a space of higher dimension.

References

[1] A. Agarwal, D. Foster, D. Hsu, S. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

[2] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33(6):1926–1951, Nov. 1995.

[3] J. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *Proc. of COLT*, 2010.

[4] P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory*, pages 454–468. Springer, 2007.

[5] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization in x-armed bandits. In *Proc. of NIPS*, 2008.

[6] S. Bubeck, G. Stoltz, and J. Yu. Lipschitz bandits without the Lipschitz constant. In *Proc. of ALT*, 2011.

[7] R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proc. of ICML*, 2014.

[8] R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. Technical Report, http://arxiv.org/abs/1405.5096, 2014.

[9] E. W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. Automat. Contr.*, 54(6):1243–1253, 2009.

[10] V. Dani, T. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. In *Proc. of COLT*, 2008.

[11] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[12] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proc. of COLT*, 2011.

[13] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil' ucb : An optimal exploration algorithm for multi-armed bandits. *Proc. of COLT*, 2014.

[14] K. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In *Proc. of NIPS*, 2012.

[15] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *Proc. of ICML*, 2012.

[16] J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

[17] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proc. of ACM STOC*, pages 681–690, 2008.

[18] R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proc. of NIPS*, 2004.

[19] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–2, 1985.

[20] S. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Proc. of COLT*, 2014.

[21] S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, Dec. 2004.

[22] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. of COLT*, 2013.

[23] J. C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., 2003.

[24] J. Yu and S. Mannor. Unimodal bandits. In *Proc. of ICML*, 2011.

A Additional numerical experiments

Figure 5 compares the regret of the various algorithms for a triangular reward function $\mu(x) = 1 - (2|1/2 - x|)$, and illustrates a typical run of the SP′ algorithm for such a reward function with time horizon $T = 10^6$ and $\gamma = 0.6$.

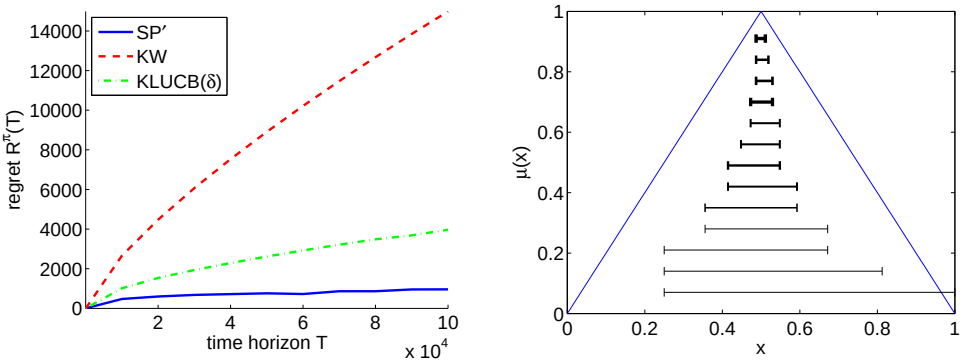

Figure 5: Reward function: $\mu(x) = 1-(2|1/2-x|)$. (Left) Regret vs time of various algorithms. (Right) Illustration of a run of SP′ with time horizon $T = 10^6$.

B Proofs

B.1 Proof of Theorem 4.1

Proof of (i) (Minimax risk). Let $\mu \in \mathcal{U}$, and consider the test $\chi =$IT$_K$. By definition, its risk is:

$$\alpha^\chi(\mu) = \sum_{u=1}^{2} \mathbf{1}\{\mu \in B_u\}\mathbb{P}_\mu[S^\chi = u].$$

If $\mu \notin B_1 \cup B_2$, then $\alpha^\chi(\mu) = 0$ so that the risk is indeed smaller than ζ. Now we assume that $\mu \in B_u$ and we derive an upper bound of $\mathbb{P}_\mu[S^\chi = u]$. By definition of IT$_K$, the event $S^\chi = u$ implies that there exists $n \leq s$ such that $\bar{t}^\chi(n) i_u(\hat{\mu}(n)) \geq f(s,\zeta)$. Using the following two facts: (a) $\mu \in B_u$ and (b) $t_k^\chi(n) \geq \bar{t}^\chi(n)$, we have

$$\begin{aligned} f(s,\zeta) \leq \bar{t}^\chi(n) i_s(\hat{\mu}(n)) = \bar{t}^\chi(n) \inf_{\lambda \in B_u} \sum_{k=1}^{K} \mathrm{KL}\left(\hat{\mu}_k(n), \lambda(x_k)\right) \\ \overset{(a)}{\leq} \bar{t}^\chi(n) \sum_{k=1}^{K} \mathrm{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right) \overset{(b)}{\leq} \sum_{k=1}^{K} t_k^\chi(n) \mathrm{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right). \end{aligned}$$

Therefore we have proven that:

$$\alpha^\chi(\mu) \leq \mathbb{P}_\mu\left[\sup_{n \leq s} \sum_{k=1}^{K} t_k^\chi(n) \mathrm{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right) \geq f(s,\zeta)\right]$$

Applying Theorem B.4 (presented at the end of the appendix) with $\delta := f(s,\zeta)$, we obtain:

$$\alpha^{\chi}(\mu) \leq e^{K+1-f(s,\zeta)}(f(s,\zeta)\lceil f(s,\zeta)\log(s)\rceil/K)^K \leq \zeta.$$

The above inequality holds for all $\mu \in \mathcal{U}$, and hence the minimax risk satisfies $\alpha^{\chi} \leq \zeta$, which concludes the proof of (i).

Proof of (ii) (Expected length).

We now consider $1 \leq k \leq K$ and we derive an upper bound of $\mathbb{E}_{\mu}[t_k^{\chi}(s)]$. Fix $\epsilon > 0$, and define $t_0 = (1+\epsilon)f(s,\zeta)/i_u(\mu(x_1),\ldots,\mu(x_K))$. Introduce the following two sets of rounds:

$$A = \{1 \leq n \leq s : x(n) = x_k, \bar{t}^{\chi}(n) \leq t_0\},$$
$$B = \{1 \leq n \leq s : x(n) = x_k, \bar{t}^{\chi}(n) \geq t_0\}.$$

We have $t_k^{\chi}(s) \leq |A| + |B|$. Furthermore, in each round $n \in A$, $t_k^{\chi}(n)$ is incremented, therefore $|A| \leq t_0$. Now let $n \in B$. By design of IT_K, this implies that: $t_0 \leq \bar{t}^{\chi}(n)$ and $\bar{t}^{\chi}(n) i_u(\hat{\mu}(n)) \leq f(s,\zeta)$. Therefore:

$$t_0 i_u(\hat{\mu}(n)) \leq f(s,\zeta),$$

and thus:

$$i_u(\hat{\mu}(n)) \leq i_u(\mu(x_1),\ldots,\mu(x_K))/(1+\epsilon). \tag{1}$$

Now one can verify that the function $(\lambda_1,\ldots,\lambda_K) \mapsto \sum_{k=1}^K \text{KL}\,(\mu(x_k),\lambda_k)$ attains its infimum on B_u. By continuity of KL in its second argument, there must exist $\lambda^\star \in B_u$ such that:

$$i_u(\mu(x_1),\ldots,\mu(x_K)) = \sum_{k=1}^K \text{KL}\,(\mu(x_k),\lambda^\star(x_k)).$$

Let $\eta > 0$ such that we have $|\hat{\mu}_k(n) - \mu(x_k)| \leq \eta$ for all k. Since $\lambda^\star \in B_u$, this implies that:

$$i_u(\hat{\mu}(n)) = \inf_{\lambda \in B_u} \sum_{k=1}^K \text{KL}\,(\hat{\mu}_k(n),\lambda(x_k)) \leq \sum_{k=1}^K \text{KL}\,(\hat{\mu}_k(n),\lambda^\star(x_k)). \tag{2}$$

Since $|\hat{\mu}_k(n) - \mu(x_k)| \leq \eta$ for all k, the r.h.s. of (2) tends to $i_u(\mu(x_1),\ldots,\mu(x_K)) < i_u(\mu(x_1),\ldots,\mu(x_K))/(1+\epsilon)$ as $\eta \to 0$. Hence the inequality (1) cannot hold for arbitrary small η.

Hence, there exists η_0 such that $n \in B$ implies $\max_k |\hat{\mu}_k(n) - \mu(x_k)| \geq \eta_0$. Note that η_0 might depend on ϵ and $\mu(x_1),\ldots,\mu(x_K)$. Using Lemma B.5, we get $\mathbb{E}[|B|] = o(\log(s))$.

Therefore we have:

$$\mathbb{E}[t_k^{\chi}(s)] \leq \frac{(1+\epsilon)f(s,\zeta)}{i_u(\mu(x_1),\ldots,\mu(x_K))} + o(\log(s)).$$

As noted in remark 1, when considering $\zeta = s^{-\gamma}$, we may use $f(s,\zeta) = \gamma\log(s) + o(\log(s))$, hence:

$$\limsup_{s\to\infty} \frac{\mathbb{E}[t_k^{\chi}(s)]}{\log(s)} \leq \frac{(1+\epsilon)\gamma}{i_u(\mu(x_1),\ldots,\mu(x_K))}.$$

Since the above inequality holds for all $\epsilon > 0$, we obtain the announced result:

$$\limsup_{s\to\infty} \frac{\mathbb{E}[t_k^{\chi}(s)]}{\log(s)} \leq \frac{\gamma}{i_u(\mu(x_1),\ldots,\mu(x_K))}.$$

which concludes the proof of (ii).

B.2 Proof of Theorem 4.2

We start by proving Lemma B.1 which shows that i_u can be lower bounded by the KL* function.

Lemma B.1 *Consider Bernoulli rewards. Define $m = 1$ if $u = 1$ and $m = 3$ otherwise. Then we have for all $u \in \{1, 2\}$:*

$$i_u(\mu(x_1), \mu(x_2), \mu(x_3)) \geq \textit{KL}^\star(\mu(x_m), \mu(x_2)).$$

Proof. We only prove the statement for $u = 1$, as the case $u = 2$ follows by symmetry. By a slight abuse of notation we denote $\mu(x_k)$ and $\lambda(x_k)$ by μ_k and λ_k respectively.

First note that if $\mu_2 < \mu_1$, we have KL$^\star(\mu_1, \mu_2) = 0$ and the statement holds because $i_u(\mu(x_1), \mu(x_2), \mu(x_3)) \geq 0$, since the KL divergence is positive.

Now consider the case $\mu_2 \geq \mu_1$. We have the inequality:

$$i_1(\mu_1, \mu_2, \mu_3) = \inf_{\lambda \in B_1} \sum_{k=1}^{3} \text{KL}\,(\mu_k, \lambda_k) \geq \inf_{\lambda \in B_1} \sum_{k=1}^{2} \text{KL}\,(\mu_k, \lambda_k).$$

Define function $\phi : [0, 1]^2 \to \mathbb{R}$ by $\phi(\lambda_1, \lambda_2) = \sum_{k=1}^{2} \text{KL}\,(\mu_k, \lambda_k)$. Define the set $\Lambda = \{(\lambda_1, \lambda_2) : \lambda_1 \geq \lambda_2\}$. Consider $\lambda \in B_1$, then $x \mapsto \lambda(x)$ attains its maximum in $[\underline{x}, x_1]$, and since λ is unimodal we must have $\lambda_1 \geq \lambda_2$. Therefore:

$$i_1(\mu_1, \mu_2, \mu_3) \geq \min_{(\lambda_1, \lambda_2) \in \Lambda} \phi(\lambda_1, \lambda_2). \tag{3}$$

Consider $(\lambda_1^\star, \lambda_2^\star) \in \arg\min_{(\lambda_1, \lambda_2) \in \Lambda} \phi(\lambda_1, \lambda_2)$. We are going to prove that we must have $\lambda_1^\star = \lambda_2^\star$. Consider two subcases (a) $0 \leq \lambda_2^\star \leq \mu_1$ and (b) $\mu_1 \leq \lambda_2^\star \leq 1$. In case (a) we must have $\lambda_1^\star = \mu_1$ since $\lambda_1 \mapsto \text{KL}\,(\mu_1, \lambda_1)$ attains its minimum at μ_1. In turn we must have $\lambda_2^\star = \mu_1 = \lambda_1^\star$ since $\lambda_1 \mapsto \text{KL}\,(\mu_1, \lambda_1)$ is decreasing for $\lambda_1 \leq \mu_1 \leq \mu_2$. In case (b), we must have $\lambda_1^\star = \lambda_2^\star$ because $\lambda_1 \mapsto \text{KL}\,(\mu_1, \lambda_1)$ is increasing for $\lambda_1 \geq \lambda_2^\star \geq \mu_1$. In both cases we have proven that $\lambda_1^\star = \lambda_2^\star$.

Define function $\tilde{\phi}(\lambda) = \phi(\lambda, \lambda)$, from the reasoning above we have that:

$$\min_{(\lambda_1, \lambda_2) \in \Lambda} \phi(\lambda_1, \lambda_2) = \min_{\lambda \in [0,1]} \tilde{\phi}(\lambda).$$

- If $\mu_1 = \mu_2 = 0$, then $\phi(0, 0) = 0$ so that the optimum is $\lambda^\star = 0$.
- If $\mu_1 = \mu_2 = 1$, then $\phi(1, 1) = 0$, so that the optimum is $\lambda^\star = 1$.
- Otherwise, denote by $\tilde{\phi}'$ the first derivative of $\tilde{\phi}$. We have:

$$\tilde{\phi}'(\lambda) = \frac{2 - (\mu_1 + \mu_2)}{1 - \lambda} - \frac{\mu_1 + \mu_2}{\lambda}.$$

and $\tilde{\phi}'(0^+) = -\infty$ and $\tilde{\phi}'(1^-) = +\infty$ so that $\tilde{\phi}$ attains its maximum in the interior of $[0, 1]$. Solving for $\tilde{\phi}'(\lambda^\star) = 0$ we obtain the unique solution $\lambda^\star = (\mu_1 + \mu_2)/2$.

We observe that in the three above cases, the optimum is $\lambda^\star = (\mu_1 + \mu_2)/2$. We have proven the announced inequality:

$$i_1(\mu(x_1), \ldots, \mu(x_K)) \geq \min_{(\lambda_1, \lambda_2) \in \Lambda} \phi(\lambda_1, \lambda_2) = \min_{\lambda \in [0,1]} \tilde{\phi}(\lambda) = \tilde{\phi}((\mu_1 + \mu_2)/2) = \text{KL}^\star(\mu_1, \mu_2).$$

□

Proof of Theorem 4.2.

(i) Minimax risk of IT$_3$. The minimax risk of IT$_3$ is upper bounded by ζ by Theorem 4.1.

(i)' Minimax risk of χ =IT$'_3$. Let $\mu \in B_u$, let us upper bound $\mathbb{P}_\mu[S^\chi = u]$. Without loss of generality consider $u = 1$ and a time instant $n \leq s$ such that $S^\chi(n) = 1$. By definition of IT'$_3$ this implies that $\bar{t}^\chi(n)\text{KL}^\star(\hat{\mu}_1(n), \hat{\mu}_2(n)) \geq f(s,\zeta)$. We deduce that:

$$\begin{aligned} f(s,\zeta) &\leq \bar{t}^\chi(n)\text{KL}^\star(\hat{\mu}_1(n), \hat{\mu}_2(n)) \\ &\overset{(a)}{\leq} \bar{t}^\chi(n) i_1(\hat{\mu}_1(n), \ldots, \hat{\mu}_K(n)) \\ &= \bar{t}^\chi(n) \inf_{\lambda \in B_1} \sum_{k=1}^K \text{KL}\left(\hat{\mu}_k(n), \lambda(x_k)\right) \\ &\overset{(b)}{\leq} \bar{t}^\chi(n) \sum_{k=1}^K \text{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right) \\ &\overset{(c)}{\leq} \sum_{k=1}^K t_k^\chi(n)\text{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right). \end{aligned}$$

where we have used (a) Lemma B.1, (b) the fact that $\mu \in B_1$ (c) the fact that $\bar{t}^\chi(n) \leq t_k^\chi(n)$ for all k. Applying theorem B.4 once again:

$$\alpha^\chi(\mu) \leq \mathbb{P}\left[\sup_{n \leq s} \sum_{k=1}^K t_k^\chi(n)\text{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right) \geq f(s,\zeta)\right] \leq \zeta$$

which proves that $\alpha^\chi(\mu) \leq \zeta$ for all $\mu \in \mathcal{U}$ and concludes the proof of (i)'.

(ii) Expected duration of χ =IT$_3$. The proof of (ii) for IT$'_3$ follows by the same arguments. By a slight abuse of notation we denote $\mu(x_k)$ by μ_k. Without loss of generality, consider μ such that $x^\star \in [x_2, \overline{x}]$. Therefore we have that $\mu_2 > \mu_1$ since μ is unimodal. Fix $0 < \epsilon < \delta/2$, and define $t_0 = f(s,\zeta)/\text{KL}^{\star,\epsilon}(\mu_1,\mu_2)$. Introduce the two sets of instants:

$$A = \{1 \leq n \leq s : x(n) = x_k, \bar{t}^\chi(n) \leq t_0\} \ , \ B = \{n \geq 1 : x(n) = x_k, \max_{k' \in \{1,2\}} |\hat{\mu}_{k'}(n) - \mu_{k'}| \geq \epsilon\}.$$

We prove that $x(n) = x_k$ implies that $n \in A \cup B$. Consider n such that $\bar{t}^\chi(n) \geq t_0$ and $|\hat{\mu}_{k'}(n) - \mu_{k'}| \leq \epsilon$, $k' \in \{1,2\}$. Since $\epsilon < \delta/2 \leq (\mu_2 - \mu_1)/4$ we have:

$$\begin{aligned} \hat{\mu}_1(n) &\leq \mu_1 + \epsilon \leq (\mu_1 + \mu_2)/2 - \epsilon \leq (\hat{\mu}_1(n) + \hat{\mu}_2(n))/2 \\ \hat{\mu}_2(n) &\geq \mu_2 - \epsilon \geq (\mu_1 + \mu_2)/2 + \epsilon \geq (\hat{\mu}_1(n) + \hat{\mu}_2(n))/2 \end{aligned}$$

so that KL $^\star(\hat{\mu}_1(n), \hat{\mu}_2(n)) \geq$ KL $^{\star,\epsilon}(\mu_1,\mu_2)$. Applying Lemma B.1, we have:

$$\bar{t}(n) i_1(\hat{\mu}(n)) \geq \bar{t}^\chi(n)\text{KL}^\star(\hat{\mu}_1(n), \hat{\mu}_2(n)) \geq t_0 \text{KL}^{\star,\epsilon}(\mu_1,\mu_2) = \overline{f}(s).$$

Therefore we cannot have $x(n) = x_k$.

We have proven that $t_k^\chi(s) \leq |A| + |B|$. Furthermore, at each instant $n \in A$, $\bar{t}^\chi(n)$ is incremented, therefore $|A| \leq t_0$. Let us upper bound the expected size of B. Decompose $B = B^1 \cup B^2$, with:

$$B^{k'} = \{n \geq 1 : x(n) = x_k, |\hat{\mu}_{k'}(n) - \mu_{k'}| \geq \epsilon\}.$$

Let $n \in B^{k'}$ and define $a = \sum_{n' \leq n} \mathbf{1}\{n' \in B^{k'}\}$ so that n is the a-th instant of $B^{k'}$. Then we have that $t_{k'}^\chi(n) \geq a$ and applying [8][Lemma 2.2] we have that for $k' \in \{1,2\}$, $\mathbb{E}[|B^{k'}|] \leq \epsilon^{-2}$. Therefore $\mathbb{E}[|B|] \leq 2\epsilon^{-2}$. So statement (ii) is proven:

$$\mathbb{E}_\mu[t_k^\chi(s)] \leq t_0 + 2\epsilon^{-2} = \frac{\overline{f}(s)}{\text{KL}^{\star,\epsilon}(\mu(x_1), \mu(x_2))} + 2\epsilon^{-2}.$$

(iii) Further bounds on the duration of χ =IT$_3$. The proof of (iii) for IT$'_3$ follows by the same arguments. To establish the announced inequalities, we will use the following fact: from Pinsker's inequality KL $(\alpha, \beta) \geq 2(\alpha - \beta)^2$ for all $(\alpha, \beta) \in [0, 1]^2$, so that:

$$\text{KL}^{\star,\epsilon}(\mu_1, \mu_2) \geq 4((\mu_2 - \mu_1)/2 - 2\epsilon)^2 \geq 4(\delta - 2\epsilon)^2,$$

In particular for $\epsilon = \delta/4$ we have KL $^{\star,\epsilon}(\mu_1, \mu_2) \geq \delta^2$.
Inequality (a). Define $t_0 = 8\overline{f}(s)\delta^{-2}$ and $n_0 = 3t_0$. By design of IT$_3$, for all k we have $t_k^\chi(n_0) = t_0$. Set $\epsilon = \delta/4$. If both $\hat{\mu}_1(n_0) \leq \mu_1 + \epsilon$ and $\hat{\mu}_2(n_0) \geq \mu_2 - \epsilon$ then we have:

$$\overline{t}^\chi(n_0)\text{KL}^\star(\hat{\mu}_1(n_0), \hat{\mu}_2(n_0)) \geq t_0\text{KL}^{\star,\epsilon}(\mu_1, \mu_2) \geq 8\overline{f}(s)\delta^{-2}\delta^2 = 8\overline{f}(s) > \overline{f}(s).$$

so that IT$_3$ must terminate at time n_0 or before. Hence, applying Hoeffding's inequality:

$$\mathbb{P}_\mu[t_k^\chi(s) \geq t_0] \leq \mathbb{P}[\hat{\mu}_k(n_0) \geq \mu_1 + \epsilon] + \mathbb{P}[\hat{\mu}_2(n_0) \leq \mu_2 - \epsilon] \leq 2e^{-2t_0\epsilon^2} = 2e^{-\overline{f}(s)}.$$

which is the announced result.
Inequality (b). Once again setting $\epsilon = \delta/4$, and using both KL $^{\star,\epsilon}(\mu_2, \mu_1) \geq \delta^2$ and statement (ii), we obtain the second claim:

$$\mathbb{E}_\mu[t_k^\chi(s)] \leq \frac{\overline{f}(s) + 32}{\delta^2}$$

Inequality (c). By statement (ii), and using the fact that $\overline{f}(s) = \gamma \log(s) + o(\log(s))$, for all $\epsilon > 0$, we have:

$$\limsup_{s\to\infty} \frac{\mathbb{E}_\mu[t_k^\chi(s)]}{\log(s)} \leq \frac{\gamma}{\text{KL}^{\star,\epsilon}(\mu(x_1), \mu(x_2))},$$

so that letting $\epsilon \to 0$ in the above expression yields:

$$\limsup_{s\to\infty} \frac{\mathbb{E}_\mu[t_k^\chi(s)]}{\log(s)} \leq \frac{\gamma}{\text{KL}^{\star}(\mu(x_1), \mu(x_2))},$$

which concludes the proof of statement (iii). □

B.3 Proof of Theorem 4.3

Fix N throughout the proof. We introduce the following notations. The algorithm proceeds in phases, each phase corresponding to a call of IT$_3$ (or IT$'_3$) subroutine. We define $I^{N'}$ the interval output after the N'-th call of IT$_3$, with $I^0 = [0, 1]$. We define $\tau^{N'}$ the duration of the N'-th call of IT$_3$. Define the event:

$$A = \cap_{N'=0}^{N}\{x^\star \in I^{N'}\},$$

which corresponds to sample paths where the first N-th calls of IT$_3$ have returned an interval containing the optimal arm $x^\star$. We denote by A^c the complement of A.

The regret due to sample paths in A^c is upper bounded by $\mu^\star T\mathbb{P}[A^c]$. The regret due to the N'-th phase for sample paths in A is upper bounded by $\mathbb{E}[\tau^{N'}\mathbf{1}\{A\}(\mu^\star - \min_{x\in I^{N'}} \mu(x))]$. This is true because the N'-th phase has duration $\tau^{N'}$, and during that phase only arms in $I^{N'}$ are sampled so that the regret of a sample in $I^{N'}$ is upper bounded by $\mu^\star - \min_{x\in I^{N'}} \mu(x)$. Therefore the regret admits the following upper bound:

$$R^\pi(T) \leq \mu^\star T\mathbb{P}[A^c] + \sum_{N'\geq 0} \mathbb{E}[\tau^{N'}\mathbf{1}\{A\}(\mu^\star - \min_{x\in I^{N'}} \mu(x))].$$

Consider a sample path in A, and $N' \leq N$, then we have $|I^{N'}| \leq \psi^{N'}$ and $x^\star \in I^{N'}$. Therefore $\mu^\star - \min_{x \in I^{N'}} \mu(x) \leq g_\mu(\psi^{N'})$ by definition of g_μ. Similarly, consider a sample path in A, and $N' > N$. Then we have $I^{N'} \subset I^N$, $|I^N| \leq \psi^N$ and $x^\star \in I^N$. Therefore:

$$\mu^\star - \min_{x \in I^{N'}} \mu(x) \leq \mu^\star - \min_{x \in I^N} \mu(x) \leq g_\mu(\psi^N),$$

and the regret satisfies:

$$\begin{aligned} R^\pi(T) &\leq \mu^\star T \mathbb{P}[A^c] + \sum_{N'=0}^{N} g_\mu(\psi^{N'}) \mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] + g_\mu(\psi^N) \sum_{N'>N} \mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] \\ &\leq \mu^\star T \mathbb{P}[A^c] + \sum_{N'=0}^{N} g_\mu(\psi^{N'}) \mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] + g_\mu(\psi^N) \mathbb{E}[\sum_{N'>N} \tau^{N'}], \\ &\leq \mu^\star T \mathbb{P}[A^c] + \sum_{N'=0}^{N} g_\mu(\psi^{N'}) \mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] + T g_\mu(\psi^N), \end{aligned}$$

where we have used the fact that $\sum_{N'>N} \tau^{N'} \leq \sum_{N' \geq 0} \tau^{N'} = T$.

We now upper bound the probability of event A^c. Since $x^\star \in I^0 = [0,1]$, the occurrence of A^c implies that there exists $N' < N$ such that $x^\star \in I^{N'}$ and $x^\star \notin I^{N'+1}$ so that we have the inclusion:

$$A^c \subset \cup_{N'=0}^{N-1} \{x^\star \in I^{N'}, x^\star \notin I^{N'+1}\}.$$

Since the event $\{x^\star \in I^{N'}, x^\star \notin I^{N'+1}\}$ corresponds to an incorrect decision taken under IT_3, we have $\mathbb{P}[x^\star \in I^{N'}, x^\star \notin I^{N'+1}] \leq T^{-\gamma}$, because of Theorem 4.2. Using a union bound we obtain the upper bound:

$$\mathbb{P}[A^c] \leq \sum_{N'=0}^{N-1} \mathbb{P}[x^\star \in I^{N'}, x^\star \notin I^{N'+1}] \leq N T^{-\gamma}.$$

The regret upper bound becomes:

$$R^\pi(T) \leq \mu^\star N T^{1-\gamma} + T g_\mu(\psi^N) + \sum_{N'=0}^{N} g_\mu(\psi^{N'}) \mathbb{E}[\tau^{N'} \mathbf{1}\{A\}].$$

Finally, from Theorem 4.2, we have that $\mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] \leq 3(\overline{f}(T) + 32)(\delta(I^{N'}))^{-2}$ (we sample from 3 arms) where $\delta(I^{N'})$ is the quantity δ defined in the statement of Theorem 3, when the interval considered by IT_3 is $I^{N'}$. Since we are considering a sample path in A, and $N' \leq N$ we have once again that $|I^{N'}| \leq \psi^{N'}$ and $x^\star \in I^{N'}$ so that $\delta(I^{N'}) \geq h_\mu(\psi^{N'})$ by definition of h_μ. Therefore: $\mathbb{E}[\tau^{N'} \mathbf{1}\{A\}] \leq 3(\overline{f}(T) + 32)(h_\mu(\psi^{N'}))^{-2}$. We obtain finally:

$$R^\pi(T) \leq \mu^\star N T^{1-\gamma} + T g_\mu(\psi^N) + 3(\overline{f}(T) + 32) \sum_{N'=0}^{N} g_\mu(\psi^{N'})(h_\mu(\psi^{N'}))^{-2},$$

which is the announced result and concludes the proof.

B.4 Proof of Theorem 4.5

To prove Theorem 4.5, we use the following intermediate result.

Proposition 1 *For all $\mu \in \mathcal{U}(C_1, C_2)$:*
(a) $g_\mu(\Delta) \leq C_2 \Delta^\xi$;
(b) $h_\mu(\Delta) \geq C_1 a_\xi \Delta^\xi$, with $a_\xi = 4^{-\xi} \min(1, 2^\xi - 1)$

Proof. (a) By definition of g_μ and since $\mu \in \mathcal{U}(C_1, C_2)$, we have:

$$g_\mu(\Delta) = \mu^\star - \min(\mu(x^\star - \Delta), \mu(x^\star + \Delta)) \leq C_2 \Delta^\xi.$$

(b) Consider x such that $x^\star \leq x \leq x^\star + \Delta/4$. Since since $\mu \in \mathcal{U}(C_1, C_2)$, we have:

$$\mu(x) - \mu(x + \Delta/4) \geq C_1((x + \Delta/4 - x^\star)^\xi - (x - x^\star)^\xi).$$

Fix Δ, and define the function $l(x) = (x + \Delta/4 - x^\star)^\xi - (x - x^\star)^\xi$. Its first derivative is:

$$l'(x) = \xi((x + \Delta/4 - x^\star)^{\xi - 1} - (x - x^\star)^{\xi - 1}).$$

Therefore the function $x \mapsto l(x)$ on interval $[x^\star, x^\star + \Delta/4]$ is increasing if $\xi \geq 1$ and decreasing if $\xi < 1$ so we get the lower bound:

$$\min_{x \in [x^\star, x^\star + \Delta/4]} \mu(x) - \mu(x + \Delta/4) \geq \begin{cases} C_1 l(x^\star) = C_1 (\Delta/4)^\xi & \text{if } \xi \geq 1 \\ C_1 l(x^\star + \Delta/4) = C_1 (2^\xi - 1)(\Delta/4)^\xi & \text{if } \xi < 1 \end{cases}$$

so that $h_\mu(\Delta) \geq C_1 \min(1, 2^\xi - 1)(\Delta/4)^\xi$ as announced. □

Let us now prove Theorem 4.5. From Theorem 4.3, we can decompose the regret upper bound into three terms:

$$\begin{aligned} R^\pi(T) &\leq r_1(T) + r_2(T) + r_3(T) \\ r_1(T) &= \mu^\star N T^{1-\gamma} \\ r_2(T) &= T g_\mu(\psi^N) \\ r_3(T) &= 3(\gamma \log(T) + 32) \sum_{N'=0}^{N} g_\mu(\psi^{N'})(h_\mu(\psi^{N'}))^{-2}. \end{aligned}$$

We proceed to upper bound each term. The first term $r_1(T)$ is explicit. By Proposition 1, the second term is upper bounded as: $r_2(T) \leq T C_2 \psi^{\xi N}$. As for the third term $r_3(T)$, by Proposition 1, we have that $g_\mu(\psi^{N'}) \leq C_2 \psi^{\xi N'}$ and $h_\mu(\psi^{N'}) \geq C_1 a_\xi \psi^{\xi N'}$, so that:

$$\sum_{N'=0}^{N} g_\mu(\psi^{N'})(h_\mu(\psi^{N'}))^{-2} \leq \sum_{N'=0}^{N} C_2 (C_1 a_\xi)^{-2} \psi^{-\xi N'} \leq \frac{C_2 \psi^{-\xi(N+1)}}{C_1^2 a_\xi^2 (\psi^{-\xi} - 1)}.$$

Finally, we get:

$$R^\pi(T) \leq \mu^\star N T^{1-\gamma} + T C_2 \psi^{\xi N} + \frac{3(\overline{f}(T) + 32) C_2 \psi^{-\xi(N+1)}}{C_1^2 a_\xi^2 (\psi^{-\xi} - 1)}$$

Define $M \geq 0$ (not necessarily an integer) such that the last two terms in the r.h.s. of the above inequality are equal:

$$C_2 T \psi^{\xi M} = \frac{3(\overline{f}(T) + 32) C_2 \psi^{-\xi(M+1)}}{C_1^2 a_\xi^2 (\psi^{-\xi} - 1)}.$$

We have that:

$$\psi^{-2\xi M} = \frac{TC_1^2 a_\xi^2(\psi^{-\xi}-1)}{3(\overline{f}(T)+32)\psi^{-\xi}} \leq TC_1^2$$

since $a_\xi \leq 1$, $\psi^{-\xi}-1 \leq \psi^{-\xi}$ and $\overline{f}(T) \geq 1$. Taking logarithms we deduce that:

$$M \leq \frac{\log(TC_1)}{\xi \log(1/\psi)}.$$

Now set $N \equiv \lceil M \rceil$ for the remainder of the proof. We obtain the announced upper bound:

$$\begin{aligned}
R^\pi(T) &\leq \mu^\star(M+1)T^{1-\gamma} + TC_2\psi^{\xi M} + \psi^{-\xi}\frac{3(\overline{f}(T)+32)C_2\psi^{-\xi(M+1)}}{C_1^2 a_\xi^2(\psi^{-\xi}-1)} \\
&\leq \mu^\star T^{1-\gamma}\left(\frac{\log(TC_1)}{\xi\log(1/\psi)}+1\right) + (1+\psi^{-\xi})TC_2\psi^{\xi M} \\
&\leq \mu^\star T^{1-\gamma}\frac{\log(TC_1\psi^{-\xi})}{\xi\log(1/\psi)} + \frac{2\psi^{-3\xi/2}C_2}{C_1 a_\xi}\sqrt{\frac{3T(\overline{f}(T)+32)}{\psi^{-\xi}-1}}.
\end{aligned}$$

This concludes the proof.

B.5 Proof of Theorem 4.6

The proof proceeds along the same lines as the proof of Theorem 4.5. Define $M \geq 0$ such that:

$$\frac{24\overline{f}(T)\psi^{-2\xi(M+1)}}{a_\xi^2 C_1^2(\psi^{-2\xi}-1)} = T.$$

Let us first upper bound M. We have that:

$$\psi^{-2\xi(M+1)} = \frac{C_1^2 a_\xi^2 T(\psi^{-2\xi}-1)}{24\overline{f}(T)} \leq C_1^2 T\psi^{-2\xi}.$$

using the fact that $a_\xi \leq 1$, $\overline{f}(T) \geq 1$ and $\psi^{-2\xi}-1 \leq \psi^{-2\xi}$. Hence, taking logarithms:

$$M \leq M+1 \leq \frac{\log(TC_1\psi^{-\xi})}{\xi\log(1/\psi)}.$$

We now fix $N = \lfloor M \rfloor$ for the remainder of the proof. Once again the algorithm proceeds in phases, each phase corresponding to a call to IT_3 (or IT_3'). We define $I^{N'}$ the interval output by the N'-th call of IT_3, with $I^0 = [0,1]$. We define $\tau^{N'}$ the duration of the N'-th call of IT_3. We define two events:

$$\begin{aligned}
A &= \cap_{N'=0}^{N}\{x^\star \in I^{N'}\}, \\
B &= \cap_{N'=0}^{N}\{\tau^{N'} \leq 24\overline{f}(T)h_\mu(\psi^{N'})^{-2}\}
\end{aligned}$$

A corresponds to sample paths where the first N-th calls of IT_3 have returned an interval containing the optimal arm $x^\star$. B corresponds to sample paths where the first N-th calls to IT_3 have not lasted more than their "typical length" (as prescribed by Theorem 4.2). The optimization error can hence be decomposed according to the occurrence of A and B:

$$\begin{aligned}
E^\pi(T) &= \mathbb{E}[(\mu^\star - \mu(x(T))\mathbf{1}\{A\cap B\}] + \mathbb{E}[(\mu^\star-\mu(x(T))\mathbf{1}\{(A\cap B)^c\}] \\
&\leq \mathbb{E}[(\mu^\star - \mu(x(T))\mathbf{1}\{A\cap B\}] + \mu^\star\mathbb{E}[\mathbf{1}\{(A\cap B)^c\}] \\
&\leq \mathbb{E}[(\mu^\star - \mu(x(T))\mathbf{1}\{A\cap B\}] + \mu^\star(\mathbb{P}[A^c] + \mathbb{P}[B^c\cap A]).
\end{aligned}$$

We will establish two facts:

(a) $\mathbb{P}[A^c] + \mathbb{P}[B^c \cap A] \leq 3MT^{-\gamma}$

(b) $(\mu^\star - \mu(x(T))\mathbf{1}\{A \cap B\} \leq C_2\psi^{\xi(M+1)}$ a.s.

If (a) and (b) hold we have that:

$$E^\pi(T) \leq C_2\psi^{\xi(M+1)} + 3\mu^\star MT^{-\gamma} \leq \frac{C_2}{C_1 a_\xi}\sqrt{\frac{24\overline{f}(T)}{T(\psi^{-2\xi}-1)}} + \frac{3T^{-\gamma}\mu^\star \log(TC_1\psi^{-\xi})}{\xi\log(1/\psi)}.$$

which is precisely the announced result.

Fact (a) From Theorem 4.2, statement (i), we know that $\mathbb{P}[A^c] \leq NT^{-\gamma}$ since the risk of IT_3 is upper bounded by $T^{-\gamma}$. Furthermore, from Theorem 4.2, statement (iii a), we know that $\mathbb{P}[B^c \cap A] \leq 2NT^{-\gamma}$ since test IT_3 applied to an interval of size $\psi^{N'}$ that contains the optimal arm has length greater than $24\overline{f}(T)h_\mu(\psi^{N'})^{-2}$ with probability less than $2e^{-\overline{f}(T)} \leq 2T^{-\gamma}$. Hence $\mathbb{P}[A^c] + \mathbb{P}[B^c \cap A] \leq 3NT^{-\gamma} \leq 3MT^{-\gamma}$ as announced.

Fact (b) Let us prove that if B occurs, then the first N-th calls to IT_3 terminate before the time horizon T. Indeed, if B occurs, applying Proposition 1, one has:

$$\begin{aligned}\sum_{N'=0}^{N}\tau^{N'} &\leq 24\overline{f}(T)\sum_{N'=0}^{N}h_\mu(\psi^{N'})^{-2} \leq \frac{24\overline{f}(T)}{a_\xi^2 C_1^2}\sum_{N'=0}^{N}\psi^{-2\xi N'} \\ &\leq \frac{24\overline{f}(T)\psi^{-2\xi(N+1)}}{a_\xi^2 C_1^2(\psi^{-2\xi}-1)} \leq \frac{24\overline{f}(T)\psi^{-2\xi(M+1)}}{a_\xi^2 C_1^2(\psi^{-2\xi}-1)} = T\end{aligned}$$

so that the first N tests do terminate before T. Furthermore, if A occurs, the N-th test returns an arm x such that $|x - x^\star| \leq \psi^{M+1}$. In turn, by proposition 1, one has $|\mu^\star - \mu(x)| \leq g_\mu(\psi^{M+1}) \leq C_2\psi^{\xi(M+1)}$. Hence we have proven that, if both A and B occur one has $(\mu^\star - \mu(x(T)) \leq C_2\psi^{\xi(M+1)}$, so that $(\mu^\star - \mu(x(T))\mathbf{1}\{A \cap B\} \leq C_2\psi^{\xi(M+1)}$ a.s. as announced. This concludes the proof.

B.6 Proof of Theorem 5.1

We work with a given sequential test χ throughout the proof and we omit the superscript $^\chi$ for clarity. Without loss of generality, let $u = 1$. We work with a fixed parameter $\lambda \in B_1$. We denote by $Y(s) = (X_1(x(1)), \ldots, X_s(x(s)))$ the observed rewards from round 1 to round s. We denote by P_s and Q_s the probability distribution of $Y(s)$ under μ and λ respectively. From Lemma B.3 (stated and proved at the end of the appendix), we have:

$$\mathrm{KL}\,(P_s||Q_s) = \sum_{k=1}^{K}\mathbb{E}[t_k(s)]\mathrm{KL}\,(\mu(x_k), \lambda(x_k)). \tag{4}$$

Consider the event $S = 1$. Since the sequential test χ has minimax risk smaller than α, and $\lambda \in B_1$, we have $\mathbb{P}_\lambda[S = 1] \leq \alpha$. Recall that by assumption $\mathbb{P}_\mu[S = 1] = \beta$ and $\alpha \leq \beta$. Now S is a function of $Y(s)$. Using Lemma B.2 (stated at the end of the appendix):

$$\mathrm{KL}\,(P_s||Q_s) \geq \mathrm{KL}_2(\mathbb{P}_\mu[S(s) = 1], \mathbb{P}_\lambda[S(s) = 1]) \geq \mathrm{KL}_2(\beta, \alpha). \tag{5}$$

where we have used the fact that $\alpha \mapsto \mathrm{KL}_2(\beta, \alpha)$ is decreasing for $\alpha \leq \beta$. Putting (4) and (5) together, we obtain:

$$\sum_{k=1}^{K}\mathbb{E}[t_k(s)]\mathrm{KL}\,(\mu(x_k), \lambda(x_k)) \geq \mathrm{KL}_2(\beta, \alpha).$$

Taking the infimum over $\lambda \in B_1$, we obtain the claimed result:

$$\inf_{\lambda \in B_1} \sum_{k=1}^{K} \mathbb{E}[t_k(s)] \mathrm{KL}\left(\mu(x_k), \lambda(x_k)\right) \geq \mathrm{KL}_2(\beta, \alpha).$$

B.7 Proof of Corollary 5.2

Let us denote $\beta_s = \mathbb{P}_\mu[S_s = 1]$, where S_s is the final decision taken under test χ^s. Since $\beta_s \to_{s\to\infty} \beta > 0$ there exists s_0 such that for all $s \geq s_0$ we have $\beta_s \geq s^{-\gamma}$. Since χ_s has minimax risk $\alpha = s^{-\gamma}$, for all $s \geq s_0$, applying Theorem 5.1, we obtain:

$$\inf_{\lambda \in B_1} \sum_{k=1}^{K} \mathbb{E}[t_k(s)] \mathrm{KL}\left(\mu(x_k), \lambda(x_k)\right) \geq \mathrm{KL}_2(\beta_s, \alpha) = \mathrm{KL}_2(\beta_s, s^{-\gamma}). \tag{6}$$

Now by definition of KL_2, we have that:

$$\mathrm{KL}_2(\beta_s, s^{-\gamma}) = \beta_s \log(\beta_s) + \beta_s \gamma \log(s) + (1-\beta_s)\log(1-\beta_s) + (1-\beta_s)\log(1-s^{-\gamma}).$$

Since $\beta_s \to_{s\to\infty} \beta > 0$, we have that $\mathrm{KL}_2(\beta_s, s^{-\gamma}) \sim_{s\to\infty} \gamma\beta \log(s)$. Letting $s \to \infty$ in (6) we have:

$$\lim \inf_{s\to\infty} \inf_{\lambda \in B_1} \sum_{k=1}^{K} \frac{\mathbb{E}_\mu[t_k(s)]}{\log(s)} \mathrm{KL}\left(\mu(x_k), \lambda(x_k)\right) \geq \gamma\beta,$$

which concludes the proof.

B.8 Proof of Corollary 5.3

The proof is constructive: we exhibit a function μ such that $\mathbb{P}_\mu[S^\chi \neq 0] \geq 1/2$. Without loss of generality we consider interval $I = [0, 1]$. Consider the function $\mu(x) = 1 - 2|1/2 - x|$. μ is clearly unimodal, with $x^\star = 1/2$ and $\mu^\star = 1$.

We proceed by contradiction. Consider a test χ such that $\mathbb{P}_\mu[S^\chi \neq 0] \geq 1/2$. Since $S^\chi \in \{0, 1, 2\}$, there exists $u \in \{1, 2\}$ such that $\mathbb{P}_\mu[S^\chi = u] \geq 1/4$. Without loss of generality consider $u = 1$. Let $\epsilon > 0$, and define the function λ^ϵ which is linear on intervals $\{[x_1, x_2], [x_2, x_3], [x_2, (x_3 + x_4)/2], [(x_3 + x_4)/2, x_4]$ with $\lambda(x_k) = \mu(x_k)$, $k \neq 3$ and $\lambda(x_3) = \mu(x_2) + \epsilon$, and $\lambda((x_3 + x_4)/2) = 1$. One can check that λ^ϵ is unimodal, and attains its maximum in $[x_3, x_4]$. We recall that $\alpha < 1/4$ and applying Theorem 5.1, we obtain the following inequality:

$$\sum_{k=1}^{K} \mathbb{E}_\mu[t_k(s)] \mathrm{KL}\left(\mu(x_k), \lambda(x_k)\right) \geq \mathrm{KL}_2(1/4, \alpha).$$

Since $\mathrm{KL}\left(\mu(x_k), \lambda(x_k)\right) = \mathrm{KL}\left(\mu(x_k), \mu(x_k)\right) = 0$, for $k \neq 3$, and $t_3(s) \leq s$ we obtain:

$$s\mathrm{KL}\left(\mu(x_3), \mu(x_3) + \epsilon\right) \geq \mathrm{KL}_2(1/4, \alpha). \tag{7}$$

Since $\alpha < 1/4$ we have that $\mathrm{KL}_2(1/4, \alpha) > 0$. On the other hand $\epsilon \mapsto \mathrm{KL}\left(\mu(x_3), \mu(x_3) + \epsilon\right)$ is continuous, and $\mathrm{KL}\left(\mu(x_3), \mu(x_3)\right) = 0$. Therefore inequality (7) cannot hold for all $\epsilon > 0$. This is a contradiction and proves that a test χ as considered here cannot exist, which concludes the proof.

B.9 Technical results

Lemma B.2 gives a lower bound of the KL divergence of probability measures using the KL divergence between two Bernoulli distributions.

Lemma B.2 *Let P and Q be two probability measures on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that P and Q are both absolutely continuous with respect to measure $m(dx)$. Then:*

$$KL\left(P||Q\right) \geq \sup_{A \in \mathcal{F}} KL_2(P(A), Q(A)).$$

Proof. The proof is based on the log-sum inequality. We recall the derivation of the log-sum inequality here. Consider $f(x) = x\log(x)$. We have that $f''(x) = 1/x$, so that f is convex. We define p, q the densities of P, Q with respect to measure m. Then for all $A \in \mathcal{F}$:

$$\begin{aligned}
\int_A \log\left(\frac{p(x)}{q(x)}\right) p(x)m(dx) &= \int_A f\left(\frac{p(x)}{q(x)}\right) q(x)m(dx) \\
&= Q(A)\int_A f\left(\frac{p(x)}{q(x)}\right)\frac{q(x)}{Q(A)}m(dx) \\
&\overset{(a)}{\geq} Q(A) f\left(\int_A \frac{p(x)}{q(x)}\frac{q(x)}{Q(A)}m(dx)\right) \\
&= Q(A) f\left(\frac{P(A)}{Q(A)}\right) = P(A)\log\left(\frac{P(A)}{Q(A)}\right).
\end{aligned}$$

and (a) holds because of Jensen's inequality. Applying the reasoning above to A and $A^c = \Omega \setminus A$:

$$\begin{aligned}
\text{KL}\left(P||Q\right) &= \int_\Omega \log\left(\frac{p(x)}{q(x)}\right) p(x)m(dx) \\
&= \int_A \log\left(\frac{p(x)}{q(x)}\right) p(x)m(dx) + \int_{A^c} \log\left(\frac{p(x)}{q(x)}\right) p(x)m(dx) \\
&\geq P(A)\log\left(\frac{P(A)}{Q(A)}\right) + P(A^c)\log\left(\frac{P(A^c)}{Q(A^c)}\right) \\
&= P(A)\log\left(\frac{P(A)}{Q(A)}\right) + (1-P(A))\log\left(\frac{1-P(A)}{1-Q(A)}\right) \\
&= \text{KL}_2(P(A), Q(A)).
\end{aligned}$$

So for all A we have:

$$\text{KL}\left(P||Q\right) \geq \text{KL}_2(P(A), Q(A)),$$

and taking the supremum over $A \in \mathcal{F}$ concludes the proof. □

Lemma B.3 evaluates the KL divergence between sample paths of a given test under two different parameters. The proof follows from a straightforward conditioning argument and is omitted here.

Lemma B.3 *We denote by $Y(s) = (X_1(x(1)), \ldots, X_s(x(s)))$ the observed rewards from time 1 to s. Consider $\mu, \lambda \in \mathcal{U}$, and denote by P_s and Q_s the probability distribution of $Y(s)$ under μ and λ respectively. Then we have:*

$$KL\left(P_s||Q_s\right) = \sum_{k=1}^{K} \mathbb{E}[t_k(s)] KL\left(\mu(x_k), \lambda(x_k)\right).$$

Theorem B.4 is a concentration inequality for sums of KL divergences. It was derived derived in [20], and is stated here for completeness.

Theorem B.4 *[20] For all $\delta \geq (K+1)$ and $s \geq 1$ we have:*

$$\mathbb{P}\left[\sup_{n \leq s} \sum_{k=1}^{K} t_k(n) \boldsymbol{KL}\left(\hat{\mu}_k(n), \mu(x_k)\right) \geq \delta\right] \leq e^{K+1-\delta} \left(\frac{\lceil \delta \log(s) \rceil \delta}{K}\right)^K. \tag{8}$$

Lemma B.5 is a technical result showing that the expected number of times the empirical mean of i.i.d. variables deviates by more than δ from its expectation is $o(\log(n))$, n being the time horizon.

Lemma B.5 *Let $\{X_n\}_{n \geq 1}$ be a family of i.i.d. random variables with common expectation μ and finite second moment. Define $\hat{\mu}(n) = (1/n)\sum_{n'=1}^{n} X_{n'}$. For $\delta > 0$ define*
$D^\delta(s) = \sum_{n=1}^{s} \boldsymbol{1}\{|\hat{\mu}(n) - \mu| \geq \delta\}$. Then we have that for all δ:

$$\frac{\mathbb{E}[D^\delta(s)]}{\log(s)} \to_{s \to \infty} 0.$$

Proof. We define $v^2 = \mathbb{E}[(X_1 - \mu)^2]$ the variance. Using the fact that $\{X_n\}_{n \geq 1}$ are independent, we have that $\mathbb{E}[(\hat{\mu}(n) - \mu)^2] = v^2/n$. Applying Chebychev's inequality we have that:

$$\mathbb{P}[|\hat{\mu}(n) - \mu| \geq \delta] \leq \frac{\mathbb{E}[(\hat{\mu}(n) - \mu)^2]}{\delta^2} = \frac{v^2}{n\delta^2}.$$

Therefore, we recognize the harmonic series:

$$\mathbb{E}[D^\delta(s)] = \sum_{n=1}^{s} \mathbb{P}[|\hat{\mu}(n) - \mu| \geq \delta] \leq \frac{v^2}{\delta^2} \sum_{n=1}^{s} \frac{1}{n} \leq \frac{v^2(\log(s)+1)}{\delta^2},$$

so that $\sup_s \mathbb{E}[D^\delta(s)]/\log(s) < \infty$.

Applying the law of large numbers, we have that $\hat{\mu}(n) \to_{n \to \infty} \mu$ a.s., so that $|\hat{\mu}(n) - \mu|$ occurs only finitely many times a.s. Hence $\sup_s D^\delta(s) < \infty$ a.s and $D^\delta(s)/\log(s) \to 0$ a.s.

We have proven that $\sup_s \mathbb{E}[D^\delta(s)]/\log(s) < \infty$ and $D^\delta(s)/\log(s) \to 0$ a.s. so applying Lebesgue's dominated convergence theorem we get the announced result:

$$\frac{\mathbb{E}[D^\delta(s)]}{\log(s)} \to_{s \to \infty} 0,$$

which concludes the proof. □