# Profitable Bandits

**Mastane Achab** FIRST.LAST@TELECOM-PARISTECH.FR
**Stephan Clémençon**
*LTCI, Télécom ParisTech, Paris, France*

**Aurélien Garivier**
*UMPA & LIP, ENS de Lyon, Lyon, France*

**Editors:** Jun Zhu and Ichiro Takeuchi

## Abstract

Originally motivated by default risk management applications, this paper investigates a novel problem, referred to as the *profitable bandit problem* here. At each step, an agent chooses a subset of the $K \geq 1$ possible actions. For each action chosen, she then respectively pays and receives the sum of a random number of costs and rewards. Her objective is to maximize her cumulated profit. We adapt and study three well-known strategies in this purpose, that were proved to be most efficient in other settings: KL-UCB, BAYES-UCB and THOMPSON SAMPLING. For each of them, we prove a finite time regret bound which, together with a lower bound we obtain as well, establishes asymptotic optimality in some cases. Our goal is also to *compare* these three strategies from a theoretical and empirical perspective both at the same time. We give simple, self-contained proofs that emphasize their similarities, as well as their differences. While both Bayesian strategies are automatically adapted to the geometry of information, the numerical experiments carried out show a slight advantage for THOMPSON SAMPLING in practice.

**Keywords:** credit risk, multi-armed bandits, thresholding bandits, index policy, bayesian policy

## 1. Introduction

### 1.1. Motivation

A general and well-known problem for lenders and investors is to choose which prospective clients they should grant loans to, so as to manage credit risk and maximize their profit. A classical supervised learning approach, referred to as *credit risk scoring* consists in ranking all the possible profiles of potential clients, viewed through a collection of socio-economic features $Z$ by means of a (real valued) scoring rule $s(Z)$: ideally, the higher the score $s(Z)$, the higher the default probability. A wide variety of learning algorithms have been proposed to build, from a historical database, a scoring function optimizing ranking performance measures such as the ROC curve or its summary, the AUC criterion, see *e.g.* West (2000), Thomas (2000), Li et al. (2004), Yang (2007) or Creamer and Freund (2004): the *credit risk screening* process then consists in selecting the prospects whose score is below a certain threshold. However, this approach has a serious drawback in general, insofar as new scoring rules are often constructed from truncated information only, namely historical data (the input features $X$ and the observed debt payment behavior) corresponding to past clients, eligible prospects who have been selected by means of a previous scoring rule, jeopardizing

thus the screening procedure when applied to prospects who would have been previously non eligible for credit. Hence, the credit risk problem leads to an exploration vs exploitation dilemma there is no way around for: should clients be used for improving the credit risk estimates, or should they be treated according to the level of risk estimated when they arrive? Lenders thus need sequential strategies able to solve this dilemma.

For simplicity, here we consider the very stylized situation, where each individual from a given category applies for a loan of the same amount in expectation. Extension of the general ideas developed in this paper to more realistic situations will be the subject of further research. In this article, we propose a mathematical model that addresses this issue. We propose several strategies, prove their optimality (by giving a lower bound on the inefficiency of any *uniformly efficient* strategy, together with tight regret analyses) and empirically compare their performance in numerical experiments.

### 1.2. Model

We assume that the population (of credit applicants) is stratified according to $K \geq 1$ categories $a \in \{1, \dots, K\}$. For each category $a$, the credit risk is modelled by a probability distribution $\nu_a$. We assume that at each step $t \in \{1, \dots, T\}$, where $T$ denotes the total number of time steps (or time horizon), the agent is presented a random number $C_a(t) \geq 1$ of clients of each category $a$. She must choose a subset $A_t \subset \{1, \dots, K\}$ of categories to which they grant the loans. We denote by $X_{a,c,t} - L_{a,c,t}$ the profit brought by the client number $c$ of category $a$ at step $t$, $L_{a,c,t}$ being the loan amount and $X_{a,c,t}$ the corresponding reimbursement. In addition, we assume that all loans $L_{a,c,t}$ for the same category $a$ have the same known expectation $\tau_a$. We assume that the variables $\{X_{a,c,t}\}$ are independent, and that $X_{a,c,t}$ has distribution $\nu_a$ and expectation $\mu_a$. We further assume that for any category $a \in \{1, \dots, K\}$, the $C_a(t)$'s are bounded i.e. there exist two positive integers $(c_a^-, c_a^+) \in \mathbb{N}^{*2}$ such that: $c_a^- \leq C_a(t) \leq c_a^+$ for all $t \geq 1$.

Here and throughout, a *sequential strategy* is a set of mappings specifying for each $t$ which categories to choose at time $t$ given the past observations only. In other words, denoting by $I_t = \big(X_{a,c,s}, C_a(s)\big)_{1 \leq s \leq t, a \in A_s, 1 \leq c \leq C_a(s)}$ the vector of variables observed up to time $t \geq 1$, a strategy specifies a sequence $(A_t)_{t \geq 1}$ of random subsets such that, for each $t \geq 2$, $A_t$ is $\sigma(I_{t-1})$-measurable.

It is the goal pursued in this work to define a strategy maximizing the expected cumulated profit given by

$$S_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{I}\{a \in A_t\} \sum_{c=1}^{C_a(t)} X_{a,c,t} - L_{a,c,t}\right].$$

This is equivalent to minimizing the *expected regret*

$$\begin{aligned} R_T &= \sum_{a \in \mathcal{A}^*} \Delta_a \tilde{C}_a(T) - S_T \\ &= \sum_{a \in \mathcal{A}^*} \Delta_a \Big(\tilde{C}_a(T) - \mathbb{E}[N_a(T)]\Big) + \sum_{a \notin \mathcal{A}^*} |\Delta_a| \, \mathbb{E}[N_a(T)], \end{aligned}$$

where $\tilde{C}_a(T) = \mathbb{E}\left[\sum_{t=1}^{T} C_a(t)\right]$ is the expected total number of clients from category $a$ over the $T$ rounds, $N_a(t) = \sum_{s=1}^{t} C_a(s)\,\mathbb{I}\{a \in A_s\}$ is the number of observations obtained from category $a$ up to time $t \geq 1$, $\Delta_a = \mu_a - \tau_a$ is the (unknown) expected profit provided by a client of category $a$ and $\mathcal{A}^* = \left\{a \in \{1, \dots, K\}, \Delta_a > 0\right\}$ is the set of profitable categories.

### 1.3. Illustrative example

Let us consider the credit risk problem in which a bank wants to identify categories of the population they should accept to loan. It may be naturally formulated as a bandit problem with $K$ arms representing the $K$ categories of the population considered. The bank pays $\tau_a$ when loaning to any member of some category $a \in \{1, \dots, K\}$. Each client $c \in \{1, \dots, C_a(t)\}$ of category $a$ receiving a loan from the bank at time step $t$ is characterized by her capacity to reimburse it, namely the Bernoulli r.v. $B_{a,c,t} \sim \mathcal{B}(p_a)$ with $p_a \in [0, 1]$:

- $B_{a,c,t} = 0$ in case of credit default, occurs with probability $1 - p_a$: the bank gets no refunding,
- $B_{a,c,t} = 1$ otherwise, occurs with probability $p_a$: the bank gets refunded $(1 + \rho_a)\tau_a$ with $\tau_a$ the loan amount and $\rho_a$ the interest rate.

All individuals from the same category are considered as independent i.e. the $B_{a,c,t}$'s are i.i.d. realizations of $\mathcal{B}(p_a)$. Hence the refunding $X_{a,c,t}$ received by the bank writes as follows: $X_{a,c,t} = (1 + \rho_a)\tau_a B_{a,c,t}$. Therefore the bank should accept to loan to people belonging to all categories $a \in \{1, \dots, K\}$ such that $\mathbb{E}[X_{a,1,1}] > \tau_a$. This condition rewrites:

$$p_a > \frac{1}{1 + \rho_a}. \tag{1}$$

Hence the role of the bank is to sequentially identify categories verifying Eq. (1) in order to maximize its cumulative profit over the $T$ rounds.

### 1.4. State of the art

In the multi-armed bandit (MAB) problem, a learner has to sequentially explore and exploit different sources in order to maximize the cumulative gain. In the stochastic setting, each source (or *arm*) is associated with a distribution generating random rewards. The optimal strategy in hindsight then consists in always pulling the arm with highest expectation. Many approaches have been proposed for solving this problem such as the UCB1 algorithm (Auer et al. (2002)) for bounded rewards or the Thompson Sampling heuristic first proposed in Thompson (1933). More recently many algorithms have been proven to be asymptotically optimal, particularly in the case of exponential family distributions, such as kl-UCB (Garivier and Cappé (2011)), Bayes-UCB (Kaufmann (2016)) and Thompson Sampling (Kaufmann et al. (2012), Korda et al. (2013)). In this paper we consider a variation of the MAB problem, where, at each time step, the learner may pull several arms simultaneously or no arm at all. To each arm is associated a known threshold and the goal is to maximize the cumulative profit which sums, for each arm pulled by the learner, the difference between the mean reward and the corresponding threshold. This threshold is typically the price to pay for observing a reward from a given arm, e.g. a coin that has to

be inserted in a slot machine. Here the optimal strategy consists in always pulling the arms whose expectations are above their respective thresholds. The case where all arms share the same threshold is studied in Reverdy et al. (2017) with a different definition of regret, which only penalizes pulls of non-profitable arms and hence do not refer to the notion of profit. A similar problem has been tackled in Locatelli et al. (2016) in a best arm identification setting with fixed time horizon and for a unique threshold, where rate-optimal strategies are studied. The purpose of this paper is however different, and we argue that the strategies proposed here are more relevant in many applications (e.g. bank loan management, see Section 1.1).

Indeed, in this paper we mainly focus on deriving asymptotically optimal strategies in the case of one-dimensional exponential family distributions. Section 2 contains an asymptotic lower bound for the profitable bandit problem for any *uniformly efficient* policy. The three following sections (respectively 4, 5 and 6) are devoted to the adaptation of three celebrated MAB strategies (respectively KL-UCB, Bayes-UCB and Thompson Sampling) to the present problem. We provide in each case a finite-time regret analysis. Asymptotical optimality properties of these algorithms are discussed in Section 7. The final Section 8 contains an empirical comparison of the three strategies through numerical experiments.

## 2. Lower Bound

The goal of this section is to give an asymptotic lower bound on the expected regret of any *uniformly efficient* strategy. In this purpose, we adapt the argument of Lai and Robbins (1985), rewritten by Garivier et al. (2016), on asymptotic lower bounds for the expected regret in MAB problems. First we define a model $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_K$ where, for a any arm $a \in \{1, \dots, K\}$, $\mathcal{D}_a$ is the set of candidates for distribution-threshold pairs $(\nu_a, \tau_a)$. Then, we introduce the class of *uniformy efficient* policies that we focus on.

**Definition 1** *A strategy is uniformly efficient if, for any profitable bandit problem* $(\nu_a, \tau_a)_{1 \leq a \leq K} \in \mathcal{D}$*, it satisfies for all arms* $a \in \{1, \dots, K\}$ *and for all* $\alpha \in ]0, 1]$*,* $\mathbb{E}[N_a(T)] = o(\tilde{C}_a(T)^\alpha)$ *if* $\mu_a < \tau_a$ *or* $\tilde{C}_a(T) - \mathbb{E}[N_a(T)] = o(\tilde{C}_a(T)^\alpha)$ *if* $\mu_a > \tau_a$*.*

Now we can state our lower bound which applies to these strategies.

**Theorem 2** *For all models* $\mathcal{D}$*, for all uniformly efficient strategies, for all profitable bandit problems* $(\nu_a, \tau_a)_{1 \leq a \leq K} \in \mathcal{D}$*, for all non-profitable arms* $a$ *such that* $\mu_a < \tau_a$*,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \tau_a)},$$

*where* $\mathcal{K}_{\inf}(\nu_a, \tau_a) = \inf\{KL(\nu_a, \nu'_a), (\nu'_a, \tau_a) \in \mathcal{D}_a, \mu'_a > \tau_a\}$ *with* $KL(\nu_a, \nu'_a)$ *the Kullback-Leibler divergence between distributions* $\nu_a$ *and* $\nu'_a$ *and* $\mu'_a$ *the expectation of* $\nu'_a$*.*

In the remainder of the article, we mainly focus on proposing asymptotically optimal strategies inspired by classical algorithms for MAB, namely KL-UCB (Garivier and Cappé (2011) and Cappé et al. (Jun. 2013)), Bayes-UCB (Kaufmann (2016)) and Thompson Sampling (Kaufmann et al. (2012) and Korda et al. (2013)). For each policy, we prove a corresponding upper bound on its expected regret which will be hopefully tight with respect to the lower bound stated above.

## 3. Preliminaries

### 3.1. Comparison with the classical bandit framework

**One-armed problems.** We point out that the objective of a profitable bandit problem, characterized by $K$ pairs of reward distributions and thresholds $\{(\nu_1, \tau_1), \ldots, (\nu_K, \tau_K)\}$, can be equivalently reformulated as simultaneously solving $K$ independent instances of one-armed subproblems: $\{(\nu_1, \tau_1)\}$, $\ldots$, $\{(\nu_K, \tau_K)\}$. In other words, we could without loss of generality only consider one-armed instances of the profitable bandit problem i.e. the case $K = 1$. Nevertheless, we will still write this paper in the general case $K \geq 1$ in order to refer to MAB notations and to our main motivating application, credit risk, which naturally formulates with several categories. As a consequence of this 'separation' property, the theoretical guarantees on the expected regret that we provide for different policies come with simpler proofs than in MAB: the proofs proposed in this paper contain all core ideas of regret analyses of some of the most sucessful bandit strategies (Thompson Sampling, Bayes-UCB and kl-UCB) with a somewhat simpler and thus more accessible setting.
**Per round numbers of observations.** Another difference with the classical MAB model (where at each round $t \geq 1$ the learner observes only one reward drawn from pulled arm $a$) is that we consider here a more general setting where a random number $C_a(t)$ of i.i.d. rewards sampled from $\nu_a$ are observed. On the other hand, a multiplicative constant (larger than or equal to 1) appears in the upper bounds on the expected regret that we propose for different policies and some parts of their proofs become more intricate.

### 3.2. One-dimensional exponential family

We consider arms with distributions belonging to a one-dimensional exponential family. It should be noted that the kl-UCB-4P algorithm presented next, as kl-UCB, can be shown to apply to the non-parametric setting of bounded distributions, although the resulting approach has weaker optimality properties (see Section 4.3).
**Definitions and properties.** A one-dimensional canonical exponential family is a set of probability distributions $\mathcal{P}_\Theta = \{\nu_\theta, \theta \in \Theta\}$ indexed by a *natural parameter* $\theta$ living in the parameter space $\Theta = ]\theta^-, \theta^+[ \subseteq \mathbb{R}$ and where for all $\theta \in \Theta$, $\nu_\theta$ has a density $f_\theta(x) = A(x)\exp(G(x)\theta - F(\theta))$ with respect to a reference measure $\xi$. $A(x)$ and the sufficient statistic $G(x)$ are functions that characterize the exponential family and $F(\theta) = \log \int A(x)\exp(G(x)\theta)d\xi(x)$ is the normalization function. For notational simplicity, we only consider families with $G(x) = x$, which includes many usual distributions (*e.g.* normal, Bernoulli, gamma among others) but not heavy-tailed distributions, commonly used in financial models, such as Pareto ($G(x) = \log(x)$) or Weibull ($G(x) = x^\ell$ with $\ell > 0$). Nevertheless generalizing all the results proved in this paper to a general sufficient statistic $G(x)$ is straightforward and boils down to considering empirical sufficient statistics $\hat{g}(n) = (1/n)\sum_{s=1}^n G(X_s)$ instead of empirical means. We additionally assume that $F$ is twice differentiable with a continuous second derivative (classic assumption, see e.g. Wasserman (2013)) which implies that $\mu : \theta \mapsto \mathbb{E}_{X\sim\nu_\theta}[X]$ is strictly increasing and thus one-to-one in $\theta$. We denote $\mu^- = \mu(\theta^-)$ and $\mu^+ = \mu(\theta^+)$. The Kullback-Leibler divergence between two distributions $\nu_\theta$ and $\nu_{\theta'}$ in the same exponential family admits the following

closed form expression as a function of the natural parameters $\theta$ and $\theta'$:

$$K(\theta, \theta') := KL(\nu_\theta, \nu_{\theta'}) = F(\theta') - [F(\theta) + F'(\theta)(\theta' - \theta)].$$

We also introduce the KL-divergence between two distributions $\nu_{\mu^{-1}(x)}$ and $\nu_{\mu^{-1}(y)}$:

$$\begin{aligned} d(x, y) :&= K(\mu^{-1}(x), \mu^{-1}(y)) \\ &= \sup_\lambda \{\lambda x - \log \mathbb{E}_{\mu^{-1}(y)}[\exp(\lambda X)]\}, \end{aligned} \tag{2}$$

where the last equality comes from the proof of Lemma 3 in Korda et al. (2013). This last expression of $d$ allows to build a confidence interval on $x$ based on a fixed number of i.i.d. samples from $\nu_{\mu^{-1}(x)}$ by applying the Cramér-Chernoff method (see e.g. Boucheron et al. (2013)).

**Examples.** In Figure 1 we recall some usual examples of one-dimensional exponential families. For some of these distributions that are characterized by two parameters (namely normal, gamma, Pareto and Weibull), one of the two parameters is fixed to define one-dimensional families.

| Distribution | Density | Parameter $\theta$ |
|---|---|---|
| Bernoulli $\mathcal{B}(\lambda)$ | $\lambda^x(1-\lambda)^{1-x}\mathbb{I}_{\{0,1\}}(x)$ | $\log\left(\frac{\lambda}{1-\lambda}\right)$ |
| Normal $\mathcal{N}(\lambda, \sigma^2)$ | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$ | $\frac{\lambda}{\sigma^2}$ |
| Gamma $\Gamma(k, \lambda)$ | $\frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} \mathbb{I}_{\mathbb{R}^+}(x)$ | $-\lambda$ |
| Poisson $\mathcal{P}(\lambda)$ | $\frac{\lambda^x e^{-\lambda}}{x!} \mathbb{I}_{\mathbb{N}}(x)$ | $\log(\lambda)$ |
| Pareto$(x_m, \lambda)$ | $\frac{\lambda x_m^\lambda}{x^{\lambda+1}} \mathbb{I}_{[x_m, +\infty[}(x)$ | $-\lambda - 1$ |
| Weibull$(\ell, \lambda)$ | $\ell\lambda(x\lambda)^{\ell-1} e^{-(\lambda x)^\ell} \mathbb{I}_{\mathbb{R}^+}(x)$ | $-\lambda^\ell$ |

Figure 1: Usual examples of one-dimensional exponential families (parameters $\sigma^2$, $k$, $x_m$ and $\ell$ are fixed).

We mainly investigate the profitable bandit problem in the parametric setting, where all distributions $\{\nu_{\theta_a}\}_{1 \le a \le K}$ belong to a known one-dimensional canonical exponential family $\mathcal{P}_\Theta$ as defined above.

### 3.3. Index policies

All bandit strategies considered in this paper are *index policies*: they are fully characterized by an index $u_a(t)$ which is computed at each round $t \ge 1$ for each arm separately; only arms with an index larger than the threshold $\tau_a$ are chosen. Index policies are formally described in Algorithm 1.

## 4. The kl-UCB-4P Algorithm

We introduce the KL-UCB-4P algorithm, '4P' meaning 'for profit', as a variant of the UCB1 algorithm (Auer et al. (2002)) and more precisely of its improvement KL-UCB introduced

**Algorithm 1** Generic index policy

**Require:** time horizon $T$, thresholds $(\tau_a)_{a\in\{1,\dots,K\}}$.
Pull all arms: $A_1 = \{1,\dots,K\}$.
**for** $t = 1$ **to** $T-1$ **do**
  Compute $u_a(t)$ for all arms $a \in \{1,\dots,K\}$.
  Choose $A_{t+1} \leftarrow \{a \in \{1,\dots,K\}, u_a(t) \geq \tau_a\}$.
**end for**

in Garivier and Cappé (2011). It is defined by the index

$$u_a(t) \quad = \quad \sup\Big\{q \quad > \quad \hat{\mu}_a(t) \quad : \quad N_a(t)d(\hat{\mu}_a(t), q) \quad \leq \quad \log t \; + \; c\log\log t\Big\} \quad ,$$

where $\hat{\mu}_a(t) = (1/N_a(t))\sum_{s=1}^{t}\sum_{c=1}^{C_a(s)} X_{a,c,s}$ is the empirical average reward at time $t$, $d$ is the divergence induced by the Kullback-Leibler divergence defined in Equation (2) and $c$ is a positive constant typically smaller than 3. Due to its special importance for bounded rewards, we name KL-BERNOULLI-UCB-4P the case $d = d_{\mathrm{Bern}} : (x,y) \mapsto x\log(x/y) + (1-x)\log((1-x)/(1-y))$ and KL-GAUSSIAN-UCB-4P the choice $d = d_{\mathrm{Gauss}} : (x,y) \mapsto 2(x-y)^2$.

### 4.1. Analysis for one-dimensional exponential family

We show for the KL-UCB-4P algorithm a finite-time regret bound that proves its asymptotic optimality up to a multiplicative constant $c_a^+/c_a^-$ (see Section 7 for further discussion). To this purpose, we upper-bound the expected number of times non-profitable arms are pulled and profitable ones are not. The analysis is sketched below, while detailed proofs are deferred to the Supplementary Material.

**Theorem 3** *The* KL-UCB-4P *algorithm satisfies the following properties.*
*(i). For any non-profitable arm $a \in \{1,\dots,K\} \setminus \mathcal{A}^*$ and all $\epsilon > 0$,*

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)\frac{c_a^+(\log T + c\log\log T)}{c_a^- d(\mu_a,\tau_a)} + c_a^+\left\{1 + \frac{H_1(\epsilon)}{T^{\beta_1(\epsilon)}}\right\},$$

*where $H_1(\epsilon)$ and $\beta_1(\epsilon)$ are positive functions of $\epsilon$ depending on $c_a^-, \mu_a$ and $\tau_a$.*
*(ii). For any profitable arm $a \in \mathcal{A}^*$, if $T \geq \max(3, c_a^+)$ and $c \geq 3$, we have:*

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+\big\{e(2c+3)\log\log T + c_a^+ + 3\big\}.$$

### 4.2. Sketch of proof

The analysis goes as follows:

(i). For a **non-profitable arm** $a \in \{1,\dots,K\} \setminus \mathcal{A}^*$, we must upper bound $\mathbb{E}[N_a(T)]$. At first, a sub-optimal arm is drawn because its confidence bonus is large. But after some $K_T \approx \kappa\log(T)$ draws (where $\kappa$ is the information constant given in the theorem), the index $u_a(t)$ can be large only when the empirical mean of the observations deviates from its expectation, which has small probability. Thus, we write

$$\mathbb{E}[N_a(T)] \leq c_a^+\left\{K_T + \sum_{t\geq 1}\mathbb{P}(a \in A_{t+1}, N_a(t) > K_T)\right\} .$$

One obtains that $K_T$ gives the main term in the regret. The contribution of the remaining sum is negligible: denoting $d^+(x,y) = d(x,y)\mathbb{I}\{x < y\}$, we observe that:

$$\begin{aligned}(a \in A_{t+1}) &= (u_a(t) \geq \tau_a) \\ &\subset (d^+(\hat{\mu}_a(t), \tau_a) \leq d(\hat{\mu}_a(t), u_a(t))) \\ &= \left(d^+(\hat{\mu}_a(t), \tau_a) \leq \frac{\log(t) + c\log\log(t)}{N_a(t)}\right) .\end{aligned}$$

As a deviation from the mean, the last event proved to have small probability when $N_a(t) > K_T$. Summing over these probabilities produces a term negligible compared to $K_T$.

(ii). For a **profitable arm** $a \in \mathcal{A}^*$, we must upper bound $\tilde{C}_a(T) - \mathbb{E}[N_a(T)]$. We write

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin A_{t+1}) ,$$

and we control the defavorable events by noting that

$$(a \notin A_{t+1}) = (u_a(t) < \tau_a) \subset (u_a(t) < \mu_a) ,$$

where the probability of the last event can be upper bounded by means of a self-normalized deviation inequality such as in Lemma 10 in Cappé et al. (Jun. 2013).

### 4.3. Extension to general bounded rewards

In this subsection, rewards bounded in $[0,1]$ are considered and we build confidence intervals $u_a(t)$ with Bernoulli and Gaussian KL divergence, i.e. $d = d_{\text{Bern}}$ or $d = d_{\text{Gauss}}$, which respectively define KL-Bernoulli-UCB-4P and KL-Gaussian-UCB-4P algorithms. Then, with the same proof as in the one-dimensional exponential family setting, we obtain similar guarantees as in Theorem 3 except that the divergence $d$ is either $d_{\text{Bern}}$ or $d_{\text{Gauss}}$. By Pinsker's inequality, $d_{\text{Bern}}(\mu_a, \tau_a) > d_{\text{Gauss}}(\mu_a, \tau_a)$, which implies that KL-Bernoulli-UCB-4P performs always better than KL-Gaussian-UCB-4P. However, this upper bound is not tight w.r.t. the lower bound stated in Theorem 2 obtained for general bounded distributions. Hence, none of these two approaches is asymptotically optimal. A truly non-parametric, optimal strategy might be obtained by the use of Empirical-Likelihood (EL) confidence intervals, as in Cappé et al. (Jun. 2013), but this is beyond the scope of this article.

## 5. The Bayes-UCB-4P Algorithm

### 5.1. Analysis

We now propose a Bayesian index policy which is derived from Bayes-UCB (Kaufmann (2016)). For all arms $a \in \{1, \dots, K\}$, a prior distribution is chosen for the unknown mean $\mu_a$. At each round $t \geq 1$, we compute the posterior distribution $\pi_{a,t}$ using the $N_a(t)$ observed realizations of $\nu_a$. We compute the quantile $q_a(t) = Q(1 - 1/(t(\log t)^c); \pi_{a,t})$, where $Q(\alpha, \pi)$ denotes the quantile of order $\alpha$ of the distribution $\pi$. The Bayes-UCB-4P is the index policy defined by $u_a(t) = q_a(t)$. In other words, arm $a$ is pulled ($a \in A_{t+1}$) whenever the quantile $q_a(t)$ of the posterior is larger than the threshold $\tau_a$. The following results, proven in the Supplementary Material, show that Bayes-UCB-4P is asymptotically optimal up to a multiplicative constant $c_a^+/c_a^-$ (see Section 7).

**Theorem 4** *When running the* Bayes-UCB-4P *algorithm the following assertions hold.*
*(i). For any non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ and for all $\epsilon > 0$ there exists a problem-dependent constant $N_a(\epsilon)$ such that for all $T \geq N_a(\epsilon)$,*

$$\mathbb{E}[N_a(T)] \leq \left(\frac{1+\epsilon}{1-\epsilon}\right) \frac{c_a^+(\log T + c \log\log T)}{c_a^- d(\mu_a, \tau_a)} + c_a^+ \left\{1 + H_2 + \frac{H_3(\epsilon)}{T^{\beta_2(\epsilon)}}\right\},$$

*where $H_2$, $H_3(\epsilon)$ and $\beta_2(\epsilon)$ are respectively a constant and two positive functions of $\epsilon$ depending on $c_a^-, \tau_a, \mu_a$ and a constant $\mu_0^-$ verifying $\mu^- < \mu_0^- \leq \mu_a$.*
*(ii). For any profitable arm $a \in \mathcal{A}^*$, if $T \geq t_a$ and $c \geq 5$,*

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+ \left\{\frac{e(2(c-2)+4)}{A} \log\log T + t_a + 1\right\},$$

*where $t_a = \max(e/A, 3, A, c_a^+, Ac_a^+)$ and $A$ is a constant depending on the chosen prior distribution.*

### 5.2. Sketch of proof

We present the main steps of the proof of Theorem 4 (see the Supplementary Material for the complete version). The idea is to capitalize on the analysis of kl-UCB-4P, and to relate the quantiles of the posterior distributions to the Kullback-Leibler upper-confidence bounds.

(i). For a non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, we want to upper bound $\mathbb{E}[N_a(T)]$. Again, we use the following decomposition:

$$\mathbb{E}[N_a(T)] \leq c_a^+ \left\{K_T + \sum_{t \geq 1} \mathbb{P}(a \in A_{t+1}, N_a(t) > K_T)\right\},$$

where $K_T \approx \kappa \log(T)$ of the same order of magnitude as the asymptotic lower bound derived in Theorem 2. This cut-off $K_T$ is expected to be the dominant term in our upper bound, since the contribution of the remaining sum is negligible compared to $K_T$: when $N_a(t) > K_T$, we first observe that

$$(a \in A_{t+1}) = (q_a(t) \geq \tau_a) = \left(\pi_{a,t}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c}\right), \tag{3}$$

where the $\pi_{a,t}$ is the posterior distribution on $\mu_a$ at round $t$ and $q_a(t)$ is, under $\pi_{a,t}$, the quantile of order $1 - \frac{1}{t(\log t)^c}$. The key ingredient here is Lemma 4 from Kaufmann (2016), which relates a quantile of the posterior to an upper confidence bound on the empirical mean:

$$\pi_{a,t}([\tau_a, \mu^+[) \lesssim \sqrt{N_a(t)} e^{-N_a(t) d(\hat{\mu}_a(t), \tau_a)}.$$

This permits to conclude as for kl-UCB-4P.

(ii). For a profitable arm $a \in \mathcal{A}^*$, we must upper bound $\tilde{C}_a(T) - \mathbb{E}[N_a(T)]$. We write

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin A_{t+1}).$$

Then we note that for all $t \geq 1$,

$$(a \notin A_{t+1}) = (q_a(t) < \tau_a) = \left(\pi_{a,t}([\tau_a, \mu^+[) < \frac{1}{t(\log t)^c}\right).$$

Using again the bridge between posterior quantiles and upper-confidence bounds of Lemma 4 in Kaufmann (2016):

$$\pi_{a,t}([\tau_a, \mu^+[) \gtrsim \frac{e^{-N_a(t)d(\hat{\mu}_a(t),\tau_a)}}{N_a(t)},$$

we can again argue as for KL-UCB-4P.

## 6. The TS-4P Algorithm

### 6.1. Analysis

The TS-4P algorithm described in this section is inspired from the analysis of Thompson Sampling provided in Korda et al. (2013). Although the guarantees given in Section 5 for Bayes-UCB-4P are valid for any prior distribution, the Bayesian approach proposed in this section will be analyzed only for Jeffreys priors (see Korda et al. (2013) for more details). $\pi_a(0)$ will refer to the prior distribution on $\theta_a$ and $\pi_a(t)$ to the posterior distribution updated with the $N_a(t)$ observations collected from arm $a$ up to time $t$. At each time step $t \geq 1$, sample $\theta_a(t) \sim \pi_a(t)$ and define the TS-4P algorithm (see Algorithm 1) pulling arm $a$ (i.e. $a \in A_{t+1}$) if $u_a(t) = \mu(\theta_a(t))$ is larger than or equal to $\tau_a$.

**Theorem 5** *When running the* TS-4P *algorithm the following assertions hold.*
*(i). For any non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ and for all $\epsilon \in ]0, 1[$,*

$$\mathbb{E}[N_a(T)] \leq \left(\frac{1+\epsilon}{1-\epsilon}\right) \frac{c_a^+ \log T}{c_a^- \, d(\mu_a, \tau_a)} + H_4,$$

*where $H_4$ is a problem dependent constant.*
*(ii). For any profitable arm $a \in \mathcal{A}^*$,*

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq H_5,$$

*with $H_5$ a problem dependent constant.*

### 6.2. Sketch of proof

Here we give the main steps of the proof of Theorem 5 (see the Supplementary Material for complete proof).

(i). For a non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, we must upper bound $\mathbb{E}[N_a(T)]$. We first write:

$$\mathbb{E}[N_a(T)] \lesssim c_a^+ \left\{ K_T + \sum_{t \geq 1} \mathbb{P}\left(a \in A_{t+1}, E_a(t), N_a(t) > K_T\right) \right\},$$

where $K_T \approx \kappa \log(T)$ is, as in the proofs of KL-UCB-4P and Bayes-UCB-4P, a cut-off corresponding to the main term in our bound as suggested by the asymptotic lower bound

in Theorem 2 and $E_a(t)$ is a high probability event ensuring that the current empirical mean at times $t$, namely $\hat{\mu}_a(t)$, is well concentrated around the true mean $\mu_a$. It remains to prove that the sum of defavorable events (for $N_a(t) > K_T$ and under $E_a(t)$) is negligible compared to $K_T$. Observe that the following holds:

$$\mathbb{P}(a \in A_{t+1}, E_a(t), N_a(t) > K_T) \leq \mathbb{P}(\mu(\theta_a(t)) \geq \tau_a, E_a(t), N_a(t) > K_T), \tag{4}$$

where $\theta_a(t)$ is sampled from the posterior distribution $\pi_a(t)$. Then we upper bound the right-hand side expression in Eq. (4) thanks to the deviation inequality stated in Theorem 4 in Korda et al. (2013) and that we recall in Lemma 8 in the Supplementary Material. Summing over these probabilities produces a term negligible compared to $K_T$.

(ii). For a profitable arm $a \in \mathcal{A}^*$, we must upper bound $\tilde{C}_a(T) - \mathbb{E}[N_a(T)]$, which we decompose as follows:

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin A_{t+1}).$$

Then, we control the defavorable events: for all $t \geq 1$ and $b \in ]0, 1[$,

$$\sum_{t=1}^{T-1} \mathbb{P}(a \notin A_{t+1}) \lesssim \sum_{t=1}^{+\infty} \mathbb{P}\left(\mu(\theta_a(t)) < \tau_a, E_a(t) \middle| N_a(t) > t^b\right) + \sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right),$$

where the first series is proved to converge thanks to Lemma 8 and the second too by Lemma 9 provided in the Supplementary. We point out that our proof of Lemma 9, which is a much simplified version of the proof of Proposition 5 in Korda et al. (2013), takes advantage of the independence of arms in our objective (see Section 3.1).

## 7. Asymptotic Optimality

A direct consequence of theorems 3, 4 and 5 is the following asymptotic upper bound on the regret of KL-UCB-4P (with $c \geq 3$), Bayes-UCB-4P (with $c \geq 5$) and TS-4P:

$$\limsup_{T \to \infty} \frac{R_T}{\log T} \leq \sum_{a,\, \mu_a < \tau_a} \frac{c_a^+ |\Delta_a|}{c_a^- \, d(\mu_a, \tau_a)}.$$

Observe that this asymptotic upper bound on the regret is tight with the asymptotic lower bound in Section 2 when $c_a^+ = c_a^-$ for all non-profitable arms $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, which is achieved if and only if the $C_a(t)$'s are constant. In this particular case these three algorithms are asymptotically optimal.

## 8. Numerical Experiments

We perform three series of numerical experiments for three different one-dimensional exponential families: Bernoulli, Poisson and exponential. In each scenario, we consider five arms ($K = 5$) with distributions belonging to the same one-dimensional exponential family. For all arms $a \in \{1, \dots, 5\}$ and time steps $t \in \{1, \dots, T\}$, $C_a(t) - 1$ is sampled from a

Poisson distribution $\mathcal{P}(\lambda_a)$, where $(\lambda_1, \dots, \lambda_5) = (3, 4, 5, 6, 7)$. Moreover, the time horizon is chosen equal to $T = 10000$ and the regret is empirically averaged over 10000 independent trajectories. Our experiments also include algorithms, all index policies, whose theoretical properties have not been discussed in this article, namely:

- UCB-V-4P: same index as UCB-V introduced in Audibert et al. (2009) and using empirical estimates of the variance of each distribution,
- KL-Emp-UCB-4P: same index as empirical KL-UCB introduced in Cappé et al. (Jun. 2013) and using the empirical likelihood principle,
- kl-UCB$^+$-4P: derived from kl-UCB$^+$ introduced in Kaufmann (2016) and defined by the index $u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log(t(\log t)^c / N_a(t))\right\}$.

We also define kl-Bernoulli-UCB$^+$-4P by replacing the divergence $d$ by $d_{\text{Bern}}$ in the index of kl-UCB$^+$-4P.

### 8.1. Scenario 1: Bernoulli

In the first scenario, the $K = 5$ categories have Bernoulli distributions $\mathcal{B}(p_a)$ with parameters $(p_1, \dots, p_5) = (0.1, 0.3, 0.5, 0.5, 0.7)$ and thresholds $(\tau_1, \dots, \tau_5) = (0.2, 0.2, 0.4, 0.6, 0.8)$. Hence the profitable arms are the second and the third ones. Notice that although arms 3 and 4 have the same distribution, namely $\mathcal{B}(0.5)$, their thresholds are different such that arm 3 is profitable but not arm 4.

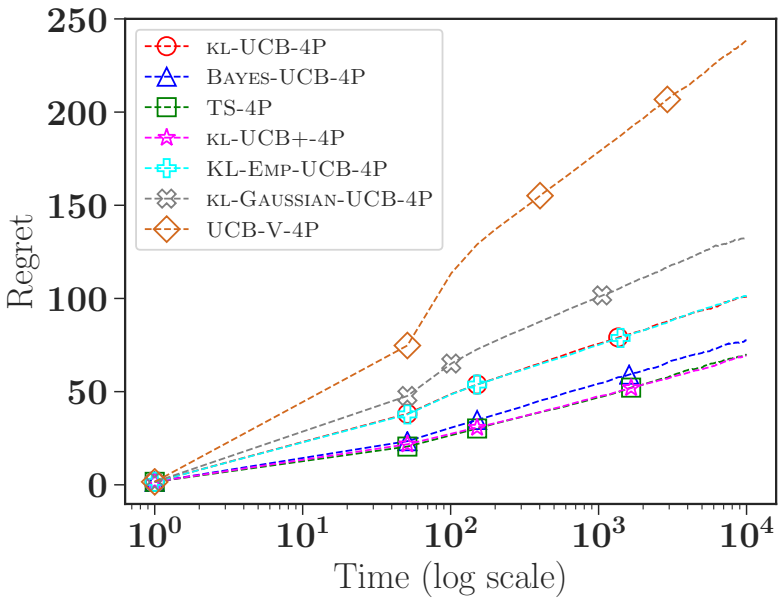


Figure 2: Regret of various algorithms as a function of time in the Bernoulli scenario.

Observe that kl-Gaussian-UCB-4P produces large regret, which confirms the discussion in Section 4.3 stating that it always performs worse than kl-Bernoulli-UCB-4P, which here coincides with kl-UCB-4P.

### 8.2. Scenario 2: Poisson

In the second scenario, the five categories $a \in \{1, \dots, 5\}$ have Poisson distributions $\mathcal{P}(\theta_a)$ with respective mean parameters $(\theta_1, \dots, \theta_5) = (1, 2, 3, 4, 5)$ and thresholds $(\tau_1, \dots, \tau_5) = (2, 1, 4, 3, 6)$: the profitable arms are 2 and 4. In order to run KL-Emp-UCB-4P which assumes boundedness, the rewards are truncated at a maximal value equal to 100.

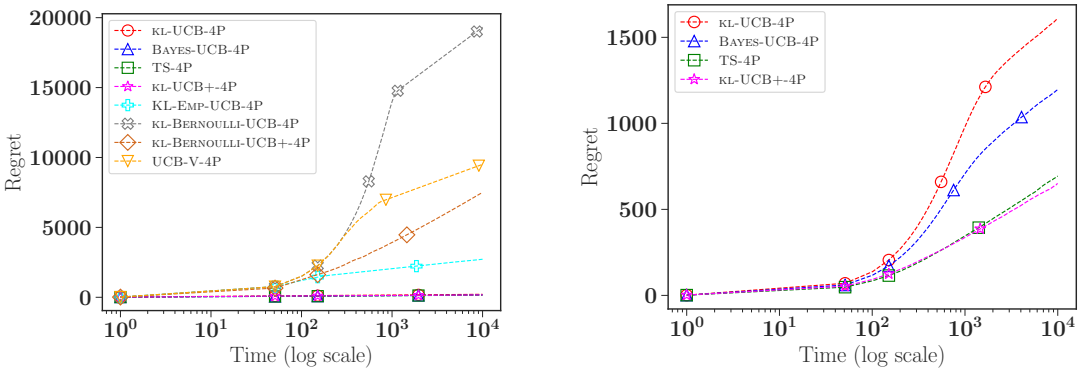


Figure 3: Regret of various algorithms as a function of time in the Poisson scenario. The right hand-side plot only displays the best performing policies on a harder problem.

The right-hand side plot in Figure 3 only displays the regret of the best performing strategies on a harder problem with same distributions but thresholds closer to expectations: $(\tau_1, \dots, \tau_5) = (1.1, 1.9, 3.1, 3.9, 5.1)$.

### 8.3. Scenario 3: exponential

In the third scenario, we consider exponential distributions $\mathcal{E}(\lambda_a)$ with respective mean values $(\lambda_1^{-1}, \dots, \lambda_5^{-1}) = (1, 2, 3, 4, 5)$ and thresholds $(\tau_1, \dots, \tau_5) = (2, 1, 4, 3, 6)$. As in the Poisson scenario, the rewards are truncated at a maximal value of 100.

The right-hand side plot in Figure 4 only displays the best performing strategies. Here again, the distributions are kept the same but the problem is made harder with new thresholds: $(\tau_1, \dots, \tau_5) = (1.1, 1.9, 3.1, 3.9, 5.1)$.

### 8.4. Interpretation

In each scenario and for each algorithm, the regret curve presents a linear regime corresponding to a logarithmic growth as a function of time. We observe that the best performing policies (i.e. with small regret) are those adapting to the parametric family of the reward distributions: through the Kullback-Leibler divergence for kl-UCB-4P and

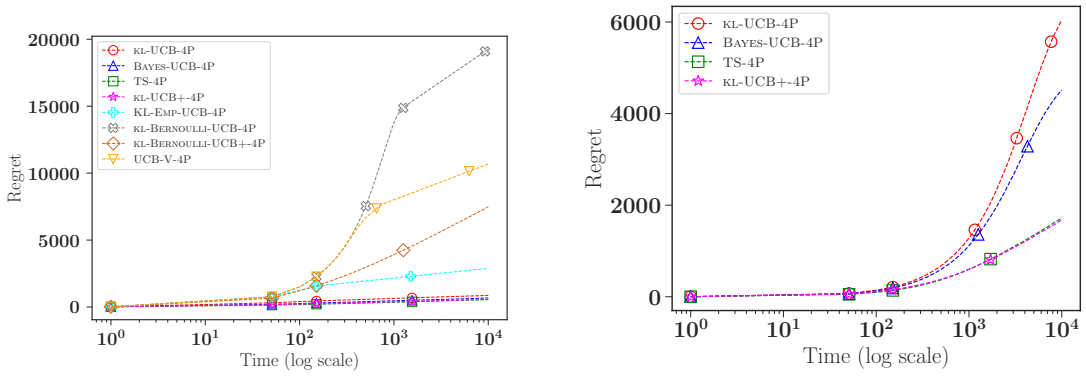


Figure 4: Regret of various algorithms as a function of time in the exponential scenario. The right hand-side plot only displays the best performing policies on a harder problem.

KL-UCB$^+$-4P, or through prior distributions for Bayes-UCB-4P and TS-4P. By contrast, KL-Gaussian-UCB-4P always uses the Gaussian Kullback-Leibler divergence, both KL-Bernoulli-UCB-4P and KL-Bernoulli-UCB$^+$-4P the Bernoulli divergence and KL-Emp-UCB-4P only assumes that the rewards are bounded. Hence we see that prior knowledge on reward distributions is critical in the efficiency of these algorithms.

## 9. Conclusion

Motivated by credit risk evaluation of different populations in a sequential context, this paper introduces the *profitable bandit problem*, evaluates its difficulty by giving an asymptotic lower bound on the expected regret and proposes and theoretically analyzes three algorithms, KL-UCB-4P, Bayes-UCB-4P and TS-4P, by giving finite-time upper bounds on their expected regret for reward distributions belonging to a one-dimensional exponential family. All three algorithms are proven to be asymptotically optimal in the particular setting where for each catefory, a same number of clients is presented to the loaner at each time step. An extension to general bounded distributions is proposed through two algorithms KL-Bernoulli-UCB-4P and KL-Gaussian-UCB-4P coming with finite-time analysis directly derived from the analysis of KL-UCB-4P. We finally compare all these strategies empirically and also against other policies inspired from other multi-armed bandits algorithms. Bayes-UCB-4P and TS-4P perform the best in our numerical experiments and we observe that policies having prior information on the distributions, through appropriate prior distribution for Bayes-UCB-4P and TS-4P or Kullback-Leibler divergence for KL-UCB-4P, perform much better than non-adaptive strategies like KL-Bernoulli-UCB-4P and KL-Gaussian-UCB-4P.

## Acknowledgments

This work was supported by a public grant (*Investissement d'avenir* project, reference ANR-11-LABX-0056-LMH, LabEx LMH) and by the industrial chair *Machine Learning for Big Data* from Télécom ParisTech.

## References

J.Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Cappé, Auréien Garivier, Odalric-Ambrym Maillard, Rmi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, Jun. 2013.

G.G. Creamer and Y. Freund. Predicting performance and quantifying corporate governance risk for latin american adrs and banks. *FINANCIAL ENGINEERING AND APPLICATIONS, MIT, Cambridge*, 2004.

A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. *ArXiv e-prints*, February 2011.

A. Garivier, P. Ménard, and G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *ArXiv e-prints*, February 2016.

E. Kaufmann. On Bayesian index policies for sequential resource allocation. *ArXiv e-prints*, January 2016.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, volume 12, pages 199–213. Springer, 2012.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

X. Li, Ying, W., J. Tuo, W. Li, and W. Liu. Applications of classification trees to consumer credit scoring methods in commercial banks. *Systems, Man and Cybernetics SMC*, 5: 4112–4117, 2004.

A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the Thresholding Bandit Problem. *ArXiv e-prints*, May 2016.

P. Reverdy, V. Srivastava, and N. Leonard. Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, 62(8):3788–3803, 2017.

L.C. Thomas. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16:149–172, 2000.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

D. West. Neural network credit scoring models. *Computers and Operations Research*, 27: 1131–1152, 2000.

Y. Yang. Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, 183(3):1521–1536, 2007.

## Appendix A. Technical Proofs

### A.1. Proof of Theorem 2

We use the inequality $(F)$ in Section 2 in Garivier et al. (2016), a consequence of the contraction of entropy property, which straightforwardly extends from the classical multi-armed bandit setting to ours where several arms can be pulled at each round $t$ and a number $C_a(t) \geq 1$ of observations are observed simultaneously for each pulled arm $a$. Then we have

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) \geq \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]), \tag{5}$$

where $Z$ is any $\sigma(I_T)$-measurable random variable with values in $[0,1]$. Consider a thresholding bandit problem $(\nu, \tau) \in \mathcal{D}$ with at least one non-profitable arm $a \in \{1, \dots, K\}$, we define a modified problem $(\nu', \tau)$ such that $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in \mathcal{D}_a$ verifies $\mu'_a > \tau_a$. Then, considering $Z = N_a(T)/\tilde{C}_a(T)$, Eq. (5) rewrites as follows:

$$\begin{aligned}
\mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) &\geq \mathrm{kl}(\mathbb{E}_\nu[N_a(T)]/\tilde{C}_a(T), \mathbb{E}_{\nu'}[N_a(T)]/\tilde{C}_a(T)) \\
&\geq \left(1 - \frac{\mathbb{E}_\nu[N_a(T)]}{\tilde{C}_a(T)}\right) \log\left(\frac{\tilde{C}_a(T)}{\tilde{C}_a(T) - \mathbb{E}_{\nu'}[N_a(T)]}\right) - \log(2),
\end{aligned}$$

where we used for the last inequality that for all $(p,q) \in [0,1]^2$,

$$\mathrm{kl}(p,q) \geq (1-p)\log\left(\frac{1}{1-q}\right) - \log(2).$$

Then, by uniform efficiency it holds: $\mathbb{E}_\nu[N_a(T)] = o(\tilde{C}_a(T))$ and $\tilde{C}_a(T) - \mathbb{E}_{\nu'}[N_a(T)] = o(\tilde{C}_a(T)^\alpha)$ for all $\alpha \in (0,1]$. Hence for all $\alpha \in (0,1]$,

$$\liminf_{T\to\infty} \frac{1}{\log T}\mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) \geq \liminf_{T\to\infty} \frac{1}{\log T} \log\left(\frac{\tilde{C}_a(T)}{\tilde{C}_a(T)^\alpha}\right) = 1 - \alpha.$$

Taking the limit $\alpha \to 0$ in the right-hand side and taking the infimum over all distributions $\nu'_a \in \mathcal{D}_a$ such that $\mu'_a > \tau_a$ in the left-hand side conclude the proof.

### A.2. Proof of Theorem 3

For any arm $a \in \{1, \dots, K\}$, the average reward at time $t$ is denoted by $\hat{\mu}_a(t) = S_a(t)/N_a(t)$ where $S_a(t) = \sum_{s=1}^{t} \sum_{c=1}^{C_a(s)} X_{a,c,s}\, \mathbb{I}\{a \in A_s\}$ and $N_a(t) = \sum_{s=1}^{t} C_a(s)\, \mathbb{I}\{a \in A_s\}$. For every positive integer $s$, we also denote by $\hat{\mu}_{a,s} = (X_{a,1} + \cdots + X_{a,s})/s$ with $X_{a,1}, \dots, X_{a,s}$ the first $s$ samples pulled from arm $a$, so that $\hat{\mu}_t(a) = \hat{\mu}_{a,N_a(t)}$. The upper confidence bound for $\mu_a$ appearing in KL-UCB-4P is then given by:

$$u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t\right\}.$$

For all $(x,y) \in [\mu^-, \mu^+]^2$, define $d^+(x,y) = d(x,y)\mathbb{I}\{x < y\}$.

(i). Let $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ be a non-profitable arm i.e. such that $\mu_a < \tau_a$. Given $\epsilon \in ]0, 1[$, we upper bound the expectation of $N_a(T)$ as follows,

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} C_a(t)\mathbb{I}\{a \in A_t\}\right] \leq c_a^+ \, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{a \in A_t\}\right].$$

Now observe for $t \geq 1$ that $a \in A_{t+1}$ implies $u_a(t) \geq \tau_a$ and hence,

$$d^+(\hat{\mu}_a(t), \tau_a) \leq d(\hat{\mu}_a(t), u_a(t)) = \frac{\log t + c \log\log t}{N_a(t)}.$$

Then,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{I}\{a \in A_t\} \\
&= 1 + \sum_{t=1}^{T-1} \mathbb{I}\left\{a \in A_{t+1}\right\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{a \in \bigcap_{i \in \{i_1, \dots, i_s\}} A_i, a \notin \bigcup_{i \in \{1, \dots, t\} \setminus \{i_1, \dots, i_s\}} A_i\right\} \\
&\qquad \times \mathbb{I}\left\{(C_a(i_1) + \cdots + C_a(i_s)) d^+\left(\hat{\mu}_{a, C_a(i_1) + \cdots + C_a(i_s)}, \tau_a\right) \leq \log t + c \log\log t\right\}.
\end{aligned} \tag{6}$$

Given $\epsilon \in ]0, 1[$, we upper bound the last indicator function appearing in Eq. (6) by

$$\begin{aligned}
&\mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, k d^+(\hat{\mu}_{a,k}, \tau_a) \leq \log T + c \log\log T\right\} \\
&\leq \mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{1+\epsilon}\right\},
\end{aligned} \tag{7}$$

where $K_T = \left\lceil (1+\epsilon) \frac{\log T + c \log\log T}{c_a^- \, d(\mu_a, \tau_a)} \right\rceil$. The last expression in Eq. (7) is not using the indices $t, i_1, \dots, i_s$ which allows us to exchange the sums over $t$ and $s$ in Eq. (6) and to obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{I}\{a \in A_t\} \\
&\leq 1 + \sum_{s=1}^{T} \left( \mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{1+\epsilon}\right\} \right) \\
&\qquad \times \sum_{t=1}^{T-1} \mathbb{I}\left\{a \in A_{t+1}\right\} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{a \in \bigcap_{i \in \{i_1, \dots, i_s\}} A_i, a \notin \bigcup_{i \in \{1, \dots, t\} \setminus \{i_1, \dots, i_s\}} A_i\right\} \\
&\leq K_T + \sum_{s=K_T}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{1+\epsilon}\right\},
\end{aligned}$$

where the last inequality is implied by

$$\sum_{t=1}^{T-1} \mathbb{I}\{a \in A_{t+1}\} \sum_{1\le i_1<\cdots<i_s\le t} \mathbb{I}\left\{a \in \bigcap_{i\in\{i_1,\ldots,i_s\}} A_i, a \notin \bigcup_{i\in\{1,\ldots,t\}\setminus\{i_1,\ldots,i_s\}} A_i\right\} \le 1. \quad (8)$$

Hence,

$$\begin{aligned}\mathbb{E}[N_a(T)] &\le c_a^+ \left\{ K_T + \sum_{s=K_T}^{+\infty} \sum_{k=c_a^- s}^{+\infty} \mathbb{P}\left( d^+(\hat{\mu}_{a,k}, \tau_a) \le \frac{d(\mu_a,\tau_a)}{1+\epsilon}\right)\right\} \\ &\le (1+\epsilon)\frac{c_a^+}{c_a^-}\frac{\log T + c\log\log T}{d(\mu_a,\tau_a)} + c_a^+\left\{1 + \frac{H_1(\epsilon)}{T^{\beta_1(\epsilon)}}\right\},\end{aligned}$$

comes from Lemma 6 with $H_1(\epsilon)$ and $\beta_1(\epsilon)$ positive functions of $\epsilon$.

(ii). Now consider $a \in A^*$ i.e. verifying $\mu_a > \tau_a$. It follows,

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=2}^{T} C_a(t)\mathbb{I}\{a \notin A_t\}\right] \le c_a^+ \sum_{t=1}^{T-1} \mathbb{P}\left(u_a(t) < \mu_a\right).$$

Let $t \in \{1,\ldots,T-1\}$ and observe that $(u_a(t) < \mu_a) \subset (d^+(\hat{\mu}_a(t),\mu_a) > d(\hat{\mu}_a(t), u_a(t)))$. Hence for $c \ge 3$ and $t \ge \max(3, c_a^+)$,

$$\begin{aligned}&\mathbb{P}\left(u_a(t) < \mu_a\right) \\ &\le \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t),\mu_a) > \delta_t\right) \le (\delta_t \log(c_a^+ t) + 1)\exp(-\delta_t + 1) \\ &= \frac{e((\log t)^2 + c\log(t)\log\log(t) + \log(c_a^+)\log(t) + c\log(c_a^+)\log\log(t) + 1)}{t(\log t)^c} \\ &\le \frac{e(2c+3)}{t\log t},\end{aligned}$$

where $\delta_t = \log t + c\log\log t > 1$ and the second inequality results from the self-normalized concentration inequality stated in Lemma 10 in Cappé et al. (Jun. 2013). Then by summing over $t$,

$$\begin{aligned}\tilde{C}_a(T) - \mathbb{E}[N_a(T)] &\le c_a^+\left\{2 + c_a^+ + e(2c+3)\sum_{t=3}^{T-1}\frac{1}{t\log t}\right\} \\ &\le c_a^+\left\{e(2c+3)\log\log T + c_a^+ + 3\right\}.\end{aligned}$$

### A.3. Lemma 6

**Lemma 6** *Let $a \in \{1,\ldots,K\}\setminus\mathcal{A}^*$ a non-profitable arm (i.e. $\mu_a < \tau_a$), $\epsilon \in ]0,1[$ and $K_T = \left\lceil f(\epsilon)\frac{\log T + c\log\log T}{c_a^- d(\mu_a,\tau_a)}\right\rceil$ with $f$ a function such that $f(\epsilon') > 1$ for all $\epsilon' \in ]0,1[$. Then there exist $H(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that*

$$\sum_{s=K_T}^{+\infty}\sum_{k=c_a^- s}^{+\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,k},\tau_a) \le \frac{d(\mu_a,\tau_a)}{f(\epsilon)}\right) \le \frac{H(\epsilon)}{T^{\beta(\epsilon)}},$$

*where $H(\epsilon)$ and $\beta(\epsilon)$ are positive functions of $\epsilon$ depeding on $\mu_a, \tau_a$ and $c_a^-$.*

**Proof** Observe that $d^+(\hat{\mu}_{a,k}, \tau_a) \leq d(\mu_a, \tau_a)/f(\epsilon)$ if and only if $\hat{\mu}_{a,k} \geq r(\epsilon)$ where $r(\epsilon) \in ]\mu_a, \tau_a[$ verifies $d(r(\epsilon), \tau_a) = d(\mu_a, \tau_a)/f(\epsilon)$. Thus,

$$\mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{f(\epsilon)}\right) = \mathbb{P}\left(\hat{\mu}_{a,k} \geq r(\epsilon)\right) \leq e^{-kd(r(\epsilon),\mu_a)}$$

and

$$\begin{aligned}
\sum_{s=K_T}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{f(\epsilon)}\right) &\leq \sum_{s=K_T}^{+\infty} \sum_{k=c_a^- s}^{+\infty} e^{-kd(r(\epsilon),\mu_a)} \\
&= \frac{1}{1 - e^{-d(r(\epsilon),\mu_a)}} \sum_{s=K_T}^{+\infty} e^{-c_a^- s d(r(\epsilon),\mu_a)} \\
&= \frac{e^{-c_a^- d(r(\epsilon),\mu_a) K_T}}{\left(1 - e^{-d(r(\epsilon),\mu_a)}\right)\left(1 - e^{-c_a^- d(r(\epsilon),\mu_a)}\right)} \\
&\leq \frac{H(\epsilon)}{T^{\beta(\epsilon)}},
\end{aligned}$$

where $H(\epsilon) = \left[\left(1 - e^{-d(r(\epsilon),\mu_a)}\right)\left(1 - e^{-c_a^- d(r(\epsilon),\mu_a)}\right)\right]^{-1}$ and $\beta(\epsilon) = f(\epsilon)d(r(\epsilon), \mu_a)/d(\mu_a, \tau_a)$. ■

### A.4. Proof of Theorem 4

We first recall that the posterior distribution on the mean of a distribution belonging to an exponential family only depends on the number of observations $n$ and the empirical mean $x$ (see e.g. Lemma 1 in Kaufmann (2016)): for a given arm $a \in \{1, \ldots, K\}$, we denote this posterior by $\pi_{a,n,x}$. Given two constants $\mu_0^- > \mu^-$ and $\mu_0^+ < \mu^+$ verifying $\mu_0^- \leq \mu_a \leq \mu_0^+$ for all arms $a \in \{1, \ldots, K\}$, we define the truncated empirical mean: $\bar{\mu}_a(t) = \min(\max(\hat{\mu}_a(t), \mu_0^-), \mu_0^+)$. Then, for any arm $a \in \{1, \ldots, K\}$ and time step $t \geq 1$, the posterior distribution involved in Bayes-UCB-4P defines as follows:

$$\pi_{a,t} = \pi_{a,N_a(t),\bar{\mu}_a(t)}.$$

(i). Let $a \in \{1, \ldots, K\} \setminus \mathcal{A}^*$ be a non-profitable arm (i.e. $\mu_a < \tau_a$). We upper bound the expectation of $N_a(T)$ as follows:

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} C_a(t)\mathbb{I}\{a \in A_t\}\right] \leq c_a^+ \,\mathbb{E}\left[1 + \sum_{t=1}^{T-1} \mathbb{I}\{q_a(t) \geq \tau_a\}\right]$$

$$= c_a^+ \,\mathbb{E}\left[1 + \sum_{t=1}^{T-1} \mathbb{I}\left\{\pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c}, a \in A_{t+1}\right\}\right]$$

$$\leq c_a^+ \,\mathbb{E}\Bigg[1 + \sum_{t=1}^{T-1} \mathbb{I}\left\{\bar{\mu}_a(t) < \tau_a, \pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c}, a \in A_{t+1}\right\} \tag{9}$$

$$+ \sum_{t=1}^{T-1} \mathbb{I}\{\bar{\mu}_a(t) \geq \tau_a, a \in A_{t+1}\}\Bigg]. \tag{10}$$

Using Lemma 4 in Kaufmann (2016), the first sum in (9) is upper bounded by

$$\sum_{t=1}^{T-1} \mathbb{I}\left\{B\sqrt{N_a(t)}e^{-N_a(t)d^+(\bar{\mu}_a(t),\tau_a)} \geq \frac{1}{t(\log t)^c}, a \in A_{t+1}\right\}$$

$$= \sum_{t=1}^{T-1} \mathbb{I}\{a \in A_{t+1}\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{a \in \bigcap_{i \in \{i_1,\ldots,i_s\}} A_i, a \notin \bigcup_{i \in \{1,\ldots,t\}\setminus\{i_1,\ldots,i_s\}} A_i\right\}$$

$$\times\, \mathbb{I}\left\{B\sqrt{C_a(i_1) + \cdots + C_a(i_s)}e^{-(C_a(i_1)+\cdots+C_a(i_s))d^+\left(\bar{\mu}_{a,C_a(i_1)+\cdots+C_a(i_s)},\tau_a\right)} \geq \frac{1}{t(\log t)^c}\right\}, \tag{11}$$

where $B$ is a constant depending on $\mu_0^-$, $\mu_0^+$ and on prior densities. Then we upper bound the last indicator function appearing in Eq. (11) by

$$\mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, kd^+(\bar{\mu}_{a,k}, \tau_a) \leq \log T + c\log\log T + \frac{1}{2}\log k + \log B\right\}$$

$$\leq \mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, kd^+(\hat{\mu}_{a,k}, \tau_a) \leq \log T + c\log\log T + \frac{1}{2}\log k + \log B\right\}$$

$$+ \mathbb{I}\{\hat{\mu}_{a,k} < \mu_0^-\}. \tag{12}$$

We are now able to upper bound the right-hand side expression in Eq. (11) by injecting Eq. (12) and switching the sums on indices $t$ and $s$, which leads to

$$\sum_{t=1}^{T-1} \mathbb{I}\left\{\bar{\mu}_a(t) < \tau_a, \pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c}, a \in A_{t+1}\right\}$$

$$\leq K_T - 1 + \sum_{s=1}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{s \geq K_T, kd^+(\hat{\mu}_{a,k}, \tau_a) \leq \log T + c\log\log T + \frac{1}{2}\log k + \log B\right\}$$

$$+ \mathbb{I}\{\hat{\mu}_{a,k} < \mu_0^-\}, \tag{13}$$

where we used the same argument as in Eq. (8) to get rid of the sum over $t$.

Given $\epsilon \in ]0,1[$ we define $K_T = \left\lceil \frac{1+\epsilon}{1-\epsilon} \frac{\log T + c \log\log T}{c_a^- d(\mu_a, \tau_a)} \right\rceil$ and denote by $N_a(\epsilon)$ the constant such that $T \geq N_a(\epsilon)$ implies:

$$K_T \geq \left\lceil \frac{3}{c_a^-} \right\rceil \quad \text{and} \quad \frac{1}{c_a^- K_T} \left( \frac{1}{2} \log(c_a^- K_T) + \log(B) \right) \leq \frac{\epsilon}{1+\epsilon} d(\mu_a, \tau_a), \tag{14}$$

where the first inequality ensures that for all $k \geq c_a^- K_T$, the function $k \mapsto \log(x)/x$ decreases. Hence, the first indicator function appearing in the right-hand side in Eq. (13) is upper bounded by

$$\mathbb{I}\left\{ s \geq K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{1-\epsilon}{1+\epsilon} d(\mu_a, \tau_a) \right\}. \tag{15}$$

By combining equations (9), (13) and (15) we obtain

$$\begin{aligned} \mathbb{E}[N_a(T)] \leq c_a^+ \Bigg\{ K_T + \sum_{s=K_T}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}\left( d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{1-\epsilon}{1+\epsilon} d(\mu_a, \tau_a) \right) \\ + \sum_{s=1}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} < \mu_0^-) + \sum_{t=1}^{T-1} \mathbb{P}(\bar{\mu}_a(t) \geq \tau_a, a \in A_{t+1}) \Bigg\}, \end{aligned} \tag{16}$$

where the first sum can be upper bounded by $H_3(\epsilon) T^{-\beta_2(\epsilon)}$ with $H_3(\epsilon) > 0$ and $\beta_2(\epsilon) > 0$ thanks to Lemma 6. We upper bound the second sum in Eq. (16) with Chernoff inequality:

$$\begin{aligned} \sum_{s=1}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} < \mu_0^-) &\leq \sum_{s=1}^{+\infty} \sum_{k=c_a^- s}^{+\infty} e^{-k d(\mu_0^-, \mu_a)} \\ &= \frac{e^{-c_a^- d(\mu_0^-, \mu_a)}}{\left(1 - e^{-d(\mu_0^-, \mu_a)}\right)\left(1 - e^{-c_a^- d(\mu_0^-, \mu_a)}\right)}. \end{aligned}$$

Finally, we upper bound the third sum in Eq. (16) by

$$\begin{aligned} &\mathbb{E}\left[ \sum_{t=1}^{T-1} \mathbb{I}\{\hat{\mu}_{a,s} \geq \tau_a, a \in A_{t+1}\} \right] \\ &\leq \mathbb{E}\Bigg[ \sum_{t=1}^{T-1} \mathbb{I}\{a \in A_{t+1}\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1, \ldots, i_s\}} A_i, a \notin \bigcup_{i \in \{1, \ldots, t\} \setminus \{i_1, \ldots, i_s\}} A_i \right\} \\ &\qquad \times \mathbb{I}\left\{ \hat{\mu}_{a, C_a(i_1) + \cdots + C_a(i_s)} \geq \tau_a \right\} \Bigg] \\ &\leq \sum_{s=1}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} \geq \tau_a) \leq \frac{e^{-c_a^- d(\tau_a, \mu_a)}}{\left(1 - e^{-d(\tau_a, \mu_a)}\right)\left(1 - e^{-c_a^- d(\tau_a, \mu_a)}\right)}, \end{aligned} \tag{17}$$

where we respectively used Eq. (8) and Chernoff inequality in the two last inequalities.

(ii). Now consider $a \in A^*$. We have,

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T-1} C_a(t+1)\mathbb{I}\{a \notin A_{t+1}\}\right] = c_a^+ \sum_{t=1}^{T-1} \mathbb{P}(q_a(t) < \tau_a)$$
$$\leq c_a^+ \left\{ t_0 - 1 + \sum_{t=t_0}^{T-1} \mathbb{P}\left(\hat{\mu}_a(t) < \tau_a, N_a(t) \geq (\log t)^2\right) + \sum_{t=1}^{T-1} \mathbb{P}\left(q_a(t) < \tau_a, N_a(t) \leq (\log t)^2\right) \right\}, \tag{18}$$

where $t_0 = \max(t_1, t_2)$ with $t_1$ the smallest integer verifying $C^2 t_0 (\log t_0)^{2c} \geq 1$, which implies for all $t \geq t_1$ that $\bar{\mu}_a(t) \leq q_a(t)$, and $t_2 = \lceil \exp(2/d(\tau_a, \mu_a)) \rceil$ to ensure that $d(\tau_a, \mu_a)(\log t)^2 \geq 2 \log t$ for all $t \geq t_2$. To upper bound the first sum in Eq. (18) we write for $t \geq t_0$,

$$\mathbb{P}\left(\hat{\mu}_a(t) < \tau_a, N_a(t) \geq (\log t)^2\right) \leq \sum_{s=\lceil (\log t)^2 \rceil}^{t} \mathbb{P}(\hat{\mu}_{a,s} < \tau_a) \leq \sum_{s=\lceil (\log t)^2 \rceil}^{+\infty} e^{-sd(\tau_a, \mu_a)}$$
$$\leq e^{-d(\tau_a,\mu_a)(\log t)^2} \leq \frac{1}{t^2}.$$

To upper bound the second sum in Eq. (18) use again Lemma 4 in Kaufmann (2016),

$$\mathbb{P}\left(q_a(t) < \tau_a, N_a(t) \leq (\log t)^2\right) = \mathbb{P}\left(\pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a, \mu^+[) < \frac{1}{t(\log t)^c}, N_a(t) \leq (\log t)^2\right)$$
$$\leq \mathbb{P}\left(\frac{Ae^{-N_a(t)d(\bar{\mu}_a(t),\tau_a)}}{N_a(t)} < \frac{1}{t(\log t)^c}, N_a(t) \leq (\log t)^2\right)$$
$$= \mathbb{P}\left(N_a(t)d^+(\hat{\mu}_a(t), \tau_a) > \log\left(\frac{At(\log t)^c}{N_a(t)}\right), N_a(t) \leq (\log t)^2\right)$$
$$\leq \mathbb{P}\left(N_a(t)d^+(\hat{\mu}_a(t), \tau_a) > \log(At) + (c-2)\log\log t\right),$$

where $A$ is a constant depending on $\mu_0^-$, $\mu_0^+$ and on prior densities. Then for $c \geq 5$, using the self-normalized deviation inequality stated in Lemma 10 in Cappé et al. (Jun. 2013), we have,

$$\mathbb{P}(N_a(t)d^+(\hat{\mu}_a(t), \tau_a) > \log(At) + (c-2)\log\log t) \leq (\delta_t \log(c_a^+ t) + 1)\exp(-\delta_t + 1)$$
$$= \frac{e((\log(t))^2 + (c-2)\log(t)\log\log(t) + \log(Ac_a^+)\log(t) + (c-2)\log(c_a^+)\log\log(t) + \log(A)\log(c_a^+) + 1)}{At(\log(t))^{c-2}}$$
$$\leq \frac{e(2(c-2)+4)}{At\log(t)},$$

where we assumed $t \geq t_a = \max(e/A, 3, A, c_a^+, Ac_a^+)$ to ensure the last inequality and that $\delta_t = \log(At) + (c-2)\log\log(t) > 1$. Then by summing over $t$,

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] \leq c_a^+ \left\{ t_a + \frac{e(2(c-2)+4)}{A} \sum_{t=3}^{T-1} \frac{1}{t\log t} \right\}$$
$$\leq c_a^+ \{e(2(c-2)+4)\log\log T + t_a + 1\}.$$

### A.5. Proof of Theorem 5

We first introduce some notations. Denote by $(X_{a,s})_{s\geq 1}$ i.i.d. samples from distribution $\nu_a$. Let $L(\theta) = (1/2)\min(1, \sup_x p(x|\theta))$ and for any $\delta_a > 0$,

$$E_{a,s} = \left( \exists s' \in \{1,\dots,s\}, p(X_{a,s'}|\theta_a) \geq L(\theta_a), \left| \frac{\sum_{u=1,u\neq s'}^{s} X_{a,u}}{s-1} - \mu_a \right| \leq \delta_a \right)$$

is an event where there is at least one 'likely' observation of arm $a$ (namely $X_{a,s'}$) and such that the empirical sufficient statistic is close to its true mean. We also define $E_a(t) = E_{a,N_a(t)}$.

**Remark 7** *In the definition of $E_{a,s}$, the 'likely' observation $X_{a,s'}$ is only needed for technical reasons when the Jeffreys prior $\pi_a(0)$ is improper (see Remark 8 in Korda et al. (2013) for further discussion).*

We now recall the Theorem 4 in Korda et al. (2013), an important result on the posterior concentration under the event $E_a(t)$.

**Lemma 8** *There exists problem-dependent constants $C_{1,a}$ and $N_{1,a}$ and a function $\Delta \mapsto C_{2,a}(\Delta)$ such that for $\delta_a \in ]0,1[$ and $\Delta > 0$ verifying $1 - \delta_a C_{2,a}(\Delta) > 0$, it holds whenever $N_a(t) \geq N_{1,a}$:*

$$\mathbb{P}\left(\mu(\theta_a(t)) \geq \mu_a + \Delta, E_a(t) | (X_{a,s})_{1\leq s\leq N_a(t)}\right) \leq C_{1,a} N_a(t) e^{-(N_a(t)-1)(1-\delta_a C_{2,a}(\Delta))d(\mu_a,\mu_a+\Delta)}$$

*and*

$$\mathbb{P}\left(\mu(\theta_a(t)) \leq \mu_a - \Delta, E_a(t) | (X_{a,s})_{1\leq s\leq N_a(t)}\right) \leq C_{1,a} N_a(t) e^{-(N_a(t)-1)(1-\delta_a C_{2,a}(\Delta))d(\mu_a,\mu_a-\Delta)}.$$

Thanks to these concentration inequalities we can derive bounds on the expected number of pulls of any arm.

For all arms $a \in \{1,\dots,K\}$ and $t \geq 1$, $\theta_a(t)$ is a r.v. sampled from the posterior distribution $\pi_a(t)$ on $\theta_a$ obtained after $N_a(t)$ observations. For all $s \geq 1$, we also denote by $\theta_{a,s}$ a r.v. sampled from the posterior distribution resulting from the first $s$ observations pulled from arm $a$ (with arbitrary choice when some of these random variables are pulled together), so that $\theta_a(t) = \theta_{a,N_a(t)}$.

We now prove Theorem 5.

(i). Let $a \in \{1,\dots,K\} \setminus \mathcal{A}^*$ be a non-profitable arm (i.e. $\mu_a < \tau_a$). We upper bound the expectation of $N_a(T)$ as follows:

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[C_a(t) \sum_{t=1}^{T} \mathbb{I}\{a \in A_t\}\right] \leq c_a^+ \left\{1 + \sum_{t=1}^{T-1} \mathbb{P}\left(a \in A_{t+1}, E_a(t)\right) + \mathbb{P}\left(a \in A_{t+1}, E_a(t)^c\right)\right\}. \tag{19}$$

First observe that the first sum in the right-hand side in Eq. (19) is equal to

$$\mathbb{E}\Bigg[\sum_{t=1}^{T-1} \mathbb{I}\left\{a \in A_{t+1}\right\} \sum_{s=1}^{t} \sum_{1\leq i_1<\dots<i_s\leq t} \mathbb{I}\left\{a \in \bigcap_{i\in\{i_1,\dots,i_s\}} A_i, a \notin \bigcup_{i\in\{1,\dots,t-1\}\setminus\{i_1,\dots,i_s\}} A_i\right\}$$
$$\times\, \mathbb{I}\left\{\mu(\theta_{a,C_a(i_1)+\dots+C_a(i_s)}) \geq \tau_a, E_{a,C_a(i_1)+\dots+C_a(i_s)}\right\}\Bigg].$$

Then, given $\epsilon \in ]0,1[$, by choosing $\delta_a \leq \epsilon / C_{2,a}(|\Delta_a|)$, defining $K_T = \left\lceil \frac{1+\epsilon}{1-\epsilon} \frac{\log T}{c_a^- d(\mu_a, \tau_a)} \right\rceil$ and observing that $\mathbb{I}\{\mu(\theta_{a,C_a(i_1)+\cdots+C_a(i_s)}) \geq \tau_a, E_{a,C_a(i_1)+\cdots+C_a(i_s)}\}$ is upper bounded by $\mathbb{I}\{s < K_T\} + \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\{s \geq K_T, \mu(\theta_{a,k}) \geq \tau_a, E_{a,k}\}$, we obtain:

$$\begin{aligned}
\sum_{t=1}^{T-1} \mathbb{P}\left(a \in A_{t+1}, E_a(t)\right) &\leq K_T - 1 + \sum_{s=K_T}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}\left(\mu(\theta_{a,k}) \geq \tau_a, E_{a,k}\right) \\
&\leq K_T - 1 + \sum_{s=K_T}^{T} \sum_{k=c_a^- s}^{c_a^+ s} C_{1,a} k e^{-(k-1)(1-\epsilon)d(\mu_a,\tau_a)} \\
&\leq \frac{1+\epsilon}{1-\epsilon} \frac{\log T}{c_a^- d(\mu_a, \tau_a)} + C_{1,a} T (c_a^+ K_T)^2 e^{-(c_a^- K_T - 1)(1-\epsilon)d(\mu_a,\tau_a)} \\
&\leq \frac{1+\epsilon}{1-\epsilon} \frac{\log T}{c_a^- d(\mu_a, \tau_a)} + C_{1,a} e^{(1-\epsilon)d(\mu_a,\tau_a)} \frac{(c_a^+ K_T)^2}{T^\epsilon},
\end{aligned}$$

where we used in the first inequality Eq. (8). In the second and third inequalities we assumed $T$ larger than $N_a(\epsilon)$ verifying $T \geq N_a(\epsilon) \Rightarrow K_T \geq \max(N_{1,a}/c_a^-, N_{2,a})$ with $N_{1,a}$ defined in Lemma 8 and $N_{2,a}$ such that the function $u \mapsto u^2 e^{-(c_a^- u - 1)(1-\epsilon)d(\mu_a,\tau_a)}$ is decreasing for $u \geq N_{2,a}$.

In order to upper bound the second sum in the right-hand side in Eq. (19) we first introduce the following events:

$$B_{a,s} = (\forall s' \in \{1, \ldots, s\}, p(X_{a,s'}|\theta_a) \leq L(\theta_a))$$

and

$$D_{a,s} = \left( \exists s' \in \{1, \ldots, s\}, \left| \frac{\sum_{u=1, u \neq s'}^{s} X_{a,u}}{s-1} - \mu_a \right| > \delta_a \right).$$

Then observing that $E_a(t)^c \subset B_{a,N_a(t)} \bigcup D_{a,N_a(t)}$ and it holds

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{P}(a \in A_{t+1}, E_a(t)^c) \\
&\leq \mathbb{E}\left[ \sum_{t=1}^{T-1} \mathbb{I}\left\{a \in A_{t+1}\right\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1, \ldots, i_s\}} A_i, a \notin \bigcup_{i \in \{1, \ldots, t-1\} \setminus \{i_1, \ldots, i_s\}} A_i \right\} \right. \\
&\qquad \left. \times \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{I}\left\{B_{a,k}\right\} + \mathbb{I}\left\{D_{a,k}\right\} \right] \\
&\leq \sum_{s=1}^{T} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}\left(B_{a,k}\right) + \mathbb{P}\left(D_{a,k}\right) \\
&\leq \sum_{s=1}^{+\infty} c_a^+ s \mathbb{P}\left(p(X_{a,1}|\theta_a) < L(\theta_a)\right)^{c_a^- s} + \left(c_a^+ s\right)^2 \left( e^{-(c_a^- s - 1)d(\mu_a - \delta_a, \mu_a)} + e^{-(c_a^- s - 1)d(\mu_a + \delta_a, \mu_a)} \right) < +\infty,
\end{aligned}$$

where we used Eq. (8) in the second inequality.

(ii). Now consider $a \in A^*$ i.e. verifying $\mu_a > \tau_a$. Let $b \in ]0,1[$, we have:

$$\tilde{C}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=2}^{T} C_a(t)\, \mathbb{I}\{a \notin A_t\}\right] \leq c_a^+ \sum_{t=1}^{T-1} \mathbb{P}(\mu(\theta_a(t)) < \tau_a)$$
$$\leq c_a^+ \left\{ \sum_{t=1}^{T-1} \mathbb{P}\left(\mu(\theta_a(t)) < \tau_a, E_a(t) \middle| N_a(t) > t^b\right) + \sum_{t=1}^{T-1} \mathbb{P}\left(E_a(t)^c \middle| N_a(t) > t^b\right) + \sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \right\}. \tag{20}$$

By applying Lemma 8, the first sum in Eq. (20) is upper bounded by

$$N_{0,a}^{1/b} + \sum_{t=\left\lceil N_{0,a}^{1/b} \right\rceil}^{+\infty} C_{1,a} t^b e^{-(t^b-1)(1-\delta_a C_{2,a}(|\Delta_a|))d(\mu_a,\tau_a)} < +\infty,$$

where $N_{0,a} = \max(N_{1,a}, N_{3,a})$ with $N_{3,a}$ such that the function $u \mapsto u e^{-(u-1)(1-\delta_a C_{2,a}(|\Delta_a|))d(\mu_a,\tau_a)}$ is decreasing for $u \geq N_{3,a}$.

By applying Chernoff inequality we upper bound the second sum in Eq. (20) by

$$\sum_{t=1}^{T-1} \mathbb{P}\left(E_a(t)^c \middle| N_a(t) > t^b\right) \leq \sum_{t=1}^{T} \sum_{s=\lceil t^b/c_a^+ \rceil}^{t} \sum_{k=c_a^- s}^{c_a^+ s} \mathbb{P}(B_{a,k}) + \mathbb{P}(D_{a,k})$$
$$\leq \sum_{t=1}^{+\infty} c_a^+ t^2 \mathbb{P}(p(X_{a,1}|\theta_a) \leq L(\theta_a))^{\frac{c_a^-}{c_a^+} t^b} + 2(c_a^+)^2 t^3 \left( e^{-\left(\frac{c_a^-}{c_a^+} t^b - 1\right) d(\mu_a - \delta_a, \mu_a)} + e^{-\left(\frac{c_a^-}{c_a^+} t^b - 1\right) d(\mu_a + \delta_a, \mu_a)} \right) < +\infty.$$

Finally we upper bound the third sum in Eq. (20) with the following result, inspired from Proposition 5 in Korda et al. (2013). In our case its proof is simpler as there are no dependencies between arms in the objective of the profitable bandit problem.

**Lemma 9** *For any profitable arm $a \in A^*$ and any $b \in ]0,1[$, there exists a problem-dependent constant $C_b < +\infty$ such that*

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \leq C_b.$$

Then, by using the Bernstein-Von-Mises theorem telling us that $\lim_{j \to +\infty} \mathbb{P}(\mu(\theta_a(\tau_j)) < \tau_a) = 0$, we deduce that there exists a constant $C \in ]0,1[$ such that for all $j \geq 0$, $\mathbb{P}(\mu(\theta_a(\tau_j)) < \tau_a) \leq C$. Hence,

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \leq \sum_{t=1}^{+\infty} \left(t^b + 1\right) C^{t^{1-b}-1} < +\infty.$$

### A.6. Proof of Lemma 9

In all this proof we consider a fixed profitable arm $a \in A^*$. We follow the lines of the proof of Proposition 5 in Korda et al. (2013) : let $t_j$ be the occurence of the $j$-th play of the arm $a$ (with $t_0 = 0$ by convention). Let $\xi_j = t_{j+1} - t_j - 1$, it corresponds to the number of time steps between the $j$-th and the $(j+1)$-th play of arm $a$. Hence, $t - N_a(t) \leq \sum_{j=0}^{N_a(t)} \xi_j$ and we have

$$\begin{aligned}
\mathbb{P}\left(N_a(t) \leq t^b\right) &\leq \mathbb{P}\left(\exists j \in \left\{0, \dots, \lfloor t^b \rfloor, \xi_j \geq t^{1-b} - 1\right\}\right) \\
&\leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}\left(\xi_j \geq t^{1-b} - 1\right) \\
&\leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}\left(\mu(\theta_a(\tau_j)) < \tau_a\right)^{t^{1-b}-1}.
\end{aligned}$$