JMLR: Workshop and Conference Proceedings vol 49:1–30, 2016

Optimal Best Arm Identification with Fixed Confidence

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR
Institut de Mathématiques de Toulouse; UMR5219
Université de Toulouse; CNRS
UPS IMT, F-31062 Toulouse Cedex 9, France

Emilie Kaufmann EMILIE.KAUFMANN@INRIA.FR
Univ. Lille, CNRS, UMR 9189
CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille
F-59000 Lille, France

Abstract

We give a complete characterization of the complexity of best-arm identification in one-parameter bandit problems. We prove a new, tight lower bound on the sample complexity. We propose the 'Track-and-Stop' strategy, which we prove to be asymptotically optimal. It consists in a new sampling rule (which tracks the optimal proportions of arm draws highlighted by the lower bound) and in a stopping rule named after Chernoff, for which we give a new analysis.

Keywords: multi-armed bandits, best arm identification, MDL.

1. Introduction

A multi-armed bandit model is a paradigmatic framework of sequential statistics made of K probability distributions $\nu_1, \dots, \nu_K$ with respective means $\mu_1, \dots, \mu_K$: at every time step $t = 1, 2, \dots$ one *arm* $A_t \in \mathcal{A} = \{1, \dots, K\}$ is chosen and a new, independent *reward* X_t is drawn from ν_{A_t}. Introduced in the 1930s with motivations originally from clinical trials, bandit models have raised a large interest recently as relevant models for interactive learning schemes or recommender systems. A large part of these works consisted in defining efficient strategies for maximizing the expected cumulated reward $\mathbb{E}[X_1 + \cdots + X_t]$; see for instance Bubeck and Cesa-Bianchi (2012) for a survey. A good understanding of this simple model has allowed for efficient strategies in much more elaborate settings, for example including side information (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013), infinitely many arms (Srinivas et al., 2010; Bubeck et al., 2011), or for the search of optimal strategies in games (Munos, 2014), to name just a few.

In some of these applications, the real objective is not to maximize the cumulated reward, but rather to identify the arm that yields the largest mean reward $\mu^* = \max_{1 \le a \le K} \mu_a$, as fast and accurately as possible, regardless of the number of bad arm draws. Let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ be the sigma-field generated by the observations up to time t. A strategy is then defined by:

- a *sampling rule* $(A_t)_t$, where A_t is $\mathcal{F}_{t-1}$-measurable;
- a *stopping rule* τ, which is a stopping time with respect to $\mathcal{F}_t$;
- and a $\mathcal{F}_\tau$-measurable *decision rule* $\hat{a}_\tau$.

The goal is to guarantee that $\hat{a}_\tau \in \text{argmax}\, \mu_a$ with the highest possible probability while minimizing the number τ of draws. Two settings have been considered in the literature. In the *fixed-budget setting*, the number of draws is fixed in advance, and one aims at minimizing the probability of error $\mathbb{P}(\hat{a}_\tau \notin \text{argmax}\, \mu_a)$. In the *fixed-confidence setting* a maximal risk δ is fixed, and one looks for a strategy guaranteeing that $\mathbb{P}(\hat{a}_\tau \notin \text{argmax}\, \mu_a) \leq \delta$ (such a strategy is called *δ-PAC*) while minimizing the *sample complexity* $\mathbb{E}[\tau]$.

The aim of this paper is to propose an analysis of the best arm identification problem in the fixed-confidence setting. For the sake of clarity and simplicity, we suppose that there is a single arm with highest expectation, and without loss of generality that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. We also focus on the simple case where the distributions are parameterized by their means, as in one-parameter exponential families, and we index probabilities and expectations by the parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. The Kullback-Leibler divergence of two distributions of means μ_1 and μ_2 is a function $d : (\mu_1, \mu_2) \to \mathbb{R}^+$. Two cases are of particular interest: the Gaussian case, with $d(x, y) = (x - y)^2/(2\sigma^2)$, and binary rewards with $d(x, y) = \text{kl}(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$.

Several strategies have been proposed to minimize $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$. While racing strategies are based on successive eliminations of apparently sub-optimal arms (Even-Dar et al. (2006); Kaufmann and Kalyanakrishnan (2013)), another family of strategies exploits the use of upper and lower confidence bounds on the means of the arms (Kalyanakrishnan et al. (2012); Gabillon et al. (2012); Kaufmann and Kalyanakrishnan (2013); Jamieson et al. (2014)). They reflect some aspects of the difficulty of the problem, but are not proved to satisfy any optimality property[1]. In particular, there was still a gap between the lower bounds, involving complexity terms reflecting only partially the structure of the problem, and the upper bounds on $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ for these particular algorithms, even from an asymptotic point of view. For the particular case $K = 2$, this gap was closed in Kaufmann et al. (2014). The tools used there, however, where specific to the two-arm case and cannot be extended easily.

The first result of this paper is a tight, non-asymptotic lower bound on $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$. This bound involves a 'characteristic time' for the problem, depending on the parameters of the arms, which does not take a simple form like for example a sum of squared inverse gaps. Instead, it appears as the solution of an optimization problem, in the spirit of the bounds given by Graves and Lai (1997) in the context of regret minimization. We give a brief analysis of this optimization problem, and we provide an efficient numerical solution as well as elements of interpretation.

The second contribution is a new δ-PAC algorithm that asymptotically achieves this lower bound, that we call the Track-and-Stop strategy. In a nutshell, the idea is to sample so as to *equalize the probability of all possible wrong decisions*, and to stop as soon as possible. The stopping rule, which we name after Chernoff, can be interpreted in three equivalent ways: in statistical terms, as a generalized likelihood ratio test; in information-theoretic terms, as an application of the Minimal Description Length principle; and in terms of optimal transportation, in light of the lower bound. The sampling rule is a by-product of the lower bound analysis, which reveals the existence of optimal proportions of draws for each arm. By estimating and tracking these proportions, our algorithm asymptotically reaches the optimal sample complexity as δ goes to 0.

The paper is organized as follows. In Section 2, the lower bound is given with a short proof. Section 3 contains a commented description of our stopping and sampling rules. The analysis of the resulting algorithm is sketched in Sections 4 (validity of the stopping rule) and 5 (sample complexity analysis), establishing its asymptotic optimality. Section 6 contains practical comments and results of numerical experiments, in order to show the efficiency of our strategy even for moderate values

[1] Optimality is mentioned in several articles, with different and sometimes weak meanings (minimax, rate-optimal,...). In our view, BAI algorithms for which there exists a model with a sample complexity bounded, up to a multiplicative constant, by some quantity related to some lower bound, may not be called optimal.

of δ. We also briefly comment on the gain over racing strategies, which can be explained in light of Theorem 1. Most proofs and technical details are postponed to the Appendix.

2. Lower Bounds on the sample complexity

The pioneering work of Lai and Robbins (1985) has popularized the use of changes of distributions to show problem-dependent lower bounds in bandit problems: the idea is to move the parameters of the arms until a completely different behavior of the algorithm is expected on this alternative bandit model. The cost of such a transportation is induced by the deviations of the arm distributions: by choosing the most economical move, one can prove that the alternative behavior is not too rare in the original model. Recently, Kaufmann et al. (2014) and Combes and Proutière (2014) have independently introduced a new way of writing such a change of measure, which relies on a transportation lemma that encapsulates the change of measure and permits to use it at a higher level. Here we go one step further by combining *several changes of measures at the same time*, in the spirit of Graves and Lai (1997). This allows us to prove a non-asymptotic lower bound on the sample complexity valid for any δ-PAC algorithm on any bandit model with a unique optimal arm. We present this result in the particular case of arms that belong to a canonical exponential family,

$$\mathcal{P} = \left\{ (\nu_\theta)_{\theta\in\Theta} : \frac{d\nu_\theta}{d\xi} = \exp\big(\theta x - b(\theta)\big) \right\},$$

where $\Theta \subset \mathbb{R}$, ξ is some reference measure on $\mathbb{R}$ and $b : \Theta \to \mathbb{R}$ is a convex, twice differentiable function. A distribution $\nu_\theta \in \mathcal{P}$ can be parameterized by its mean $\dot{b}(\theta)$, and for every $\mu \in \dot{b}(\Theta)$ we denote by ν^μ be the unique distribution in $\mathcal{P}$ with expectation μ. Unfamiliar readers may simply think of Bernoulli laws, or Gaussian distributions with known variance. As explained in Cappé et al. (2013, see also references therein), the Kullback-Leibler divergence from ν_θ to $\nu_{\theta'}$ induces a divergence function d on $\dot{b}(\Theta)$ defined, if $\dot{b}(\theta) = \mu$ and $\dot{b}(\theta') = \mu'$, by

$$d(\mu,\mu') = \mathrm{KL}(\nu^\mu, \nu^{\mu'}) = \mathrm{KL}(\nu_\theta, \nu_{\theta'}) = b(\theta') - b(\theta) - \dot{b}(\theta)(\theta' - \theta)\ .$$

With some abuse of notation, an exponential family bandit model $\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$ is identified with the means of its arms $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$.

2.1. General Lower Bound

Denote by $\mathcal{S}$ a set of exponential bandit models such that each bandit model $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ in $\mathcal{S}$ has a unique optimal arm: for each $\boldsymbol{\mu} \in \mathcal{S}$, there exists an arm $a^*(\boldsymbol{\mu})$ such that $\mu_{a^*(\boldsymbol{\mu})} > \max\{\mu_a : a \neq a^*(\boldsymbol{\mu})\}$. Fixed-confidence strategies depend on the risk level and we subscript stopping rules by δ. A strategy is called δ-PAC if for every $\boldsymbol{\mu} \in \mathcal{S}$, $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty) = 1$ and $\mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_{\tau_\delta} \neq a^*) \leq \delta$. We introduce

$$\mathrm{Alt}(\boldsymbol{\mu}) := \{\boldsymbol{\lambda} \in \mathcal{S} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}\ ,$$

the set of problems where the optimal arm is not the same as in $\boldsymbol{\mu}$, and $\Sigma_K = \{\omega \in \mathbb{R}_+^k : \omega_1 + \cdots + \omega_K = 1\}$ the set of probability distributions on $\mathcal{A}$.

Theorem 1 *Let $\delta \in (0,1)$. For any δ-PAC strategy and any bandit model $\boldsymbol{\mu} \in \mathcal{S}$,*

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^*(\boldsymbol{\mu})\, \mathrm{kl}(\delta, 1-\delta),$$

where

$$T^*(\boldsymbol{\mu})^{-1} := \sup_{w\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a,\lambda_a)\right) . \tag{1}$$

Remark 2 *As* $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$ *when* δ *goes to zero, Theorem 1 yields the asymptotic lower bound*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq T^*(\boldsymbol{\mu}).$$

A non-asymptotic version can be obtained for example from the inequality $\mathrm{kl}(\delta, 1-\delta) \geq \log(1/(2.4\delta))$ *that holds for all* $\delta \in (0,1)$*, given in Kaufmann et al. (2015).*

We will see that the supremum in Equation (1) is indeed a maximum, and we call

$$w^*(\boldsymbol{\mu}) := \operatorname*{argmax}_{w\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a,\lambda_a)\right)$$

the corresponding distribution on the arms. The proof of Theorem 1 shows that w^* is the proportion of arm draws of any strategy matching this lower bound.

The particular case where $\mathcal{S}$ is the class of bandit models with Poisson rewards in which all suboptimal arms are equal is considered in the very insightful paper by Vaidhyan and Sundaresan (2015), where a closed-form formula is given for both $T^*(\boldsymbol{\mu})$ and $w^*(\boldsymbol{\mu})$. In this paper, we consider best arm identification in all possible bandit models with a single best arm, and in the rest of the paper we fix

$$\mathcal{S} := \left\{\boldsymbol{\mu} : \exists a \in \mathcal{A} : \mu_a > \max_{i\neq a} \mu_i \right\}.$$

Proof of Theorem 1. Let $\delta \in (0,1)$, $\boldsymbol{\mu} \in \mathcal{S}$, and consider a δ-PAC strategy. For every $t \geq 1$, denote by $N_a(t)$ the (random) number of draws of arm a up to time t. The 'transportation' Lemma 1 of Kaufmann et al. (2015) relates the expected number of draws of each arm and the Kullback-Leibler divergence of two bandit models with different optimal arms to the probability of error δ:

$$\forall \lambda \in \mathcal{S} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu}), \quad \sum_{a=1}^{K} d(\mu_a,\lambda_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)] \geq \mathrm{kl}(\delta, 1-\delta). \tag{2}$$

Instead of choosing for each arm a a specific instance of $\boldsymbol{\lambda}$ that yields a lower bound on $\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]$, we combine here the inequalities given by all alternatives $\boldsymbol{\lambda}$:

$$\begin{aligned}\mathrm{kl}(\delta, 1-\delta) &\leq \inf_{\lambda\in\mathrm{Alt}(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \left(\sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]} d(\mu_a,\lambda_a)\right) = \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \inf_{\lambda\in\mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]} d(\mu_a,\lambda_a)\right) \\ &\leq \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \sup_{w\in\Sigma_K} \inf_{\lambda\in\mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a,\lambda_a)\right),\end{aligned}$$

as $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] = \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau_\delta)]$.

□

In the last inequality, the strategy-dependent proportions of arm draws are replaced by their supremum so as to obtain a bound valid for any δ-PAC algorithm; one can see that a strategy may reach this bound only if it meets the $w^*(\boldsymbol{\mu})$. To make this bound useful, it remains to study T^* and w^*.

2.2. About the Characteristic Time and the Optimal Proportions

We study here the optimization problem (1), so as to better understand the function T^* and w^* (Proposition 6), and in order to provide an efficient algorithm for computing first $w^*(\boldsymbol{\mu})$ (Theorem 5), then $T^*(\boldsymbol{\mu})$ (Lemma 3). The main ideas are outlined here, while all technical details are postponed to Appendix A. Simplifying $T^*(\boldsymbol{\mu})$ requires the introduction of the following parameterized version of the Jensen-Shannon divergence (which corresponds to $\alpha = 1/2$): for every $\alpha \in [0,1]$, let

$$I_\alpha(\mu_1,\mu_2) := \alpha d\big(\mu_1, \alpha\mu_1 + (1-\alpha)\mu_2\big) + (1-\alpha) d\big(\mu_2, \alpha\mu_1 + (1-\alpha)\mu_2\big) \ . \tag{3}$$

The first step is, for any w, to identify the minimizer of the transportation cost:

Lemma 3 *For every $w \in \Sigma_K$,*

$$\inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^K w_a d(\mu_a,\lambda_a)\right) = \min_{a\neq 1} \, (w_1+w_a) I_{\frac{w_1}{w_1+w_a}}(\mu_1,\mu_a) \ .$$

It follows that

$$\begin{aligned} T^*(\boldsymbol{\mu})^{-1} &= \sup_{w\in\Sigma_K} \min_{a\neq 1} \, (w_1+w_a) I_{\frac{w_1}{w_1+w_a}}(\mu_1,\mu_a) \text{ , and} \\ w^*(\boldsymbol{\mu}) &= \operatorname*{argmax}_{w\in\Sigma_K} \min_{a\neq 1} \, (w_1+w_a) I_{\frac{w_1}{w_1+w_a}}(\mu_1,\mu_a) \ . \end{aligned}$$

It is easy to see that, at the optimum, the quantities $(w_1+w_a)I_{w_1/(w_1+w_a)}(\mu_1,\mu_a)$ are all equal.

Lemma 4 *For all $a,b \in \{2,\dots,K\}$,*

$$(w_1^*+w_a^*) I_{\frac{w_1^*}{w_1^*+w_a^*}}(\mu_1,\mu_a) = (w_1^*+w_b^*) I_{\frac{w_1^*}{w_1^*+w_b^*}}(\mu_1,\mu_b) \ .$$

This permits to obtain a more explicit formula for $w^*(\boldsymbol{\mu})$ involving only a single real parameter. Indeed, for every $a \in \{2,\dots K\}$ let

$$g_a(x) = (1+x) I_{\frac{1}{1+x}}(\mu_1,\mu_a) \ . \tag{4}$$

The function g_a is a strictly increasing one-to-one mapping from $[0,+\infty[$ onto $[0, d(\mu_1,\mu_a)[$. We define $x_a : [0, d(\mu_1,\mu_a)[\to [0,+\infty[$ as its inverse function: $x_a(y) = g_a^{-1}(y)$. Denoting x_1 the function constantly equal to 1, one obtains the following characterization of $w^*(\boldsymbol{\mu})$:

Theorem 5 *For every $a \in \mathcal{A}$,*

$$w_a^*(\boldsymbol{\mu}) = \frac{x_a(y^*)}{\sum_{a=1}^K x_a(y^*)} \ , \tag{5}$$

where y^* is the unique solution of the equation $F_{\boldsymbol{\mu}}(y) = 1$, and where

$$F_{\boldsymbol{\mu}} : y \mapsto \sum_{a=2}^K \frac{d\left(\mu_1, \frac{\mu_1 + x_a(y)\mu_a}{1+x_a(y)}\right)}{d\left(\mu_a, \frac{\mu_1 + x_a(y)\mu_a}{1+x_a(y)}\right)} \tag{6}$$

is a continuous, increasing function on $[0, d(\mu_1,\mu_2)[$ such that $F_{\boldsymbol{\mu}}(0) = 0$ and $F_{\boldsymbol{\mu}}(y) \to \infty$ when $y \to d(\mu_1,\mu_2))$.

Thus, w^* can be simply computed by applying (for example) the bisection method to a function whose evaluations requires the resolution of K smooth scalar equations. By using efficient numerical solvers, we obtain a fast algorithm of complexity, roughly speaking, proportional to the number of arms. This characterization of $w^*(\boldsymbol{\mu})$ also permits to obtain a few sanity-check properties, like for example:

Proposition 6

1. *For all $\boldsymbol{\mu} \in \mathcal{S}$, for all a, $w_a^*(\boldsymbol{\mu}) \neq 0$.*
2. *w^* is continuous in every $\boldsymbol{\mu} \in \mathcal{S}$.*
3. *If $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$, one has $w_2^*(\boldsymbol{\mu}) \geq \cdots \geq w_K^*(\boldsymbol{\mu})$.*

Observe that one may have[2] $w_2 > w_1$. In general, it is not possible to give closed-form formulas for $T^*(\boldsymbol{\mu})$ and $w^*(\boldsymbol{\mu})$. In particular, $T^*(\boldsymbol{\mu})$ cannot be written as a sum over the arms of individual complexity terms as in previous works (Mannor and Tsitsiklis (2004); Kaufmann et al. (2015)). But the following particular cases can be mentioned.

Two-armed bandits. For a two-armed bandit model $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $T^*(\boldsymbol{\mu})$ and $w^*(\boldsymbol{\mu})$ can be computed algebraically. Lemma 3 and the fact that $w_2 = 1 - w_1$ imply that

$$T^*(\boldsymbol{\mu})^{-1} \quad = \quad \sup_{\alpha \in (0,1)} I_\alpha(\mu_1, \mu_a),$$

Some algebra shows that the maximum is reached at $\alpha = \alpha_*(\mu_1, \mu_2)$ defined by the equation $d(\mu_1, \mu_*) = d(\mu_2, \mu_*)$, where $\mu_* = \alpha_*(\mu_1, \mu_2)\mu_1 + (1 - \alpha_*(\mu_1, \mu_2))\mu_2$. The value of the maximum is then the 'reversed' Chernoff information $d_*(\mu_1, \mu_2) := d(\mu_1, \mu_*) = d(\mu_2, \mu_*)$. This permits to recover the bound already given in Kaufmann et al. (2014):

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \frac{\mathrm{kl}(\delta, 1-\delta)}{d_*(\mu_1, \mu_2)} \ .$$

The Gaussian case. When $d(x, y) = (x - y)^2/(2\sigma^2)$, $T^*(\mu)$ and $w^*(\boldsymbol{\mu})$ can be computed by solving a rational equation. Indeed, Equation (6) and Lemma 4 imply that

$$\sum_{a=2}^{K} x_a^2 = 1 \quad \text{and} \quad \frac{x_a}{1 + x_a} = \frac{\lambda}{(\mu_1 - \mu_a)^2}, \quad \text{thus} \quad \sum_{a=2}^{K} \frac{\lambda^2}{\left((\mu_1 - \mu_a)^2 - \lambda\right)^2} = 1$$

for some $\lambda \in (0, (\mu_1 - \mu_2)^2)$. For $K = 3$, λ is the solution of a polynomial equation of degree 4 and has therefore an (obscure) algebraic expression. The following inequalities, established in Appendix A.4 give a better insight of the order of magnitude of $T^*(\boldsymbol{\mu})$: if $\Delta_1 = \Delta_2$ and $\Delta_a = \mu_1 - \mu_a$ for $a \geq 2$, then

$$\sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2} \leq T^*(\boldsymbol{\mu}) \leq 2 \sum_{a=1}^{K} \frac{2\sigma^2}{\Delta_a^2}.$$

[2] This can happen when the right-deviations of ν^{μ_2} are smaller than the left-deviations of ν^{μ_1}; for example, with Bernoulli arms of parameters $\boldsymbol{\mu} = (0.5, 0.1, 0.02)$, $\nu^*(\boldsymbol{\mu}) \approx (39\%, 42\%, 19\%)$.

3. The Track-and-Stop Strategy

We now describe a new strategy which is the first (as far as we know) to asymptotically match the lower bound of Theorem 1. Denote by $\hat{\mu}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_K(t))$ the current maximum likelihood estimate of $\boldsymbol{\mu}$ at time t: $\hat{\mu}_a(t) = N_a(t)^{-1} \sum_{s \leq t} X_s \mathbb{1}\{A_s = a\}$. As seen in Section 2, a good sampling rule should respect the optimal proportions of arm draws given by $w^*(\boldsymbol{\mu})$. There are several ways to ensure this, and we present two of them in Section 3.1. A good stopping rule should determine the earliest moment when sufficient statistical evidence has been gathered for identifying the best arm: we propose one (with several interpretations) in Section 3.2, showing in Section 4 how to tune it in order to ensure the δ-PAC property. As for the decision rule, we simply choose $\hat{a}_{\tau_\delta} = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ \hat{\mu}_a(\tau_\delta)$. The optimality of the Track-and-Stop strategy is shown in Section 5.

3.1. Sampling Rule: Tracking the Optimal Proportions

The first idea for matching the proportions $w^*(\boldsymbol{\mu})$ is to track the plug-in estimates $w^*(\hat{\mu}(t))$. In bandit settings, using plug-in estimates is always hazardous, because bad initial estimate may lead to abandon an arm and to prevent further observations that would correct the initial error. Indeed, one may see (both theoretically and in numerical experiments) that a naive plug-in sampling rule sometimes fails. But there is a very simple fix, which consists in forcing sufficient exploration of each arm to ensure a (sufficiently fast) convergence of $\hat{\mu}(t)$.

The shortest way to do this (which we term *C-Tracking*) is to slightly alter the optimization solution: for every $\epsilon \in (0, 1/K]$, let $w^\epsilon(\boldsymbol{\mu})$ be a L^∞ projection of $w^*(\boldsymbol{\mu})$ onto $\Sigma_K^\epsilon = \big\{(w_1, \dots, w_K) \in [\epsilon, 1] : w_1 + \dots + w_K = 1\big\}$. Choosing $\epsilon_t = (K^2 + t)^{-1/2}/2$ and

$$A_{t+1} \in \underset{1 \leq a \leq K}{\operatorname{argmax}} \sum_{s=0}^{t} w_a^{\epsilon_s}(\hat{\boldsymbol{\mu}}(s)) - N_a(t) \ ,$$

we prove in Appendix B that:

Lemma 7 *For all $t \geq 1$ and $a \in \mathcal{A}$, the C-Tracking rule ensures that $N_a(t) \geq \sqrt{t + K^2} - 2K$ and that*

$$\max_{1 \leq a \leq K} \left| N_a(t) - \sum_{s=0}^{t-1} w_a^*(\hat{\boldsymbol{\mu}}(s)) \right| \leq K(1 + \sqrt{t}) \ .$$

It is slightly more efficient in practice to target directly $w^*(\hat{\boldsymbol{\mu}}(t))$, and to force exploration steps whenever an arm is in deficit. Introducing $U_t = \{a : N_a(t) < \sqrt{t} - K/2\}$, the D-Tracking rule (A_t) is sequentially defined as

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\ N_a(t) \text{ if } U_t \neq \emptyset & \textit{(forced exploration)} \\ \underset{1 \leq a \leq K}{\operatorname{argmax}}\ t\, w_a^*(\hat{\boldsymbol{\mu}}(t)) - N_a(t) & \textit{(direct tracking)} \end{cases}$$

Lemma 8 *The D-Tracking rule ensures that $N_a(t) \geq (\sqrt{t} - K/2)_+ - 1$ and that for all $\epsilon > 0$, for all t_0, there exists $t_\epsilon \geq t_0$ such that*

$$\sup_{t \geq t_0} \max_a |w_a^*(\hat{\boldsymbol{\mu}}(t)) - w_a^*(\boldsymbol{\mu})| \leq \epsilon \quad \Rightarrow \quad \sup_{t \geq t_\epsilon} \max_a \left| \frac{N_a(t)}{t} - w_a^*(\boldsymbol{\mu}) \right| \leq 3(K-1)\epsilon \ .$$

It is guaranteed under all these sampling rules that the empirical proportion of draws of each arm converges to the optimal proportion, as proved in Appendix B.3.

Proposition 9 *The C-Tracking and D-Tracking sampling rules both satisfy*

$$\mathbb{P}_w\left(\lim_{t\to\infty}\frac{N_a(t)}{t}=w_a^*(\boldsymbol{\mu})\right)=1\,.$$

We actually have a little more: a minimal convergence speed of $\hat{\mu}$ to μ, which proves useful in the analysis of the *expected* sample complexity. Of course, other tracking strategies are possible, like for example the one introduced in Antos et al. (2008) for the uniform estimation of all arms' expectations.

3.2. Chernoff's Stopping Rule

From a statistical point of view, the question of stopping at time t is a more or less classical statistical test: do the past observations allow to assess, with a risk at most δ, that one arm is larger than the others? For all arms $a, b \in \mathcal{A}$, we consider the Generalized Likelihood Ratio statistic

$$Z_{a,b}(t) := \log\frac{\max_{\mu'_a\geq\mu'_b} p_{\mu'_a}\left(\underline{X}^a_{N_a(t)}\right)p_{\mu'_b}\left(\underline{X}^b_{N_b(t)}\right)}{\max_{\mu'_a\leq\mu'_b} p_{\mu'_a}\left(\underline{X}^a_{N_a(t)}\right)p_{\mu'_b}\left(\underline{X}^b_{N_b(t)}\right)}\,,$$

where $\underline{X}^a_{N_a(t)} = (X_s : A_s = a, s \leq t)$ is a vector that contains the observations of arm a available at time t, and where $p_\mu(Z_1,\dots,Z_n)$ is the likelihood of n i.i.d. observations from w^μ:

$$p_\mu(Z_1,\dots,Z_n)=\prod_{k=1}^n \exp(\dot{b}^{-1}(\mu)Z_k - b(\dot{b}^{-1}(\mu)))\,.$$

This statistic has a convenient closed-form expression for exponential family bandit models. Introducing for all arms a, b a weighted average of their empirical mean:

$$\hat{\mu}_{a,b}(t) \quad := \quad \frac{N_a(t)}{N_a(t)+N_b(t)}\hat{\mu}_a(t)+\frac{N_b(t)}{N_a(t)+N_b(t)}\hat{\mu}_b(t)\,,$$

it is well known and easy to see that if $\hat{\mu}_a(t)\geq\hat{\mu}_b(t)$,

$$Z_{a,b}(t)=N_a(t)\,d\big(\hat{\mu}_a(t),\hat{\mu}_{a,b}(t)\big)+N_b(t)\,d\big(\hat{\mu}_b(t),\hat{\mu}_{a,b}(t)\big)\,, \tag{7}$$

and that $Z_{a,b}(t) = -Z_{b,a}(t)$. The testing intuition thus suggests the following stopping rule:

$$\begin{aligned}\tau_\delta &= \inf\left\{t\in\mathbb{N} : \exists a\in\mathcal{A}, \forall b\in\mathcal{A}\setminus\{a\}, Z_{a,b}(t)>\beta(t,\delta)\right\}\\ &= \inf\left\{t\in\mathbb{N} : Z(t):=\max_{a\in\mathcal{A}}\min_{b\in\mathcal{A}\setminus\{a\}} Z_{a,b}(t)>\beta(t,\delta)\right\}\,,\end{aligned} \tag{8}$$

where $\beta(t,\delta)$ is an exploration rate to be tuned appropriately. The form of this stopping rule can be traced back to Chernoff (1959)[3]. As $\min_{b\in\mathcal{A}\setminus a} Z_{a,b}(t)$ is non-negative if and only if $\hat{\mu}_a(t) \geq$

[3] The stopping rule τ_δ was proposed under equivalent form (with a different threshold) in the context of adaptive sequential hypothesis testing. Best arm identification in a bandit model can be viewed as a particular instance in which we test K hypotheses, $H_a : (\mu_a = \max_{i\in\mathcal{A}}\mu_i)$, based on adaptively sampling the marginal of $\nu = (\nu^{\mu_1},\dots,\nu^{\mu_K})$. However, Chernoff (1959) considers a different performance criterion, and its analysis holds when each of the hypotheses consists in a finite set of parameters, unlike the bandit setting.

$\hat{\mu}_b(t)$ for all $b \neq a$, $Z(t) = \min_{b\in\mathcal{A}\setminus\hat{a}_t} Z_{\hat{a}_t,b}(t)$ whenever there is a unique best empirical arm $\hat{a}_t = \operatorname{argmax}_{a\in\mathcal{A}} \hat{\mu}_a(t)$ at time t. Obviously, $\big(\hat{\mu}_a(\tau_\delta)\big)_a$ has a unique maximizer, which is the final decision.

In addition to the testing interpretation given above, the stopping rule can be explained in light of the lower bound $\mathbb{E}_\mu[\tau_\delta/T^*(\boldsymbol{\mu})] \geq \mathrm{kl}(\delta, 1-\delta)$. Indeed, one may write

$$Z(t) = t \min_{\lambda\in\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} \left(\frac{N_a(t)}{t}\right) d(\hat{\mu}_a(t), \lambda_a) \leq \frac{t}{T^*\big(\hat{\mu}(t)\big)} \ .$$

Theorem 1 suggests that a δ-PAC strategy cannot stop (at least in expectation) before $t/T^*\big(\hat{\mu}(t)\big)$ is larger than $\mathrm{kl}(\delta, 1-\delta) \sim \log(1/\delta)$, which suggests to stop when $Z(t) \geq \log(1/\delta)$. In Section 4, we prove that a slightly more prudent choice of the threshold $\beta(t,\delta)$ does lead to a PAC algorithm (whatever the sampling rule, even if the proportions of draws are sub-optimal). And it is shown in Section 5 that, using our sampling rule, $\mathbb{E}[\tau_\delta]$ is indeed of order $T^*(\boldsymbol{\mu})\log(1/\delta)$.

It is also possible to give a Minimum Description Length (MDL) interpretation of the stopping rule. It is well known that choosing the model that gives the shortest description of the data is a provably efficient heuristic (see Rissanen (1978), and Grünwald (2007) for a survey). In some sense, the stopping rule presented above follows the same principle. In fact, elementary algebra shows that

$$Z_{a,b}(t) = (N_a(t) + N_b(t))h\left(\hat{\mu}_{a,b}(t)\right) - \left[N_a(t)h(\hat{\mu}_a(t)) + N_b(t)h(\hat{\mu}_b(t))\right],$$

where $h(\mu) = \mathbb{E}_{X\sim\nu^\mu}[-\log p_\mu(X)] = b(\dot{b}^{-1}(\mu)) - \dot{b}^{-1}(\mu)\mu$. In the Bernoulli case, the *Shannon entropy* $h(\mu) = -\mu\log(\mu) - (1-\mu)\log(1-\mu)$ is well-known to represent an ideal code length per character for binary compression. Thus, $Z_{a,b}(t)$ appears as the difference between the ideal code length for the rewards of arms a and b coded together, and the sum of the ideal code lengths for the rewards coded separately. If this difference is sufficiently large[4], the shortest description of the data is to separate arms a and b. The stopping rule (8) thus consists in waiting for the moment when it becomes cheaper to code the rewards of the best empirical arm separately from each of the others. It is no surprise that the proof of Proposition 10 below is based on a classical information-theoretic argument.

4. Choosing the Threshold in the Stopping Rule

We now explain how to choose the threshold $\beta(t,\delta)$ so as to ensure the δ-PAC property: with probability larger than $1-\delta$, any algorithm based on the stopping rule (8) outputs the optimal arm, provided that it stops. The interpretations of the stopping rule presented in the last section suggest the presence of two ingredients: $\log(1/\delta)$ for the risk, and $\log(t)$ for the fluctuations of the counts. We present here two results: one is based on an information-theoretic argument used for consistency proofs of MDL estimators, and the second is based on the probabilistic control of self-normalized averages taken from Magureanu et al. (2014). To keep things simple, the first argument is detailed only for the case of Bernoulli rewards (the standard framework of coding theory). The second argument is more general, but a little less tight.

[4] Universal coding theory would suggest a threshold of order $\log(t)/2$, a term that will appear in Section 4.

The Informational Threshold.

Theorem 10 *Let $\delta \in (0,1)$. Whatever the sampling strategy, using Chernoff's stopping rule* (8) *on Bernoulli bandits with threshold*

$$\beta(t,\delta) = \log\left(\frac{2t(K-1)}{\delta}\right)$$

ensures that for all $\boldsymbol{\mu} \in \mathcal{S}$, $\mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty,\, \hat{a}_{\tau_\delta} \neq a^*\right) \leq \delta$.

Proof sketch. A more detailed proof is given in Appendix C.1. We proceed here similarly to Vaidhyan and Sundaresan (2015), employing an argument used for MDL consistency proofs (see Garivier (2006) and references therein). Introducing, for $a, b \in \mathcal{A}$, $T_{a,b} := \inf\{t \in \mathbb{N} : Z_{a,b}(t) > \beta(t,\delta)\}$, one has

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) &\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists a \in \mathcal{A} \setminus a^*, \exists t \in \mathbb{N} : Z_{a,a^*}(t) > \beta(t,\delta)\right) \\ &\leq \sum_{a \in \mathcal{A}\setminus\{a^*\}} \mathbb{P}_{\boldsymbol{\mu}}(T_{a,a^*} < \infty)\ .\end{aligned}$$

It is thus sufficient to show that if $\beta(t,\delta) = \log(2t/\delta)$, and if $\mu_a < \mu_b$, then $\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) \leq \delta$. For such a pair of arms, observe that on the event $\left\{T_{a,b} = t\right\}$ time t is the first moment when $Z_{a,b}$ exceeds the threshold $\beta(t,\delta)$, which implies by definition that

$$1 \leq e^{-\beta(t,\delta)} \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{X}_t^a) p_{\mu'_b}(\underline{X}_t^b)}{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(\underline{X}_t^a) p_{\mu'_b}(\underline{X}_t^b)}\ .$$

It thus holds that

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) = \sum_{t=1}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} = t) = \sum_{t=1}^{\infty} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)}\right] \\ &\leq \sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)} \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{X}_t^a) p_{\mu'_b}(\underline{X}_t^b)}{\max_{\mu'_a \leq \mu'_b} p_{\mu'_a}(\underline{X}_t^a) p_{\mu'_b}(\underline{X}_t^b)}\right] \\ &\leq \sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)} \frac{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{X}_t^a) p_{\mu'_b}(\underline{X}_t^b)}{p_{\mu_a}(\underline{X}_t^a) p_{\mu_b}(\underline{X}_t^b)}\right] \\ &= \sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \int_{\{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(x_1, \dots, x_t) \underbrace{\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{x}_t^a) p_{\mu'_b}(\underline{x}_t^b) \prod_{i \in \mathcal{A}\setminus\{a,b\}} p_{\mu_i}(\underline{x}_t^i)}_{(*)}\, dx_1 \dots dx_t\ .\end{aligned}$$

Of course the maximum likelihood $(*)$ is not a probability density. A possible workaround (sometimes referred to as *Barron's lemma*, see Barron et al. (1998) and references therein) is to use a *universal distribution* like Krichevsky and Trofimov (1981), which is known to provide a tight uniform approximation:

Lemma 11 *[Willems et al. (1995)] Let $p_u(x)$ be the likelihood of successive observations $x \in \{0,1\}^n$ of a Bernoulli random variable with mean u. Then the Krichevsky-Trofimov distribution*

$$\mathrm{kt}(x) = \int_0^1 \frac{1}{\pi\sqrt{u(1-u)}} p_u(x) \mathrm{d}u$$

is a probability law on $\{0,1\}^n$ *that satisfies*

$$\sup_{x\in\{0,1\}^n} \sup_{u\in[0,1]} \frac{p_u(x)}{\mathrm{kt}(x)} \leq 2\sqrt{n}\ .$$

Together with the inequality $\sqrt{ab} \leq (a+b)/2$, this property permits to conclude that $\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty)$ is upper-bounded by

$$\sum_{t=1}^{\infty} 2te^{-\beta(t,\delta)} \int_{\{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(x_1,\dots,x_t) \overbrace{\mathrm{kt}(\underline{x}_t^a)\mathrm{kt}(\underline{x}_t^b) \prod_{i\in\mathcal{A}\setminus\{a,b\}} p_{\mu_i}(\underline{x}_t^i)}^{:=I(x_1,\dots,x_t)} dx_1\dots dx_t$$
$$= \sum_{t=1}^{\infty} 2te^{-\beta(t,\delta)} \tilde{\mathbb{E}}\left[\mathbb{1}_{(T_{a,b}=t)}\right] \leq \delta\, \tilde{\mathbb{P}}(T_{a,b} < \infty) \leq \delta,$$

using that the partially integrated likelihood $I(x_1,\dots,x_t)$ is the density of a probability measure $\tilde{\mathbb{P}}$ (we denote the corresponding expectation by $\tilde{\mathbb{E}}$).

□

The Deviational Threshold. The universal coding argument above can be adapted to other distributions, as shown for example in Chambaz et al. (2009). It is also possible to make use of a deviation result like Magureanu et al. (2014) in order to prove PAC guarantees in any exponential family bandit model. The exploration rate involved are slightly larger and less explicit.

Proposition 12 *Let* $\boldsymbol{\mu}$ *be an exponential family bandit model. Let* $\delta \in (0,1)$ *and* $\alpha > 1$. *There exists a constant* $C = C(\alpha, K)$ *such that whatever the sampling strategy, using Chernoff's stopping rule* (8) *with the threshold*

$$\beta(t,\delta) = \log\left(\frac{Ct^\alpha}{\delta}\right)$$

ensures that for all $\boldsymbol{\mu} \in \mathcal{S}$, $\mathbb{P}_{\boldsymbol{\mu}}\big(\tau_\delta < \infty,\, \hat{a}_{\tau_\delta} \neq a^*\big) \leq \delta$.

The proof of Proposition 12 is given in Appendix C.2.

5. Sample Complexity Analysis

Combining Chernoff's stopping rule and an optimal-tracking sampling rule permits to approach the lower bound of Theorem 1 for sufficiently small values of the risk δ. We first state a simple almost-sure convergence result and give its short proof. Then, we sketch the (somewhat more technical) analysis controlling the expectation of τ_δ.

5.1. Almost-sure Upper Bound

Proposition 13 *Let* $\alpha \in [1, e/2]$ *and* $r(t) = O(t^\alpha)$. *Using Chernoff's stopping rule with* $\beta(t,\delta) = \log(r(t)/\delta)$, *and any sampling rule ensuring that for every arm* $a \in \mathcal{A}$, $N_a(t)/t$ *converges almost-surely to* w_a^* *guarantees that for all* $\delta \in (0,1)$, $\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < +\infty) = 1$ *and*

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu})\right) = 1.$$

Proof. Let $\mathcal{E}$ be the event

$$\mathcal{E} = \left\{ \forall a \in \mathcal{A}, \frac{N_a(t)}{t} \underset{t\to\infty}{\to} w_a^* \text{ and } \hat{\boldsymbol{\mu}}(t) \underset{t\to\infty}{\to} \boldsymbol{\mu} \right\} .$$

From the assumption on the sampling strategy and the law of large number, $\mathcal{E}$ is of probability 1. On $\mathcal{E}$, there exists t_0 such that for all $t \geq t_0$, $\hat{\mu}_1(t) > \max_{a\neq 1} \hat{\mu}_a(t)$ and

$$\begin{aligned} Z(t) &= \min_{a\neq 1} Z_{1,a}(t) = \min_{a\neq 1} N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,a}(t)) + N_a(t) d(\hat{\mu}_a(t), \hat{\mu}_{1,a}(t)) \\ &= t \left[\min_{a\neq 1} \left(\frac{N_1(t)}{t} + \frac{N_a(t)}{t} \right) I_{\frac{N_1(t)/t}{N_1(t)/t+N_a(t)/t}} (\hat{\mu}_1(t), \hat{\mu}_a(t)) \right] . \end{aligned}$$

For all $a \geq 2$, the mapping $(\boldsymbol{w}, \boldsymbol{\lambda}) \to (w_1 + w_a) I_{w_1/(w_1+w_a)}(\lambda_1, \lambda_a)$ is continuous at $(w^*(\boldsymbol{\mu}), \boldsymbol{\mu})$. Therefore, for all $\epsilon > 0$ there exists $t_1 \geq t_0$ such that for all $t \geq t_1$ and all $a \in \{2, \dots, K\}$,

$$\left(\frac{N_1(t)}{t} + \frac{N_a(t)}{t} \right) I_{\frac{N_1(t)/t}{N_1(t)/t+N_a(t)/t}} (\hat{\mu}_1(t), \hat{\mu}_a(t)) \geq \frac{w_1^* + w_a^*}{1+\epsilon} I_{\frac{w_1^*}{w_1^*+w_a^*}} (\mu_1, \mu_a) .$$

Hence, for $t \geq t_1$,

$$Z(t) \geq \frac{t}{1+\epsilon} \min_{a\neq 1} (w_1^* + w_a^*) I_{\frac{w_1^*}{w_1^*+w_a^*}} (\mu_1, \mu_a) = \frac{t}{(1+\epsilon) T^*(\boldsymbol{\mu})}.$$

Consequently,

$$\begin{aligned} \tau_\delta &= \inf\{t \in \mathbb{N} : Z(t) \geq \beta(t,\delta)\} \\ &\leq t_1 \vee \inf\{t \in \mathbb{N} : t(1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} \geq \log(r(t)/\delta)\} \\ &\leq t_1 \vee \inf\{t \in \mathbb{N} : t(1+\epsilon)^{-1} T^*(\boldsymbol{\mu})^{-1} \geq \log(Ct^\alpha/\delta)\}, \end{aligned}$$

for some positive constant C. Using the technical Lemma 18 in the Appendix, it follows that on $\mathcal{E}$, as $\alpha \in [1, e/2]$,

$$\tau_\delta \leq t_1 \vee \alpha(1+\epsilon) T^*(\boldsymbol{\mu}) \left[\log\left(\frac{Ce((1+\epsilon)T^*(\boldsymbol{\mu}))^\alpha}{\delta} \right) + \log\log\left(\frac{C((1+\epsilon)T^*(\boldsymbol{\mu}))^\alpha}{\delta} \right) \right] .$$

Thus τ_δ is finite on $\mathcal{E}$ for every $\delta \in (0,1)$, and

$$\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq (1+\epsilon)\, \alpha\, T^*(\boldsymbol{\mu}) .$$

Letting ϵ go to zero concludes the proof.

□

5.2. Asymptotic Optimality in Expectation

In order to prove that the lower bound of Theorem 1 is matched, we now give an upper bound on the expectation of the stopping time τ_δ. The proof of this result is to be found in Appendix D.

Theorem 14 *Let $\boldsymbol{\mu}$ be an exponential family bandit model. Let $\alpha \in [1, e/2]$ and $r(t) = O(t^\alpha)$. Using Chernoff's stopping rule with $\beta(t,\delta) = \log(r(t)/\delta)$, and the sampling rule C-Tracking or D-Tracking,*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}) \ .$$

To summarize, for every Bernoulli bandit $\boldsymbol{\mu}$, the choice $\beta(t,\delta) = \log(2(K-1)t/\delta)$ in Chernoff's stopping rule is δ-PAC (by Theorem 10); with one of the sampling rules given above, the stopping time τ_δ is almost surely finite (by Proposition 13) and when δ is small enough its expectation is close to $T^*(\boldsymbol{\mu})\log(1/\delta)$ by Theorem 14, an optimal sample complexity after Theorem 1.

More generally, for exponential family bandit models, combining Proposition 12 and Theorem 14, one obtains for every $\alpha > 1$ the existence of an exploration rate for which Chernoff's stopping rule combined with a tracking sampling rule is δ-PAC and satisfies

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}) \ .$$

6. Discussion and Numerical Experiments

We give here a few comments on the practical behaviour of the Track-and-Stop (T-a-S) strategy. Let us first emphasize that the forced exploration step are rarely useful, but in some cases really necessary and not only for the theorems: when μ_2 and μ_3 are equal, they prevent the probability that the strategy never ends from being strictly positive. Second, our simulation study suggests that the exploration rate $\beta(t,\delta) = \log((\log(t)+1)/\delta)$, though not (yet) allowed by theory, is still over-conservative in practice. Further, even though any sampling strategy ensuring that $N_a(t)/t \to w_a^*$ satisfies the optimality theorems above, we propose (without formal justification) an experimentally more efficient sampling rule: after t observations, let

$$\hat{c}_t = \underset{c \in \mathcal{A} \setminus \{\hat{a}_t\}}{\text{argmin}} \ Z_{\hat{a}_t, c}(t)$$

be the 'best challenger' of the empirical champion $\hat{a}_t$. We choose $A_{t+1} = \hat{a}_t$ if

$$\frac{N_{\hat{a}_t}(t)}{N_{\hat{a}_t}(t) + N_{\hat{c}_t}(t)} < \frac{w^*_{\hat{a}_t}(\hat{\boldsymbol{\mu}}(t))}{w^*_{\hat{a}_t}(\hat{\boldsymbol{\mu}}(t)) + w^*_{\hat{c}(t)}(\hat{\boldsymbol{\mu}}(t))}$$

and $A_{t+1} = \hat{c}_t$ otherwise (with forced explorations steps as in the D-Tracking rule).

We consider two sample scenarios $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$ and $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$, and we choose $\delta = 0.1$. This choice is meant to illustrate that our algorithm performs well even for relatively high risk values (so far, optimality is proved only for small risks). We compare the Track-and-Stop algorithm based on D-Tracking and BestChallenger to algorithms from the literature designed for Bernoulli bandit models, namely KL-LUCB and KL-Racing (Kaufmann and Kalyanakrishnan, 2013). Racing algorithms proceed in rounds: at start, all arms are active; at each round, all active arms are drawn once; at the end of a round, a rule determines if the empirically worst arm should be eliminated. Call $\hat{a}_r$ the empirically best arm after r rounds. In KL-Racing, arm b is eliminated if its upper confidence bound $\max\{q \in [0,1] : rd(\hat{\mu}_{b,r}, q) \leq \beta(r,\delta)\}$ is smaller than the best arm's lower bound $\min\{q \in [0,1] : rd(\hat{\mu}_{\hat{a}_r,r}, q) \leq \beta(r,\delta)\}$. We also introduced in the competition the 'hybrid' Chernoff-Racing strategy, which eliminates b if

$$Z_{\hat{a}_r,b} = rd\left(\hat{\mu}_{\hat{a}_r,r}, \frac{\hat{\mu}_{\hat{a}_r,r} + \hat{\mu}_{b,r}}{2}\right) + rd\left(\hat{\mu}_{b,r}, \frac{\hat{\mu}_{\hat{a}_r,r} + \hat{\mu}_{b,r}}{2}\right) > \beta(r,\delta).$$

Table 1 presents the estimated average number of draws of the five algorithms in the two scenarios. Our (Julia) code will be available online. We see that the use of the MDL stopping rule leads to a clear improvement. Moreover, Chernoff-Racing significantly improves over KL-Racing, and its performance is even close to that of our optimal algorithms.

	T-a-S (BC)	T-a-S (D-Tracking)	Chernoff-Racing	KL-LUCB	KL-Racing
$\boldsymbol{\mu}_1$	3968	4052	4516	8437	9590
$\boldsymbol{\mu}_2$	1370	1406	3078	2716	3334

Table 1: Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments: $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$, $w^*(\boldsymbol{\mu}_1) = [0.417\ 0.390\ 0.136\ 0.057]$; $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$, $w^*(\boldsymbol{\mu}_2) = [0.336\ 0.251\ 0.177\ 0.132\ 0.104]$.

It should be emphasized that a Racing-type algorithm cannot reach the lower bound in general: by construction, it forces the last two arms in the race (hopefully μ_1 and μ_2) to be drawn equally often, which is sub-optimal unless $w_1^*(\boldsymbol{\mu}) = w^*(\boldsymbol{\mu}_2)$ (a condition approximately matched only if there is a large gap between the second and the third best arms). This is illustrated in the second scenario $\boldsymbol{\mu}_2$ of Table 1, where the ratio w_1^*/w_2^* is larger.

7. Conclusion

We gave a characterization of the complexity of best arm identification in the fixed confidence-setting, for a large class of bandit models with arms parameterized by their means. Our new lower bound reveals the existence of optimal proportions of draws of the arms that can be computed efficiently. Our Track-and-Stop strategy, by combining a track of these optimal proportions with Chernoff's stopping rule, asymptotically matches the lower bound. In future work, instead of using forced exploration steps within a plugin procedure, we will investigate optimistic (or robust-to-noise) sampling strategies in order to optimize the exploration and to obtain non-asymptotic sample complexity bounds. Furthermore, we will investigate the fixed-budget setting, for which we conjecture that $P(\hat{a}_t \neq 1) \gtrsim \exp\big(-t/T_*(\boldsymbol{\mu})\big)$ with

$$T_*(\boldsymbol{\mu})^{-1} = \sup_{w\in\Sigma_K} \min_{a\in\{2,\dots,K\}} \inf_{\mu_a<m_a<\mu_1} w_1\, d(m_a,\mu_1) + w_a\, d(m_a,\mu_a)\ .$$

Acknowledgments

This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program while Emilie Kaufmann visited Toulouse in November 2015. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA). They thank Sébastien Gerchinovitz for stimulating discussions and comments.

References

Y. Abbasi-Yadkori, D.Pál, and C.Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, 2011.

S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning (ICML)*, 2013.

A. Antos, V. Grover, and C. Szepesvári. Active learning in multi-armed bandits. In *Algorithmic Learning Theory*, 2008.

A. Barron, J. Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *Information Theory, IEEE Transactions on*, 44(6):2743–2760, Oct 1998. ISSN 0018-9448. doi: 10.1109/18.720554.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Fondations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1587–1627, 2011.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

Antoine Chambaz, Aurélien Garivier, and Elisabeth Gassiat. A MDL approach to HMM with poisson and gaussian emissions. application to order identification. *Journal of Statistical Planning and Inference*, 139(3):962–977, 2009.

H. Chernoff. Sequential design of Experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

R. Combes and A. Proutière. Unimodal Bandits without Smoothness. Technical report, 2014.

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.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*, 2012.

Aurlien Garivier. Consistency of the unlimited BIC context tree estimator. *IEEE Transactions on Information Theory*, 52(10):4630–4635, 2006.

T.L. Graves and T.L. Lai. Asymptotically Efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743, 1997.

Peter D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007. ISBN 0262072815.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*, 2012.

E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceeding of the 26th Conference On Learning Theory.*, 2013.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In *Proceedings of the 27th Conference On Learning Theory*, 2014.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research (to appear)*, 2015.

Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–206, 1981. doi: 10.1109/TIT.1981.1056331. URL http://dx.doi.org/10.1109/TIT.1981.1056331.

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

S. Magureanu, R. Combes, and A. Proutière. Lipschitz Bandits: Regret lower bounds and optimal algorithms. In *Proceedings on the 27th Conference On Learning Theory*, 2014.

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

R. Munos. *From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning.*, volume 7. Foundations and Trends in Machine Learning, 2014.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465 – 471, 1978. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/0005-1098(78)90005-5. URL http://www.sciencedirect.com/science/article/pii/0005109878900055.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian Process Optimization in the Bandit Setting : No Regret and Experimental Design. In *Proceedings of the International Conference on Machine Learning*, 2010.

N.K. Vaidhyan and R. Sundaresan. Learning to detect an oddball target. *arXiv:1508.05572*, 2015.

Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.

Appendix A. Characterization of the optimal proportion of draws

A.1. Proof of Lemma 3

Let $\boldsymbol{\mu}$ such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Using the fact that

$$\mathrm{Alt}(\boldsymbol{\mu}) = \bigcup_{a \neq 1} \left\{ \boldsymbol{\lambda} \in \mathcal{S} : \lambda_a > \lambda_1 \right\},$$

one has

$$\begin{aligned} T^*(\boldsymbol{\mu})^{-1} &= \sup_{w \in \Sigma_K} \min_{a \neq 1} \inf_{\boldsymbol{\lambda} \in \mathcal{S} : \lambda_a > \lambda_1} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \\ &= \sup_{w \in \Sigma_K} \min_{a \neq 1} \inf_{\boldsymbol{\lambda} : \lambda_a \geq \lambda_1} \left[w_1 d(\mu_1, \lambda_1) + w_a d(\mu_a, \lambda_a) \right] \ . \end{aligned}$$

Minimizing

$$f(\lambda_1, \lambda_a) = w_1 d(\mu_1, \lambda_1) + w_a d(\mu_a, \lambda_a)$$

under the constraint $\lambda_a \geq \lambda_1$ is a convex optimization problem that can be solved analytically. The minimum is obtained for

$$\lambda_1 = \lambda_a = \frac{w_1}{w_1 + w_a}\mu_1 + \frac{w_a}{w_1 + w_a}\mu_a$$

and its value can be rewritten $(w_1 + w_a) I_{\frac{w_1}{w_1 + w_a}}(\mu_1, \mu_a)$, using the function I_α defined in (3).

A.2. Proof of Theorem 5

The function g_a introduced in (4) rewrites

$$g_a(x) = d\left(\mu_1, m_a(x)\right) + x d(\mu_a, m_a(x)), \ \ \text{with} \ \ m_a(x) = \frac{\mu_1 + x\mu_a}{1 + x} \ .$$

Using that $m_a'(x) = (\mu_a - \mu_1)/(1+x)^2$ and $\frac{d}{dy} d(x, y) = (y - x)/\ddot{b}(b^{-1}(y))$ one can show that g_a is strictly increasing, since $g_a'(x) = d(\mu_a, m_a(x)) > 0$. As $g_a(x)$ tends to $d(\mu_1, \mu_a)$ when x goes to infinity, the inverse function $x_a(y) = g_a^{-1}(y)$ is defined on $[0, d(\mu_1, \mu_a)[$ and satisfies

$$x_a'(y) = \frac{1}{d(\mu_a, m_a(x_a(y)))} > 0 \ .$$

Let $\boldsymbol{w}^*$ be an element in

$$\underset{w \in \Sigma_K}{\mathrm{argmax}} \ \min_{a \neq 1} \, (w_1 + w_a) I_{\frac{w_1}{w_1 + w_a}}(\mu_1, \mu_a) \quad = \quad \underset{w \in \Sigma_K}{\mathrm{argmax}} \ w_1 \min_{a \neq 1} g_a\left(\frac{w_a}{w_1}\right) \ .$$

The equality uses that $w_1^* \neq 0$ (since if $w_1 = 0$, the value of the objective is zero). Introducing $x_a^* = \frac{w_a^*}{w_1^*}$ for all $a \neq 1$, one has

$$w_1^* = \frac{1}{1 + \sum_{a=2}^{K} x_a^*} \quad \text{and, for } a \geq 2, \ \ w_a^* = \frac{x_a^*}{1 + \sum_{a=2}^{K} x_a^*}$$

and $(x_2^*, \dots, x_K^*) \in \mathbb{R}^{K-1}$ belongs to

$$\underset{(x_2,\dots,x_K)\in\mathbb{R}^{K-1}}{\text{argmax}} \frac{\min_{a\neq 1} g_a\left(x_a\right)}{1+x_2+\dots+x_K}. \tag{9}$$

We now show that all the $g_a(x_a^*)$ have to be equal (Lemma 4). Let

$$\mathcal{B} = \left\{ b \in \{2,\dots,K\} : g_b(x_b^*) = \min_{a\neq 1} g_a(x_a^*) \right\}$$

and $\mathcal{A} = \{2,\dots,K\}\backslash\mathcal{B}$. Assume that $\mathcal{A} \neq \emptyset$. For all $a \in \mathcal{A}$ and $b \in \mathcal{B}$, one has $g_a(x_a^*) > g_b(x_b^*)$. Using the continuity of the g functions and the fact that they are strictly increasing, there exists $\epsilon > 0$ such that

$$\forall a \in \mathcal{A}, b \in \mathcal{B}, \quad g_a(x_a^* - \epsilon/|\mathcal{A}|)) > g_b\left(x_b^* + \epsilon/|\mathcal{B}|\right) > g_b\left(x_b^*\right).$$

Introducing $\overline{x}_a = x_a^* - \epsilon/|\mathcal{A}|$ for all $a \in \mathcal{A}$ and $\overline{x}_b = x_b^* + \epsilon/|\mathcal{B}|$ for all $b \in \mathcal{B}$, there exists $b \in \mathcal{B}$:

$$\frac{\min_{a\neq 1} g_a\left(\overline{x}_a\right)}{1+\overline{x}_2+\dots\overline{x}_K} = \frac{g_b\left(x_b^*+\epsilon/|\mathcal{B}|\right)}{1+x_2^*+\dots+x_K^*} > \frac{g_b\left(x_b^*\right)}{1+x_2^*+\dots+x_K^*} = \frac{\min_{a\neq 1} g_a(x_a^*)}{1+x_2^*+\dots+x_K^*},$$

which contradicts the fact that $\boldsymbol{x}^*$ belongs to (9). Hence $\mathcal{A} = \emptyset$ and there exists $y^* \in [0, d(\mu_1,\mu_2)[$ such that

$$\forall a \in \{2,\dots,K\}, \; g_a(x_a^*) = y^* \quad \Leftrightarrow \quad x_a^* = x_a(y^*),$$

with the function x_a introduced above. From (9), y^* belongs to

$$\underset{y\in[0,d(\mu_1,\mu_2)[}{\text{argmax}} \; G(y) \quad \text{with} \quad G(y) = \frac{y}{1+x_2(y)+\dots+x_K(y)}.$$

G is differentiable and, using the derivative of the x_a given above, $G'(y) = 0$ is equivalent to

$$\begin{aligned}
\sum_{a=2}^K \frac{y}{d(\mu_a, m_a(x_a(y))} &= 1 + x_2(y) + \dots + x_K(y) \\
\sum_{a=2}^K \frac{d(\mu_1, m_a(x_a(y))) + x_a(y) d(\mu_a, m_a(x_a(y))}{d(\mu_a, m_a(x_a(y))} &= 1 + x_2(y) + \dots + x_K(y) \\
\sum_{a=2}^K \frac{d(\mu_1, m_a(x_a(y)))}{d(\mu_a, m_a(x_a(y))} &= 1\,. \qquad (10)
\end{aligned}$$

For the the second equality, we use that $\forall a, d(\mu_1, m_a(x_a(y))) + x_a(y) d(\mu_a, m_a(x_a(y)) = y$. Thus y^* is solution of the equation (10). This equation has a unique solution since

$$F_{\boldsymbol{\mu}}(y) = \sum_{a=2}^K \frac{d(\mu_1, m_a(x_a(y)))}{d(\mu_a, m_a(x_a(y))} \tag{11}$$

is strictly increasing and satisfies $F_{\boldsymbol{\mu}}(0) = 0$ and $\lim_{y\to d(\mu_1,\mu_2)} F_{\boldsymbol{\mu}}(y) = +\infty$. As G is positive and satisfies $G(0) = 0$, $\lim_{y\to d(\mu_1,\mu_2)} G(y) = 0$, the unique local extrema obtained in y^* is a maximum.

A.3. Proof of Proposition 6

Let $\boldsymbol{\mu} \in \mathcal{S}$, and re-label its arms in decreasing order. From Theorem 5, $w_1(\boldsymbol{\mu}) \neq 0$ and for $a \geq 2$, $w_a^*(\boldsymbol{\mu}) = 0 \Leftrightarrow x_a(y^*) = 0 \Rightarrow y^* = 0$ where y^* is the solution of (10). But 0 is not solution of (10), since the value of the left-hand side is 0. This proves that for all a, $w_a^*(\boldsymbol{\mu}) \neq 0$. For a given $\boldsymbol{\mu}$, y^* is defined by

$$F_{\boldsymbol{\mu}}(y^*) - 1 = 0 \, ,$$

where $F_{\boldsymbol{\mu}}$ is defined in (11). For all $\boldsymbol{\mu} \in \mathcal{S}$ and every $y \in [0, d(\mu_1, \mu_2)[$, it can be shown that $\frac{d}{dy} F_{\boldsymbol{\mu}}(y) \neq 0$, in particular $\frac{d}{dy} F_{\boldsymbol{\mu}}(y^*) \neq 0$. Thus y^* is a function of $\boldsymbol{\mu}$ that is continuous in every $\boldsymbol{\mu} \in \mathcal{S}$, denoted by $y^*(\boldsymbol{\mu})$. By composition, the function $\boldsymbol{\mu} \mapsto x_a(y^*(\boldsymbol{\mu}))$ are continuous in $\boldsymbol{\mu} \in \mathcal{S}$, and so does w^*.

The proof of Statement 3 relies on the fact that if a and b are such that $\mu_1 > \mu_a \geq \mu_b$, $g_a(x) \leq g_b(x)$ for all x. Thus, for all $y \in [0, d(\mu_1, \mu_2)[$, $x_a(y) \geq x_b(y)$ and particularizing for $y^*(\boldsymbol{\mu})$ yields the result.

A.4. Bounds on the characteristic time in the Gaussian case

In the Gaussian case, with $d(x, y) = (x - y)^2/(2\sigma^2)$, the expression in Lemma 3 can be made more explicit and yields

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{a \neq 1} \frac{w_1 w_a}{w_1 + w_a} \frac{\Delta_a^2}{2\sigma^2} \, .$$

Introducing $\tilde{w} \in \Sigma_K$ defined by

$$\forall a = 1 \dots K, \;\; \tilde{w}_a = \frac{1/\Delta_a^2}{\sum_{i=1}^K 1/\Delta_i^2} \, ,$$

it holds that

$$T^*(\boldsymbol{\mu})^{-1} \geq \inf_{a \neq 1} \frac{\tilde{w}_1 \tilde{w}_a}{\tilde{w}_1 + \tilde{w}_a} \frac{\Delta_a^2}{2\sigma^2} = \frac{1}{\sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2}} \inf_{a \neq 1} \frac{1}{1 + \frac{\Delta_1^2}{\Delta_a^2}} .$$

The infimum is obtained for $a = 2$, and using that $\Delta_2 = \Delta_1$ leads to the upper bound

$$T^*(\boldsymbol{\mu}) \leq 2 \left(\sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2} \right) \, .$$

The following lower bound was obtained by Kaufmann et al. (2015) for every PAC strategy:

$$\liminf_{\delta \to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \geq \sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2} \, .$$

Combining this inequality with the upper bound on $\liminf_{\delta \to 0} \mathbb{E}[\tau_\delta]/\log(1/\delta)$ obtained in Theorem 14 for the δ-PAC Track-and-Stop algorithm shows that

$$T^*(\boldsymbol{\mu}) \geq \sum_{i=1}^K \frac{2\sigma^2}{\Delta_i^2} \, ,$$

which concludes the proof.

Appendix B. Tracking results

Lemma 7 and 8 both follow from deterministic results that we can give for procedures tracking any cumulated sums of proportions (Lemma 15) or any changing sequence of proportions that concentrates (Lemma 17). We state and prove theses two results in this section, and also explain how they lead to Lemma 7 and 8. We then provide a proof of Proposition 9.

B.1. Tracking a cumulated sum of proportions

Lemma 15 *Let K be a positive integer, let Σ_K be the simplex of dimension $K-1$, and for every $i \in \{1,\dots,K\}$, let δ_i be the vertex of Σ_K with a 1 on coordinate i. For a positive integer n, let $p(1), p(2), \dots, p(n) \in \Sigma_K$ and for every $k \leq n$ let $P(k) = p(1) + \dots + p(k)$. Define $N(0) = 0$, for every $k \in \{0,\dots,n-1\}$*

$$I_{k+1} \in \underset{1 \leq i \leq K}{\mathit{argmax}} \; [P_i(k+1) - N_i(k)]$$

and $N(k+1) = N(k) + \delta_{I_{k+1}}$. Then

$$\max_{1 \leq i \leq K} \left| N_i(n) - P_i(n) \right| \leq K - 1 \, .$$

To obtain Lemma 7, we start by applying Lemma 15 with $p(k) = w^{\epsilon_{k-1}}(\hat{\boldsymbol{\mu}}(k-1))$, so that $P(k+1) = \sum_{s=0}^{k} w^{\epsilon_s}(\hat{\boldsymbol{\mu}}(s))$. One obtains

$$\max_{1 \leq a \leq K} \left| N_a(t) - \sum_{s=0}^{t-1} w_a^{\epsilon_s}(\hat{\boldsymbol{\mu}}(s)) \right| \leq K - 1 \, . \tag{12}$$

Moreover, by definition of $w^{\epsilon}(s)$,

$$\max_{1 \leq a \leq K} \left| \sum_{s=0}^{t-1} w_a^{\epsilon_s}(\hat{\boldsymbol{\mu}}(s)) - \sum_{s=0}^{t-1} w_a^{*}(\hat{\boldsymbol{\mu}}(s)) \right| \leq \sum_{s=0}^{t-1} K\epsilon_s \, .$$

Now, with the choice $\epsilon_t = (K^2+t)^{-1/2}/2$, one has

$$\sqrt{t+K^2} - K = \int_0^t \frac{ds}{2\sqrt{K^2+s}} \leq \sum_{s=0}^{t-1} \epsilon_s \leq \int_{-1}^{t-1} \frac{ds}{2\sqrt{K^2+s}} = \sqrt{t+K^2-1} - \sqrt{K^2-1} \, ,$$

which yields

$$\max_{1 \leq a \leq K} \left| N_a(t) - \sum_{s=0}^{t-1} w_a^{*}(\hat{\boldsymbol{\mu}}(s)) \right| \leq K - 1 + K\left(\sqrt{t+K^2-1} - \sqrt{K^2-1}\right) \leq K(1+\sqrt{t}) \, .$$

From (12), it also follows that

$$N_a(t) \geq \sum_{s=0}^{t-1} \epsilon_s - (K-1) \geq \sqrt{t+K^2} - K - (K-1) \geq \sqrt{t+K^2} - 2K,$$

which concludes the proof of Lemma 7.

Proof of Lemma 15. First, we prove by induction on k that

$$\max_{1\leq i\leq K} N_i(k) - P_i(k) \leq 1.$$

The statement is obviously true for $k=0$. Assume that it holds for some $k\geq 0$. For $i\neq I_{k+1}$ one has $N_i(k+1)-P_i(k+1) = N_i(k)-P_i(k)-p_i(k)\leq 1-p_i(k)\leq 1$, whereas $N_{I_{k+1}}(k+1) - P_{I_{k+1}}(k+1) = 1 + (N_{I_{k+1}}(k) - P_{I_{k+1}}(k+1) \leq 1$, using that $\sum_i (P_i(k+1)-N_i(k)) = 0$, hence the largest term in this sum (which is for $i = I_{k+1}$ by definition) is positive.

It follows that, for all k,

$$\begin{aligned}\max_{1\leq i\leq K} \big|N_i(k)-P_i(k)\big| &= \max\Big\{\max_{1\leq i\leq K} P_i(k)-N_i(k),\ \max_{1\leq i\leq K} N_i(k)-P_i(k)\Big\}\\ &\leq \max\{\sum_{1\leq i\leq K}\big(P_i(k)-N_i(k)\big)_+, 1\}\ .\end{aligned}$$

Introducing, for every $k\in\{1,\dots,n\}$

$$r_k = \sum_{i=1}^K \big(P_i(k)-N_i(k)\big)_+,$$

Lemma 15 follow from the following bound on r_k, that we prove by induction on k:

$$r_k \leq K-1.$$

Observe that $r_0 = 0$. For every $k\geq 0$, one can write

$$r_{k+1} = r_k + \sum_{i=1}^K p_i(k+1)\mathbb{1}_{(P_i(k+1)\geq N_i(k+1))} - \mathbb{1}_{(P_{I_{k+1}}(k+1)\geq N_{I_{k+1}}(k+1))}.$$

We distinguish two cases. If $(P_{I_{k+1}}(k+1)\geq N_{I_{k+1}}(k+1))$ one has

$$r_{k+1} \leq r_k + \sum_{i=1}^K p_i(k+1) - 1 = r_k \leq K-1.$$

If $(P_{I_{k+1}}(k+1) < N_{I_{k+1}}(k+1))$, which implies $P_{I_{k+1}}(k+1) - N_{I_{k+1}}(k)\leq 1$, one has

$$\begin{aligned} r_{k+1} &= \sum_{\substack{i=1\\ i\neq I_{k+1}}}^K (P_i(k+1)-N_i(k+1))_+ \leq \sum_{\substack{i=1\\ i\neq I_{k+1}}}^K \max_j (P_j(k+1)-N_j(k))\\ &= \sum_{\substack{i=1\\ i\neq I_{k+1}}}^K \big(P_{I_{k+1}}(k+1)-N_{I_{k+1}}(k)\big) \leq (K-1)\ , \end{aligned}$$

which concludes the proof.

Remark 16 *This result is probably overly pessimistic, but one cannot hope for an upper bound independent of K: for $p_i(k) = 1\{i\geq k\}/(K-k+1)$, by choosing at each step K the smallest index in the argmax defining I_k, one gets $N_K(K-1) = 0$ and $P_K(K-1) = 1/K + 1/(K-1) + \ldots + 1/2 \sim$* $\log(K)$.

B.2. Tracking a changing sequence that concentrates

Lemma 17 *Let K be a positive integer and let Σ_K be the simplex of dimension $K-1$. Let $g : \mathbb{N} \to \mathbb{R}$ be a non-decreasing function such that $g(0) = 0$, $g(n)/n \to 0$ when n tends to infinity and for all $k \geq 1$ and $\forall m \geq 1$,*

$$\inf\{k \in \mathbb{N} : g(k) \geq m\} > \inf\{k \in \mathbb{N} : g(k) \geq m-1\} + K.$$

Let $\hat{\lambda}(k)$ be a sequence of elements in Σ_K such that there exists $\lambda^* \in \Sigma_K$, there exists $\epsilon > 0$ and an integer $n_0(\epsilon)$ such that

$$\forall n \geq n_0, \quad \sup_{1\leq i\leq K} |\hat{\lambda}_i(k) - \lambda_i^*| \leq \epsilon \ .$$

Define $N(0) = 0$, and for every $k \in \{0, \dots, n-1\}$, $U_k = \{i : N_i(k) < g(k)\}$ and

$$I_{k+1} \in \begin{cases} \underset{i\in U_k}{\text{argmin}}\ N_i(k) & \textit{if } U_k \neq \emptyset, \\ \underset{i\in\{1,\dots,K\}}{\text{argmax}}\ \left[k\hat{\lambda}_i(k) - N_i(k)\right] & \textit{else}, \end{cases}$$

and for all i $N_i(k+1) = N_i(k) + \mathbb{1}_{(I_{k+1}=i)}$. Then for all $i \in \{1, \dots, K\}$, $N_i(n) > g(n) - 1$ and there exists $n_1 \geq n_0$ (that depends on ϵ) such that for all $n \geq n_1$,

$$\max_{1\leq i\leq K} \left| \frac{N_n(i)}{n} - \lambda_i^* \right| \leq 3(K-1)\epsilon \ .$$

First it is easy to check that $g(k) = (\sqrt{k} - K/2)_+$ satisfies the assumptions of Lemma 17. Lemma 8 then follows by choosing $\hat{\lambda}(k) = w^*(\hat{\mu}(k))$ and $\lambda^* = w^*(\boldsymbol{\mu})$. The constant n_1 in Lemma 17 depends on ϵ, hence the notation t_ϵ.

The proof of this Lemma 17 is inspired by the proof of Lemma 3 in Antos et al. (2008), although a different tracking procedure is analyzed.

Proof of Lemma 17. First, we justify that $N_i(n) > g(n) - 1$. For this purpose, we introduce for all $m \in \mathbb{N}$ the integer $n_m = \inf\{k \in \mathbb{N} : g(k) \geq m\}$. We also let $\mathcal{I}_m = \{n_m, \dots, n_{m+1}-1\}$. From our assumption on g, it follows that $|\mathcal{I}_m| > K$ and by definition, for all $k \in \mathcal{I}_m$, $m \leq g(k) < m+1$.

We prove by induction that the following statement holds for all m:

$$\forall k \in \mathcal{I}_m, \forall i, N_i(k) \geq m. \text{ Moreover for } k \geq n_m + K,\ U_k = \emptyset \text{ and } N_i(k) \geq m+1\ . \tag{13}$$

First, for all $k \in \mathcal{I}_0$, one has $U_k = \{i : N_i(k) = 0\}$. Therefore $(I_1, \dots, I_K)$ is a permutation of $(1, \dots, K)$, thus for $k \geq K = n_0 + K$, $N_i(k) \geq 1$, and $U_k = \emptyset$, and the statement holds for $m = 0$. Now let $m \geq 0$ such that the statement is true. From the inductive hypothesis, one has

$$\forall k \in \mathcal{I}_{m+1}, \forall i, N_i(k) \geq N_i(n_{m+1} - 1) \geq m+1\ .$$

Besides, as $g(k) < m+2$ for $k \in \mathcal{I}_{m+1}$, one has $U_k = \{i : N_i(k) = m+1\}$ and I_k is chosen among this set while it is non empty. For $k \geq n_{m+1} + K$, it holds that $U_k = \emptyset$ and $N_i(k) \geq m+2$ for all i. Thus the statement holds for $m+1$.

From the fact that (13) holds for all m, using that for $k \in \mathcal{I}_m$, $m > g(k) - 1$, it follows that for all k, for all i, $N_i(k) > g(k) - 1$.

Now for all $i \in \{1, \dots, K\}$, we introduce $E_{i,n} = N_i(n) - n\lambda_i^*$. Using that

$$\sum_{i=1}^{K} E_{i,n} = 0, \tag{14}$$

leads to

$$\sup_i |E_{i,n}| \leq (K-1) \sup_i \; E_{i,n}.$$

Indeed, for every i, one has $E_{i,n} \leq \sup_k E_{k,n}$ and

$$E_{i,n} = -\sum_{j \neq i} E_{j,n} \geq -\sum_{j \neq i} \sup_k E_{k,n} = -(K-1) \sup_k E_{k,n} \; .$$

To conclude the proof, we give an upper bound on $\sup_i E_{i,n}$, for n large enough. Let $n_0' \geq n_0$ such that

$$\forall n \geq n_0', \quad g(n) \leq 2n\epsilon \;\text{ and }\; 1/n \leq \epsilon \; .$$

We first show that for $n \geq n_0'$,

$$(I_{n+1} = i) \subseteq (E_{i,n} \leq 2n\epsilon) \tag{15}$$

To prove this, we write

$$(I_{n+1} = i) \subseteq \left(N_i(n) \leq g(n) \;\text{ or }\; i = \underset{1 \leq j \leq K}{\mathit{argmin}} \left(N_j(n) - n\hat{\lambda}_j(n) \right) \right) \; .$$

Now if $N_i(n) \leq g(n)$, one has $E_{i,n} \leq g(n) - n\lambda_i^* \leq g(n) \leq 2n\epsilon$, by definition of n_0'. In the second case, one has

$$\begin{aligned} N_i(n) - n\hat{\lambda}_i(n) &= \min_j \left(N_j(n) - n\hat{\lambda}_j(n) \right) \\ E_{i,n} + n(\lambda_i^* - \hat{\lambda}_i(n)) &= \min_j \left(E_{j,n} + n(\lambda_j^* - \hat{\lambda}_j(n) \right) \end{aligned}$$

Using the closeness of each $\hat{\lambda}_j(n)$ to the corresponding λ_j^*, as $n \geq n_0$, yields

$$E_{i,n} + n(\lambda_i^* - \hat{\lambda}_n(i)) \leq \min_j \left(E_{j,n} + n\epsilon \right) \leq n\epsilon,$$

where we use that $\min_j E_{j,n} \leq 0$ by (14). Using that $|\lambda_i^* - \hat{\lambda}_i(n)| \leq \epsilon$ as well, one obtains

$$E_{i,n} \leq 2n\epsilon \; .$$

This proves (15).

$E_{i,n}$ satisfies $E_{i,n+1} = E_{i,n} + \mathbb{1}_{(I_{n+1}=i)} - \lambda_i^*$, therefore, if $n \geq n_0'$,

$$E_{i,n+1} \quad \leq \quad E_{i,n} + \mathbb{1}_{(E_{i,n} \leq 2n\epsilon)} - \lambda_i^* \; .$$

We now prove by induction that for every $n \geq n_0'$, one has

$$E_{i,n} \leq \max(E_{i,n_0'}, 2n\epsilon + 1).$$

For $n = n_0'$, this statement clearly holds. Let $n \geq n_0'$ such that the statement holds. If $E_{i,n} \leq 2n\epsilon$, one has

$$\begin{aligned} E_{i,n+1} &\leq 2n\epsilon + 1 - \lambda_i^* \leq 2n\epsilon + 1 \leq \max(E_{i,n_0'}, 2(n)\epsilon + 1) \\ &\leq \max(E_{i,n_0'}, 2(n+1)\epsilon + 1). \end{aligned}$$

If $E_{i,n} > 2n\epsilon$, the indicator is zero and

$$E_{i,n+1} \leq \max(E_{i,n_0'}, 2n\epsilon + 1) - \lambda_i^* \leq \max(E_{i,n_0'}, 2(n+1)\epsilon + 1)\ ,$$

which concludes the induction.

For all $n \geq n_0'$, using that $E_{i,n_0'} \leq n_0'$ and $1/n \leq \epsilon$, it follows that

$$\sup_i \left|\frac{E_{i,n}}{n}\right| \leq (K-1)\max\left(2\epsilon + \frac{1}{n}, \frac{n_0'}{n}\right) \leq (K-1)\max\left(3\epsilon, \frac{n_0'}{n}\right).$$

Hence there exists $n_1 \geq n_0'$ such that, for all $n \geq n_1$,

$$\sup_i \left|\frac{E_{i,n}}{n}\right| \leq 3(K-1)\epsilon\ ,$$

which concludes the proof.

B.3. Proof of Proposition 9

Because of the forced-exploration step, both tracking strategies satisfy $\forall a, N_a(t) \to \infty$. Thus, from the law of large number, the event

$$\mathcal{E} = \left(\hat{\boldsymbol{\mu}}(t) \underset{t\to\infty}{\rightarrow} \boldsymbol{\mu}\right)$$

is of probability one. For C-Tracking, it follows from Lemma 7 that

$$\left|\frac{N_a(t)}{t} - \frac{1}{t}\sum_{s=0}^{t-1} w_a^*(\boldsymbol{\mu}(s))\right| \leq \frac{K(\sqrt{t}+1)}{t}\ .$$

By continuity of w^* in $\boldsymbol{\mu}$, for all a, $w_a^*(\hat{\boldsymbol{\mu}}(t)) \to w_a^*(\boldsymbol{\mu})$. Using moreover the Cesaro lemma, one obtains that, on $\mathcal{E}$, $N_a(t)/t \to w_a^*(\boldsymbol{\mu})$. For D-Tracking, we first use that for $\omega \in \mathcal{E}$, there exists $t_0(\epsilon)$ such that

$$\sup_{t\geq t_0(\epsilon)} \max_a |w_a^*(\hat{\boldsymbol{\mu}}(t)) - w_a^*(\boldsymbol{\mu})| \leq \frac{\epsilon}{3(K-1)}\ ,$$

by continuity of the function $\boldsymbol{\lambda} \mapsto w^*(\boldsymbol{\lambda})$ in $\boldsymbol{\mu}$. Hence, using Lemma 8, there exists $t_\epsilon \geq t_0$ such that for all $t \geq t_\epsilon$,

$$\max_a \left|\frac{N_a(t)}{t} - w_a^*(\boldsymbol{\mu})\right| \leq \epsilon\ .$$

Hence, for this $\omega \in \mathcal{E}$, $N_a(t)/t(\omega) \to w_a^*(\boldsymbol{\mu})$ for all a.

Appendix C. PAC guarantees

C.1. Proof of Proposition 10.

Recall that

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) &\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists a \in \mathcal{A} \setminus a^*, \exists t \in \mathbb{N} : Z_{a,a^*}(t) > \beta(t,\delta)\right)\\ &\leq \sum_{a \in \mathcal{A}\setminus A^*} \mathbb{P}_{\boldsymbol{\mu}}(T_{a,a^*} < \infty)\ ,\end{aligned}$$

where $T_{a,b} := \inf\{t \in \mathbb{N} : Z_{a,b}(t) > \beta(t,\delta)\}$. To conclude the proof, we now show that for $\beta(t,\delta)$ as in Proposition 10, for any a, b such that $\mu_a < \mu_b$, one has

$$\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) \leq \frac{\delta}{K-1}.$$

Let a,b be such that $\mu_a < \mu_b$. One introduces $f_{\boldsymbol{\mu}}(\underline{x}_t, \underline{a}_t)$ the likelihood of observing the sequence of rewards $\underline{x}_t = (x_1, \ldots, x_t)$ and sequence of actions $\underline{a}_t = (a_1, \ldots, a_t)$. One has

$$f_{\boldsymbol{\mu}}(\underline{x}_t, \underline{a}_t) = \prod_{i=1}^{K} p_{\mu_i}\left(\underline{x}_t^i\right) \times \left[p(a_1)\prod_{s=2}^{t} p(a_s|\underline{x}_{s-1}, \underline{a}_{s-1})\right],$$

where $\underline{x}_t^i$ is a vector that gathers the sequence of successive rewards from arm i up to time t (that is a function of both $\underline{x}_t$ and $\underline{a}_t$). $f_{\boldsymbol{\mu}}(\underline{x}_t, \underline{a}_t)$ is a probability density on $\mathcal{X}^t \times \mathcal{A}^t$. For any density h supported on $\mathcal{I}$, the (partially) integrated likelihood

$$I_h(\underline{x}_t, \underline{a}_t) = \prod_{i \in \mathcal{A}\setminus\{a,b\}} p_{\mu_i}\left(\underline{x}_t^i\right) \left(\int_{\mathbb{R}} p_u\left(\underline{x}_t^a\right) h(u)du\right) \left(\int_{\mathbb{R}} p_u\left(\underline{x}_t^b\right) h(u)du\right) \times \left[p(a_1)\prod_{s=2}^{t} p(a_s|\underline{x}_{s-1}, \underline{a}_{s-1})\right]$$

is also a probability distribution.

On the event $(T_{a,b} = t)$, $Z_{a,b}(t)$ exceeds for the first time the threshold $\beta(t,\delta)$, which implies in particular (from the definition of $Z_{a,b}(t)$) that

$$1 \leq e^{-\beta(t,\delta)} \frac{\max_{\mu_a' \geq \mu_b'} p_{\mu_a'}(\underline{X}_t^a) p_{\mu_b'}(\underline{X}_t^b)}{\max_{\mu_a' \leq \mu_b'} p_{\mu_a'}(\underline{X}_t^a) p_{\mu_b'}(\underline{X}_t^b)}\ .$$

We use this fact in the first inequality below, whereas the second inequality is based on the fact that $\boldsymbol{\mu}$ satisfies $\mu_a < \mu_b$:

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) &= \sum_{t=1}^{\infty} \mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} = t) = \sum_{t=1}^{\infty} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)}\right]\\ &\leq \sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)} \frac{\max_{\mu_a' \geq \mu_b'} p_{\mu_a'}(\underline{X}_t^a) p_{\mu_b'}(\underline{X}_t^b)}{\max_{\mu_a' \leq \mu_b'} p_{\mu_a'}(\underline{X}_t^a) p_{\mu_b'}(\underline{X}_t^b)}\right]\\ &\leq \sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \underbrace{\mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{1}_{(T_{a,b}=t)} \frac{\max_{\mu_a' \geq \mu_b'} p_{\mu_a'}(\underline{X}_t^a) p_{\mu_b'}(\underline{X}_t^b)}{p_{\mu_a}(\underline{X}_t^a) p_{\mu_b}(\underline{X}_t^b)}\right]}_{(*)}\ .\end{aligned}$$

We now explicit the expectation $(*)$ for Bernoulli distributions.

$$(*) = \sum_{\underline{a}_t \in \mathcal{A}^t} \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t, \underline{a}_t) \max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{x}_t^a) p_{\mu'_b}(\underline{x}_t^b) \prod_{i \in \mathcal{A} \backslash \{a,b\}} p_{\mu_i}(\underline{x}_t^i) \left[p(a_1) \prod_{s=2}^{t} p(a_s | \underline{x}_{s-1}, \underline{a}_{s-1}) \right]$$

Introducing, for a vector x,

$$\mathrm{kt}(x) = \int_0^1 \frac{1}{\sqrt{\pi u(1-u)}} p_u(x) \mathrm{d}u \ ,$$

we have that

$$\max_{\mu'_a \geq \mu'_b} p_{\mu'_a}(\underline{x}_t^a) p_{\mu'_b}(\underline{x}_t^b) \leq 2t \times \mathrm{kt}(\underline{x}_t^a) \mathrm{kt}(\underline{x}_t^b),$$

which follows from Lemma 11 stated in the main text and the fact that for n_a, n_b are such that $n_a + n_b \leq t$, $4\sqrt{n_a}\sqrt{n_b} \leq 2(n_a + n_b) \leq 2t$. Using this inequality to upper bound $(*)$, one recognize the integrated likelihood associated to the density $h(u) = \frac{1}{\sqrt{\pi u(1-u)}} \mathbb{1}_{[0,1]}(u)$:

$$(*) \leq 2t \sum_{\underline{a}_t \in \mathcal{A}^t} \sum_{\underline{x}_t \in \{0,1\}^t} \mathbb{1}_{(T_{a,b}=t)}(\underline{x}_t, \underline{a}_t) I_h(\underline{x}_t, \underline{a}_t) = 2t \tilde{\mathbb{P}}(T_{a,b} = t) \ ,$$

where $\tilde{\mathbb{P}}$ is an alternative probabilistic model, under which μ_a and μ_b are drawn from a $\mathrm{Beta}(1/2, 1/2)$ (prior) distribution at the beginning of the bandit game. Finally, using the explicit expression of $\beta(t, \delta)$,

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}(T_{a,b} < \infty) &\leq \sum_{t=1}^{\infty} 2t e^{-\beta(t,\delta)} \tilde{\mathbb{P}}(T_{a,b} = t) \leq \frac{\delta}{K-1} \sum_{t=1}^{\infty} \tilde{\mathbb{P}}(T_{a,b} = t) \\ &= \frac{\delta}{K-1} \tilde{\mathbb{P}}(T_{a,b} < \infty) \leq \frac{\delta}{K-1} \ , \end{aligned}$$

which concludes the proof.

C.2. Proof of Proposition 12.

The proof relies on the fact that $Z_{a,b}(t)$ can be expressed using function I_α introduced in Definition 3. An interesting property of this function, that we use below, is the following. It can be checked that if $x > y$,

$$I_\alpha(x, y) = \inf_{x' < y'} \left[\alpha d(x, x') + (1-\alpha) d(y, y') \right] .$$

For every a, b that are such that $\mu_a < \mu_b$ and $\hat{\mu}_a(t) > \hat{\mu}_b(t)$, one has the following inequality:

$$\begin{aligned} Z_{a,b}(t) &= (N_a(t) + N_b(t)) I_{\frac{N_a(t)}{N_a(t)+N_b(t)}}(\hat{\mu}_a(t), \hat{\mu}_b(t)) \\ &= \inf_{\mu'_a < \mu'_b} N_a(t) d(\hat{\mu}_a(t), \mu'_a) + N_b(t) d(\hat{\mu}_b(t), \mu'_b) \\ &\leq N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) + N_b(t) d(\hat{\mu}_b(t), \mu_b) \ . \end{aligned}$$

One has

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*) &\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists a \in \mathcal{A} \setminus a^*, \exists t \in \mathbb{N} : \hat{\mu}_a(t) > \hat{\mu}_{a^*}(t), Z_{a,a^*}(t) > \beta(t,\delta)\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\big(\exists t \in \mathbb{N} : \exists a \in \mathcal{A} \setminus a^* : N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) + N_{a^*}(t) d(\hat{\mu}_{a^*}(t), \mu_{a^*}) \geq \beta(t,\delta)\big) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \beta(t,\delta)\right) \\
&\leq \sum_{t=1}^{\infty} e^{K+1} \left(\frac{\beta(t,\delta)^2 \log(t)}{K}\right)^K e^{-\beta(t,\delta)} .
\end{aligned}$$

The last inequality follows from a union bound and Theorem 2 of Magureanu et al. (2014), originally stated for Bernoulli distributions but whose generalization to one-parameter exponential families is straightforward. Hence, with an exploration rate of the form $\beta(t,\delta) = \log(Ct^\alpha/\delta)$, for $\alpha > 1$, choosing C satisfying

$$\sum_{t=1}^{\infty} \frac{e^{K+1}}{K^K} \frac{(\log^2(Ct^\alpha) \log t)^K}{t^\alpha} \leq C$$

yields a probability of error upper bounded by δ.

Appendix D. Expected sample complexity analysis

The proof of Theorem 14 relies on two ingredients: a concentration result for the empirical mean $\hat{\boldsymbol{\mu}}(t)$, that follows from the forced exploration (Lemma 19) and the tracking lemma associated to the sampling strategy used (Lemma 7, Lemma 8). We start with a small technical lemma that can be checked directly, or that can be seen as a by-product of well-known bounds on the Lambert W function.

Lemma 18 *For every $\alpha \in [1, e/2]$, for any two constants $c_1, c_2 > 0$,*

$$x = \frac{\alpha}{c_1}\left[\log\left(\frac{c_2 e}{c_1^\alpha}\right) + \log\log\left(\frac{c_2}{c_1^\alpha}\right)\right]$$

is such that $c_1 x \geq \log(c_2 x^\alpha)$.

D.1. Proof of Theorem 14

To ease the notation, we assume that the bandit model $\boldsymbol{\mu}$ is such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. Let $\epsilon > 0$. From the continuity of w^* in $\boldsymbol{\mu}$, there exists $\xi = \xi(\epsilon) \leq (\mu_1 - \mu_2)/4$ such that

$$\mathcal{I}_\epsilon := [\mu_1 - \xi, \mu_1 + \xi] \times \cdots \times [\mu_K - \xi, \mu_K + \xi]$$

is such that for all $\boldsymbol{\mu}' \in \mathcal{I}_\epsilon$,

$$\max_a |w_a^*(\boldsymbol{\mu}') - w_a^*(\boldsymbol{\mu})| \leq \epsilon.$$

In particular, whenever $\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon$, the empirical best arm is $\hat{a}_t = 1$.

Let $T \in \mathbb{N}$ and define $h(T) := T^{1/4}$ and the event

$$\mathcal{E}_T(\epsilon) = \bigcap_{t=h(T)}^{T} \left(\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon\right).$$

The following lemma is a consequence of the "forced exploration" performed by the algorithm, that ensures that each arm is drawn at least of order $\sqrt{t}$ times at round t.

Lemma 19 *There exist two constants B, C (that depend on $\boldsymbol{\mu}$ and ϵ) such that*

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8}).$$

Then, exploiting the corresponding tracking Lemma, one can prove the following

Lemma 20 *There exists a constant T_ϵ such that for $T \geq T_\epsilon$, it holds that on $\mathcal{E}_T$, for either C-Tracking or D-Tracking,*

$$\forall t \geq \sqrt{T},\ \max_a \left| \frac{N_a(t)}{t} - w_a^*(\mu) \right| \leq 3(K-1)\epsilon$$

Proof. This statement is obvious for D-Tracking, just by definition of $\mathcal{I}_\epsilon$ and by Lemma 8. For C-Tracking, for any $t \geq \sqrt{T} = h(T)^2$, using Lemma 7, one can write, for all a,

$$\begin{aligned}
\left| \frac{N_a(t)}{t} - w_a^*(\boldsymbol{\mu}) \right| &\leq \left| \frac{N_a(t)}{t} - \frac{1}{t}\sum_{s=0}^{t-1} w_a^*(\hat{\boldsymbol{\mu}}(s)) \right| + \left| \frac{1}{t}\sum_{s=0}^{t-1} w_a^*(\hat{\boldsymbol{\mu}}(s)) - w_a^*(\boldsymbol{\mu}) \right| \\
&\leq \frac{K(\sqrt{t}+1)}{t} + \frac{h(T)}{t} + \frac{1}{t}\sum_{t=h(T)}^{t-1} |w_a^*(\hat{\boldsymbol{\mu}}(s)) - w_a^*(\boldsymbol{\mu})| \\
&\leq \frac{2K}{T^{1/4}} + \frac{1}{h(T)} + \epsilon = \frac{2K+1}{T^{1/4}} + \epsilon \leq 3\epsilon\ ,
\end{aligned}$$

whenever $T \geq ((2K+1)/2\epsilon)^4$.

□

On the event $\mathcal{E}_T$, it holds for $t \geq h(T)$ that $\hat{a}_t = 1$ and the Chernoff stopping statistic rewrites

$$\begin{aligned}
\max_a \min_{b \neq a} Z_{a,b}(t) &= \min_{a \neq 1} Z_{1,a}(t) = \min_{a \neq 1} N_1(t) d(\hat{\mu}_1(t), \hat{\mu}_{1,a}(t)) + N_a(t) d(\hat{\mu}_a(t), \mu_{1,a}(t)) \\
&= t \left[\min_{a \neq 1} \left(\frac{N_1(t)}{t} + \frac{N_a(t)}{t} \right) I_{\frac{N_1(t)/t}{N_1(t)/t + N_a(t)/t}} (\hat{\mu}_1(t), \hat{\mu}_a(t)) \right] \\
&= tg\left(\hat{\boldsymbol{\mu}}(t), \left(\frac{N_a(t)}{t} \right)_{a=1}^K \right)\ ,
\end{aligned}$$

where we introduce the function

$$g(\boldsymbol{\mu}', \boldsymbol{w}') = \min_{a \neq 1} (w_1' + w_a') I_{\frac{w_1'}{w_1' + w_a'}} (\mu_1', \mu_a').$$

Using Lemma 20, for $T \geq T_\epsilon$, introducing

$$C_\epsilon^*(\boldsymbol{\mu}) = \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}|| \leq \alpha(\epsilon) \\ \boldsymbol{w}' : ||\boldsymbol{w}' - w^*(\boldsymbol{\mu})|| \leq 2(K-1)\epsilon}} g(\boldsymbol{\mu}', \boldsymbol{w}')\ ,$$

on the event $\mathcal{E}_T$ it holds that for every $t \geq \sqrt{T}$,

$$\left(\max_a \min_{b \neq a} \; Z_{a,b}(t) \geq t C_\epsilon^*(\boldsymbol{\mu}) \right) \; .$$

Let $T \geq T_\epsilon$. On $\mathcal{E}_T$,

$$\begin{aligned} \min(\tau_\delta, T) & \leq & \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(\tau_\delta > t)} \leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{\left(\max_a \min_{b \neq a} \; Z_{a,b}(t) \leq \beta(t,\delta)\right)} \\ & \leq & \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(t C_\epsilon^*(\boldsymbol{\mu}) \leq \beta(T,\delta))} \leq \sqrt{T} + \frac{\beta(T,\delta)}{C_\epsilon^*(\boldsymbol{\mu})} \; . \end{aligned}$$

Introducing

$$T_0(\delta) = \inf \left\{ T \in \mathbb{N} : \sqrt{T} + \frac{\beta(T,\delta)}{C_\epsilon^*(\boldsymbol{\mu})} \leq T \right\},$$

for every $T \geq \max(T_0(\delta), T_\epsilon)$, one has $\mathcal{E}_T \subseteq (\tau_\delta \leq T)$, therefore

$$\mathbb{P}_{\boldsymbol{\mu}} \left(\tau_\delta > T \right) \leq \mathbb{P}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8})$$

and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \leq T_0(\delta) + T_\epsilon + \sum_{T=1}^{\infty} BT \exp(-CT^{1/8}) \; .$$

We now provide an upper bound on $T_0(\delta)$. Letting $\eta > 0$ and introducing the constant

$$C(\eta) = \inf\{T \in \mathbb{N} : T - \sqrt{T} \geq T/(1+\eta)\}$$

one has

$$\begin{aligned} T_0(\delta) & \leq & C(\eta) + \inf \left\{ T \in \mathbb{N} : \frac{1}{C_\epsilon^*(\boldsymbol{\mu})} \log \left(\frac{r(T)}{\delta} \right) \leq \frac{T}{1+\eta} \right\} \\ & \leq & C(\eta) + \inf \left\{ T \in \mathbb{N} : \frac{C_\epsilon^*(\boldsymbol{\mu})}{1+\eta} T \geq \log \left(\frac{Dt^{1+\alpha}}{\delta} \right) \right\}, \end{aligned}$$

where the constant D is such that $r(T) \leq DT^\alpha$. Using again the technical Lemma 18, one obtains, for $\alpha \in [1, e/2]$,

$$T_0(\delta) \leq C(\eta) + \frac{\alpha(1+\eta)}{C_\epsilon(\boldsymbol{\mu})} \left[\log \left(\frac{De(1+\eta)^\alpha}{\delta (C_\epsilon^*(\boldsymbol{\mu}))^\alpha} \right) + \log\log \left(\frac{D(1+\eta)^\alpha}{\delta (C_\epsilon^*(\boldsymbol{\mu}))^\alpha} \right) \right] .$$

This last upper bound yields, for every $\eta > 0$ and $\epsilon > 0$,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \frac{\alpha(1+\eta)}{C_\epsilon^*(\boldsymbol{\mu})}.$$

Letting η and ϵ go to zero and using that, by continuity of g and by definition of w^*,

$$\lim_{\epsilon \to 0} C_\epsilon^*(\boldsymbol{\mu}) = T^*(\boldsymbol{\mu})^{-1}$$

yields

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}) \; .$$

D.2. Proof of Lemma 19.

$$\mathbb{P}\left(\mathcal{E}_T^c\right) \leq \sum_{t=h(T)}^{T} \mathbb{P}\left(\hat{\boldsymbol{\mu}}(t) \notin \mathcal{I}_\epsilon\right) = \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \left[\mathbb{P}\left(\hat{\mu}_a(t) \leq \mu_a - \xi\right) + \mathbb{P}\left(\hat{\mu}_a(t) \geq \mu_a + \xi\right)\right] \ .$$

Let T be such that $h(T) \geq K^2$. Then for $t \geq h(T)$ one has $N_a(t) \geq (\sqrt{t} - K/2)_+ - 1 \geq \sqrt{t} - K$ for every arm a. Let $\hat{\mu}_{a,s}$ be the empirical mean of the first s rewards from arm a (such that $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$). Using a union bound and Chernoff inequality, one can write

$$\begin{aligned} \mathbb{P}\left(\hat{\mu}_a(t) \leq \mu_a - \xi\right) &= \mathbb{P}\left(\hat{\mu}_a(t) \leq \mu_a - \xi, N_a(t) \geq \sqrt{t}\right) \leq \sum_{s=\sqrt{t}-K}^{t} \mathbb{P}\left(\hat{\mu}_{a,s} \leq \mu_a - \xi\right) \\ &\leq \sum_{s=\sqrt{t}-K}^{t} \exp(-sd(\mu_a - \xi, \mu_a)) \leq \frac{e^{-(\sqrt{t}-K)d(\mu_a-\xi,\mu_a)}}{1 - e^{-d(\mu_a-\xi,\mu_a)}} \ . \end{aligned}$$

Similarly, one can prove that

$$\mathbb{P}\left(\hat{\mu}_a(t) \geq \mu_a + \xi\right) \leq \frac{e^{-(\sqrt{t}-K)d(\mu_a+\xi,\mu_a)}}{1 - e^{-d(\mu_a+\xi,\mu_a)}}.$$

Finally, letting

$$C = \min_a \left(d(\mu_a - \xi, \mu_a) \wedge d(\mu_a + \xi, \mu_a)\right) \ \text{ and } \ B = \sum_{a=1}^{K} \left(\frac{e^{Kd(\mu_a-\xi,\mu_a)}}{1 - e^{-d(\mu_a-\xi,\mu_a)}} + \frac{e^{Kd(\mu_a+\xi,\mu_a)}}{1 - e^{-d(\mu_a+\xi,\mu_a)}}\right) \ ,$$

one obtains

$$\mathbb{P}\left(\mathcal{E}_T^c\right) \leq \sum_{t=h(T)}^{T} B \exp(-\sqrt{t}C) \leq BT \exp(-\sqrt{h(T)}C) = BT \exp(-CT^{1/8}) \ .$$