On the Complexity of A/B Testing

Emilie Kaufmann, Olivier Cappé, Aurélien Garivier

JMLR: Workshop and Conference Proceedings vol 35:1–22, 2014

On the Complexity of A/B Testing

Emilie Kaufmann KAUFMANN@TELECOM-PARISTECH.FR
LTCI, Télécom ParisTech & CNRS

Olivier Cappé CAPPE@TELECOM-PARISTECH.FR
LTCI, Télécom ParisTech & CNRS

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR
Institut de Mathématiques de Toulouse, Université Paul Sabatier

Abstract

A/B testing refers to the task of determining the best option among two alternatives that yield random outcomes. We provide distribution-dependent lower bounds for the performance of A/B testing that improve over the results currently available both in the fixed-confidence (or δ-PAC) and fixed-budget settings. When the distribution of the outcomes are Gaussian, we prove that the complexity of the fixed-confidence and fixed-budget settings are equivalent, and that uniform sampling of both alternatives is optimal only in the case of equal variances. In the common variance case, we also provide a stopping rule that terminates faster than existing fixed-confidence algorithms. In the case of Bernoulli distributions, we show that the complexity of fixed-budget setting is smaller than that of fixed-confidence setting and that uniform sampling of both alternatives—though not optimal—is advisable in practice when combined with an appropriate stopping criterion.

Keywords: Sequential testing. Best arm identification. Bandit models. Sample complexity.

1. Introduction

A/B Testing is a popular procedure used, for instance, for website optimization: two versions of a webpage, say A and B, are empirically compared by being presented to users. Each user only sees one of the two versions, and the goal is to determine which version is preferable. We assume that the users provide a real-valued index of the quality of the pages, which is modeled by probability distributions ν_A and ν_B, with respective means μ_A and μ_B. For example, a standard objective is to determine which webpage has the highest conversion rate (probability that a user actually becomes a customer) by receiving binary feedback from the users.

Methods for A/B Testing are often viewed as statistical tests of the hypothesis $H_0 : (\mu_A \leq \mu_B)$ against $H_1 : (\mu_A > \mu_B)$. One may consider either *classical tests*, based on a number of samples n_A and n_B from each distribution fixed before the experiment, or *sequential tests*, based on paired samples (X_s, Y_s) of ν_A, ν_B and in which a randomized stopping rule determines when the experiment is to be terminated. In both of these test settings, the sampling schedule is determined in advance, which is a possible source of sub-optimality as A/B Testing algorithms could take advantage of past observations to provide a smarter choice of the page to be displayed to the next user. In the sequel, we investigate whether A/B Testing could benefit from an adaptive sampling schedule. Ignoring the possible long-term effects on users of presenting one or the other option, we consider it as a particular instance of *best arm identification in a two-armed bandit model*.

A two-armed bandit model consists of two unknown probability distributions on $\mathbb{R}$, ν_1 and ν_2, sometimes referred to as arms or options (webpages in our motivating example). Arm a has expectation μ_a. At each time $t = 1, 2, \ldots$, an agent chooses an option $A_t \in \{1, 2\}$ and receives an independent draw Z_t of the corresponding arm ν_{A_t}. We denote by $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) the probability law (resp. expectation) of the corresponding process (Z_t). We assume that the bandit model $\nu = (\nu_1, \nu_2)$ belongs to a class $\mathcal{M}$ such that for all $\nu \in \mathcal{M}$, $\mu_1 \neq \mu_2$. In order to identify the best arm, that is the arm a^* with highest expectation, the agent must use a strategy defining which arms to sample from, when to stop sampling, and which arm $\hat{a}$ to choose. The *sampling rule* determines how, at time t, the arm A_t is chosen based on the past observations; in other words, A_t is $\mathcal{F}_{t-1}$–measurable, with $\mathcal{F}_t = \sigma(A_1, Z_1, \ldots, A_t, Z_t)$. The *stopping rule* τ is a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$ satisfying $\mathbb{P}_\nu(\tau < +\infty) = 1$. The *recommendation rule* is a $\mathcal{F}_\tau$-measurable random arm $\hat{a} \in \{1, 2\}$. This triple $((A_t), \tau, \hat{a})$ entirely determines the strategy, which we denote in the sequel by $\mathcal{A}$. As discussed before, statistical tests correspond to strategies that sample the arms in a round-robin fashion, which we will refer to as *uniform sampling*.

In the bandit literature, two different settings have been considered. In the *fixed-confidence setting*, a risk parameter δ is fixed. A strategy $\mathcal{A}$ is called δ-*PAC* if, for every choice of $\nu \in \mathcal{M}$, $\mathbb{P}_\nu(\hat{a} = a^*) \geq 1 - \delta$. The goal is, among the δ-*PAC* strategies, to minimize the expected number of draws $\mathbb{E}_\nu[\tau]$. In the *fixed-budget setting*, the number of draws τ is fixed in advance ($\tau = t$ almost surely) and the goal is to choose the recommendation rule so as to minimize $p_t(\nu) := \mathbb{P}_\nu(\hat{a} \neq a^*)$. In the fixed-budget setting, a strategy $\mathcal{A}$ is called *consistent* if, for every choice of $\nu \in \mathcal{M}$, $p_t(\nu) \underset{t\to\infty}{\to} 0$.

In order to unify and compare these approaches, we define the *complexity* $\kappa_C(\nu)$ (resp. $\kappa_B(\nu)$) of best arm identification in the fixed-confidence (resp. fixed-budget) setting, as follows:

$$\kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta-\mathrm{PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}, \qquad \kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A}\ \text{consistent}} \left(\limsup_{t\to\infty} -\frac{1}{t} \log p_t(\nu) \right)^{-1}.$$

Heuristically, for a given bandit model ν and a given $\delta > 0$, a fixed-confidence optimal strategy uses an average number of samples of order $\kappa_C(\nu)\log(1/\delta)$, whereas a fixed-budget optimal strategy uses approximately $t = \kappa_B(\nu)\log(1/\delta)$ draws in order to ensure a probability of error at most equal to δ. Most of the existing performance bounds for the fixed confidence and fixed budget settings can be expressed using these complexity measures.

The main goal of this paper is to determine κ_{C} and κ_{B} for important classes of parametric bandit models, allowing for a comparison between the fixed-confidence and fixed-budget settings. Classical sequential testing theory provides a first element in that direction in the simpler case of fully specified alternatives. Consider for instance the case where ν_1 and ν_2 are Gaussian laws with the same known variance σ^2, the means μ_1 and μ_2 being known up to a permutation. Denoting by P the joint distribution of the paired samples (X_s, Y_s), one must choose between the hypotheses $H_0 : P = \mathcal{N}\left(\mu_1, \sigma^2\right) \otimes \mathcal{N}\left(\mu_2, \sigma^2\right)$ and $H_1 : P = \mathcal{N}\left(\mu_2, \sigma^2\right) \otimes \mathcal{N}\left(\mu_1, \sigma^2\right)$. It is known since Wald (1945) that among the sequential tests such that type I and type II probabilities of error are both smaller than δ, the Sequential Probability Ratio Test (SPRT) minimizes the expected number of required samples, and is such that $\mathbb{E}_\nu[\tau] = 2\sigma^2/(\mu_1 - \mu_2)^2 \log(1/\delta)$. However, the batch test that minimizes both probabilities of error is the Likelihood Ratio test; it can be shown to require a sample size of order $8\sigma^2/(\mu_1 - \mu_2)^2 \log(1/\delta)$ in order to ensure that both type I and type II error probabilities are smaller than δ. Thus, when the sampling strategy is uniform and the parameters are

known, there is a clear gain in using randomized stopping strategies. We show in the following that this conclusion is not valid anymore when the values of μ_1 and μ_2 are not assumed to be known.

Related works. Bandit models have received a considerable interest since their introduction by Thompson (1933). An important focus was set on a slightly different perspective, in which each observation is considered as a reward: the agent aims at maximizing the cumulative rewards obtained until some horizon t. Equivalently, his goal is to minimize the *regret* $R_t(\nu) = t\mu_{[1]} - \mathbb{E}_\nu\left[\sum_{s=1}^t Z_s\right]$. Regret minimization, which is paradigmatic of the so-called *exploration versus exploitation dilemma*, was introduced by Robbins (1952) and its complexity is well understood for parametric bandits. In generic one-parameter models, Lai and Robbins (1985) prove that, with a proper notion of consistency adapted to regret minimization,

$$\inf_{A \text{ consistent}} \liminf_{t\to\infty} \frac{R_t(\nu)}{\log t} \geq \sum_{a:\mu_a<\mu_{[1]}} \frac{(\mu_{[1]} - \mu_a)}{\mathrm{KL}(\nu_a, \nu_{[1]})} ,$$

where $\mathrm{KL}(\nu_i, \nu_j)$ denotes the Kullback-Leibler divergence between distributions ν_i and ν_j. Since then, non-asymptotic analyses of efficient algorithms matching this bound have been proposed. Optimal algorithms include the KL-UCB algorithm of Cappé et al. (2013)—a variant of UCB1 (Auer et al. (2002)) using informational upper bounds, Thompson Sampling (Kaufmann et al. (2012b); Agrawal and Goyal (2013)), the DMED algorithm (Honda and Takemura, 2011) and Bayes-UCB Kaufmann et al. (2012a). This paper is a first step in the attempt to similarly characterize the complexity of *pure exploration*, where the goal is to determine the best arms without trying to maximize the cumulated observations.

The problem of best arm identification has received an important interest in the 1950s as a particular case of 'ranking and identification problems'. The literature on the subject goes beyond two-armed bandit models to finding the $m > 1$ best arms among $K > 2$ arms, and sometimes introduces a relaxation parameter $\epsilon > 0$, such that arms within ϵ of the best arm should be recommended. In the sequel, we always particularize the existing results to the two-armed bandit models presented above. The first advances on this topic are summarized in the monograph by Bechhofer et al. (1968), who only consider the fixed-confidence setting. In the same setting, algorithms introduced more recently by Even-Dar et al. (2006); Kalyanakrishnan et al. (2012); Gabillon et al. (2012) can be used to find the best arm in a two-armed bounded bandit model, in which ν_1 and ν_2 are probability distributions on $[0, 1]$. Combining the upper bound on $\mathbb{E}_\nu[\tau]$ for the LUCB algorithm of Kalyanakrishnan et al. (2012) with the lower bound following from the work of Mannor and Tsitsiklis (2004), it can be shown that for bounded bandit models such that $\mu_a \in [0, 1-\alpha]$ for $a \in \{1, 2\}$, there exists a constant C_α for which

$$C_\alpha/(\mu_1 - \mu_2)^2 \leq \kappa_C(\nu) \leq 584/(\mu_1 - \mu_2)^2.$$

The fixed-budget setting has been studied recently by Audibert et al. (2010); Bubeck et al. (2013). In two-armed bandit problems, the algorithms introduced in these papers boil down to sampling each arm $t/2$ times—t denoting the total budget—and recommending the empirical best arm. A simple upper bound on the probability of error of this strategy can be derived, and this result paired with the lower bound of Audibert et al. (2010) yields, for bounded bandit models such that $\mu_a \in [\alpha; 1-\alpha]$ for $a \in \{1, 2\}$:

$$(2/5)\alpha(1-\alpha)/(\mu_1 - \mu_2)^2 \leq \kappa_B(\nu) \leq 2/(\mu_1 - \mu_2)^2.$$

Bubeck et al. (2011) show that in the fixed-budget setting any sampling strategy designed to minimize regret performs poorly with respect to the *simple regret* $r_t := \mu^* - \mu_{\hat{S}_1}$, a quantity closely related to the probability $p_t(\nu)$ of recommending the wrong arm. Therefore, good strategies for best arm identification have to be quite different from UCB-like strategies. We will show below that the complexities $\kappa_B(\nu)$ and $\kappa_C(\nu)$ of pure-exploration involve information terms that are different from the Kullback-Leibler divergence featured in Lai and Robbins' lower bound on regret.

Contents of the paper. Compared to existing results, we provide general lower bounds on $\kappa_B(\nu)$ and $\kappa_C(\nu)$ that: *(i)* are tighter, leading in specific parametric cases to a precise evaluation of these complexities; *(ii)* do not require unnecessary support assumptions; and *(*iii) are stated in terms of information divergences between the distributions ν_1 and ν_2 rather than in terms of the gap $\mu_1 - \mu_2$. As can be expected, we will indeed confirm that the inverse of the squared gap $(\mu_1 - \mu_2)^2$ is the relevant measure of complexity only in the Gaussian case, and an approximation (in the spirit of Pinsker's inequality) for sub-Gaussian distributions.

Lower bounds on the sample complexity (resp. probability of error) of algorithms using the uniform sampling strategy in the fixed-confidence (resp. fixed-budget) setting are also derived and we show that for Gaussian bandit models with different variances, there is a significant gain in using a non-uniform sampling strategy. For Bernoulli bandits however, we show that little can be gained by departing from uniform sampling, and we therefore propose close-to-optimal tests both for the batch and sequential settings. For Gaussian bandits with a known common variance the optimal algorithm uses uniform sampling. In this specific case, we propose an improved δ-PAC stopping rule, illustrating its performance through numerical experiments.

Our contributions follow from two main mathematical results: Lemma 8 provides a general relation between the expected number of draws and Kullback-Leibler divergences of the arms' distributions, which is the key element to derive the lower bounds. Lemma 9 is a tight devation inequality for martingales with sub-Gaussian increments, in the spirit of the Law of Iterated Logarithm.

The paper is structured as follows. Section 2 presents a distribution-dependent lower bound on both $\kappa_B(\nu)$ and $\kappa_C(\nu)$ under the some identifiability assumption, as well as lower bounds for algorithms using uniform sampling. Gaussian bandit models are then studied in details in Section 3, and Bernoulli bandit models in Section 4. Section 5 includes a practical illustration of the performance of matching algorithms for Gaussian bandits, as well as a practical comparison of the fixed-confidence and fixed-budget settings. The most important elements of proof are gathered in Section 6, with the rest of the proofs in the Appendix.

2. Lower Bounding the Complexity

Introducing the Kullback-Leibler divergence of any two probability distributions p and q:

$$\mathrm{KL}(p,q) = \begin{cases} \int \log\left[\frac{dp}{dq}(x)\right] dp(x) \text{ if } q \ll p, \\ +\infty \text{ otherwise,} \end{cases}$$

we make the assumption that there exists a set $\mathcal{N}$ such that for all $\nu = (\nu_1, \nu_2) \in \mathcal{M}$, for $a \in \{1,2\}, \nu_a \in \mathcal{N}$ and that $\mathcal{N}$ satisfies

$$\forall p, q \in \mathcal{N},\ p \neq q \ \Rightarrow\ 0 < \mathrm{KL}(p,q) < +\infty.$$

A class $\mathcal{M}$ of bandit models satisfying this property is called identifiable. For $\mathcal{M}$ an identifiable class of bandit models, Theorem 1 provides lower bounds on $\kappa_B(\nu)$ and $\kappa_C(\nu)$ for every $\nu \in \mathcal{M}$. The proof of this theorem is based on changes of distribution and detailed in Section 6.

Theorem 1 *Let $\nu = (\nu_1, \nu_2)$ be a two-armed bandit model such that $\mu_1 > \mu_2$. In the fixed-budget setting, any consistent algorithm satisfies*

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq c^*(\nu), \quad \textit{where} \quad c^*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \max\left\{\mathrm{KL}(\nu_1',\nu_1), \mathrm{KL}(\nu_2',\nu_2)\right\}.$$

In the fixed-confidence setting any algorithm that is δ-PAC on $\mathcal{M}$ satisfies, when $\delta \leq 0.15$,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{c_*(\nu)}\log\left(\frac{1}{2\delta}\right), \quad \textit{where} \quad c_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \max\left\{\mathrm{KL}(\nu_1,\nu_1'), \mathrm{KL}(\nu_2,\nu_2')\right\}.$$

In particular, Theorem 1 implies that $\kappa_B(\nu) \geq 1/c^*(\nu)$ and $\kappa_C(\nu) \geq 1/c_*(\nu)$. Proceeding similarly, one can obtain lower bounds for the algorithms that use uniform sampling of both arms. The proof of the following result is easily adapted from that of Theorem 1 (cf. Section 6), using that each arm is drawn $\tau/2$ times.

Theorem 2 *Let $\nu = (\nu_1, \nu_2)$ be a two-armed bandit model such that $\mu_1 > \mu_2$. In the fixed-budget setting, any consistent algorithm using a uniform sampling strategy satisfies*

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq I^*(\nu) \quad \textit{where} \quad I^*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \frac{\mathrm{KL}\left(\nu_1',\nu_1\right) + \mathrm{KL}\left(\nu_2',\nu_2\right)}{2}.$$

In the fixed-confidence setting, any algorithm that is δ-PAC on $\mathcal{M}$ and uses a uniform sampling strategy satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{I_*(\nu)}\log\frac{1}{2\delta} \quad \textit{where} \quad I_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \frac{\mathrm{KL}\left(\nu_1,\nu_1'\right) + \mathrm{KL}\left(\nu_2,\nu_2'\right)}{2}.$$

Obviously, one always has $I^*(\nu) \leq c^*(\nu)$ and $I_*(\nu) \leq c_*(\nu)$ suggesting that uniform sampling can be sub-optimal. It is possible to give explicit expressions for the quantities $c^*(\nu), c_*(\nu)$ and $I^*(\nu), I_*(\nu)$ for specific classes of parametric bandit models that will be considered in the rest of the paper. In the case of Gaussian bandits with known variance (see Section 3):

$$\mathcal{M} = \{\nu = \left(\mathcal{N}\left(\mu_1,\sigma_1^2\right), \mathcal{N}\left(\mu_2,\sigma_2^2\right)\right) : (\mu_1,\mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\}, \tag{1}$$

one obtains

$$c^*(\nu) = c_*(\nu) = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2} \quad \text{and} \quad I^*(\nu) = I_*(\nu) = \frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}.$$

Hence, the lower bounds of Theorem 1 are equal in this case, and we provide in Section 3 matching upper bounds confirming that indeed $\kappa_B(\nu) = \kappa_C(\nu)$. In addition, the observation that $2I^*(\nu) \geq c^*(\nu) \geq I^*(\nu)$ implies that, except when $\sigma_1 = \sigma_2$, strategies based on uniform sampling are sub-optimal.

The values of $c^*(\nu)$ and $c_*(\nu)$ can also be computed for *canonical one-parameter exponential families* with density with respect to some reference measure given by

$$f_\theta(x) = A(x)\exp(\theta x - b(\theta)), \quad \text{for } \theta \in \Theta \subset \mathbb{R}. \tag{2}$$

We consider the class of bandit models

$$\mathcal{M} = \{\nu = (\nu_{\theta_1}, \nu_{\theta_2}) : (\theta_1, \theta_2) \in \Theta^2, \theta_1 \neq \theta_2\}$$

where ν_{θ_a} has density f_{θ_a} given by (2). Using the shorthand $\mathrm{K}(\theta, \theta') = \mathrm{KL}(\nu_\theta, \nu_{\theta'})$ for $(\theta, \theta') \in \Theta^2$, one can show that, for ν such that $\mu_1 > \mu_2$:

$$\begin{array}{rcccl} c^*(\nu) & = & \inf_{\theta\in\Theta} \max\left(\mathrm{K}(\theta,\theta_1), \mathrm{K}(\theta,\theta_2)\right) & = & \mathrm{K}(\theta^*,\theta_1), \text{ where } \mathrm{K}(\theta^*,\theta_1) = \mathrm{K}(\theta^*,\theta_2), \\ c_*(\nu) & = & \inf_{\theta\in\Theta} \max\left(\mathrm{K}(\theta_1,\theta), \mathrm{K}(\theta_2,\theta)\right) & = & \mathrm{K}(\theta_1,\theta_*), \text{ where } \mathrm{K}(\theta_1,\theta_*) = \mathrm{K}(\theta_2,\theta_*). \end{array}$$

The coefficient $c^*(\nu)$ is known as the *Chernoff information* $\mathrm{K}^*(\theta_1, \theta_2)$ between the distributions ν_{θ_1} and ν_{θ_2} (see Cover and Thomas (2006) and Kaufmann and Kalyanakrishnan (2013) for earlier notice of the relevance of this quantity in the best arm selection problem). By analogy, we will also denote $c_*(\nu)$ by $\mathrm{K}_*(\theta_1, \theta_2) = \mathrm{K}(\theta_1, \theta_*)$.

For exponential family bandits the quantities $c^*(\nu)$ and $c_*(\nu)$ are not equal in general, although it can be shown that it is the case when the log-partition function $b(\theta)$ is (Fenchel) self-conjugate (e.g., for Gaussian and exponential variables). In Section 4, we will focus on the case of Bernoulli models for which $c^*(\nu) > c_*(\nu)$. By exhibiting a matching strategy in the fixed-budget case, we will show that this implies that $\kappa_C(\nu) > \kappa_B(\nu)$ in this case.

3. The Gaussian Case

We study in this Section the class of two-armed Gaussian bandit models with known variances defined by (1), where σ_1 and σ_2 are fixed. In this case, we observed above that the lower bounds of Theorem 1 are similar, because $c^*(\nu) = c_*(\nu)$. We prove in this section that indeed

$$\kappa_C(\nu) = \kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

by exhibiting strategies that reach these performance bounds. These strategies are based on the simple recommendation of the empirical best arm but use non-uniform sampling in cases where σ_1 and σ_2 differ. When $\sigma_1 = \sigma_2$ we provide in Theorem 3 an improved stopping rule that is δ-PAC but results in a significant reduction of the running time of fixed-confidence tests.

3.1. Fixed-Budget Setting

We consider the simple family of *static strategies* that draw n_1 samples from arm 1 followed by $n_2 = t - n_1$ samples of arm 2, and then choose arm 1 if $\hat{\mu}_{1,n_1} < \hat{\mu}_{2,n_2}$, where $\hat{\mu}_{i,n_i}$ denotes the empirical mean of the n_i samples from arm i. Assume for instance that $\mu_1 > \mu_2$. Since $\hat{\mu}_{1,n_1} - \hat{\mu}_{2,n_2} - \mu_1 + \mu_2 \sim \mathcal{N}\left(0, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$, the probability of error of such a strategy is easily upper bounded as:

$$\mathbb{P}\left(\hat{\mu}_{1,n_1} < \hat{\mu}_{2,n_2}\right) \quad \leq \quad \exp\left(-\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{-1} \frac{(\mu_1 - \mu_2)^2}{2}\right).$$

The right hand side is minimized when $n_1/(n_1+n_2) = \sigma_1/(\sigma_1+\sigma_2)$, and the static strategy drawing $n_1 = \lceil \sigma_1 t/(\sigma_1+\sigma_2) \rceil$ times arm 1 is such that

$$\liminf_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \geq \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}\ ,$$

which matches the bound of Theorem 1 for Gaussian bandit models.

3.2. Fixed-Confidence Setting

3.2.1. EQUAL VARIANCES

We start with the simpler case $\sigma_1 = \sigma_2 = \sigma$, where the quantity $I_*(\nu)$ introduced in Theorem 2 coincides with $c_*(\nu)$, which suggests that uniform sampling could be optimal. A uniform sampling strategy is equivalent to collecting paired samples (X_s, Y_s) from both arms. The difference $X_s - Y_s$ is Gaussian with mean $\mu = \mu_1 - \mu_2$ and a δ-PAC algorithm is equivalent to a sequential test of $H_0 : \mu < 0$ versus $H_1 : \mu > 0$ such that the probability of error is uniformly bounded by δ. Robbins (1970) proposes such a test that stops after a number of samples

$$\tau = \inf\left\{ t \in 2\mathbb{N}^* : \left|\sum_{s=1}^{t/2}(X_s - Y_s)\right| > \sqrt{2\sigma^2 t\beta(t,\delta)} \right\} \text{ with } \beta(t,\delta) = \frac{t+1}{t}\log\left(\frac{t+1}{2\delta}\right) \quad (3)$$

and recommends the empirical best arm. This procedure belongs to the class of *elimination strategies*, introduced by Jennison et al. (1982) who derive a lower bound on the sample complexity of any δ-PAC *elimination* strategy—whereas our lower bound applies to *any* δ-PAC algorithm—matched by Robbins' algorithm, that is, $\lim_{\delta\to 0}\mathbb{E}_\nu[\tau]/\log\frac{1}{\delta} = 8\sigma^2/(\mu_1-\mu_2)^2$. Therefore, Robbins' rule (3) yields an optimal strategy matching our general lower bound of Theorem 1 in the particular case of Gaussian distributions with common known variance.

Note that any elimination strategy that is δ-PAC and uses a threshold function smaller than Robbins' also matches our asymptotic lower bound, while being strictly more efficient than Robbins' rule. For practical purpose, it is therefore interesting to exhibit smaller *exploration rates* $\beta(t,\delta)$ leading to a δ-PAC algorithm. The probability of error of such an algorithm is upper bounded, for example for $\mu_1 < \mu_2$ by

$$\mathbb{P}_\nu\left(\exists k \in \mathbb{N} : \sum_{s=1}^{k}\frac{X_s - Y_s - (\mu_1-\mu_2)}{\sqrt{2\sigma^2}} > \sqrt{2k\beta(2k,\delta)}\right) = \mathbb{P}\left(\exists k\in\mathbb{N} : S_k > \sqrt{2k\beta(2k,\delta)}\right) \quad (4)$$

where S_k is a sum of k i.i.d. variables of distribution $\mathcal{N}(0,1)$. Robbins (1970) obtains a non-explicit confidence region of risk at most δ by choosing $\beta(2k,\delta) = \log(\log(k)/\delta) + o(\log\log(k))$. The dependency in k is in some sense optimal, because the Law of Iterated Logarithm (LIL) states that $\limsup_{k\to\infty} S_k/\sqrt{2k\log\log(k)} = 1$ almost surely. Recently, Jamieson et al. (2013) proposed an explicit confidence region inspired by the LIL. However, Lemma 1 of (Jamieson et al., 2013) cannot be used to upper bound (4) by δ and we provide in Section 6 a result derived independently (Lemma 9) that achieves this goal and yields the following result.

Theorem 3 *For δ small enough, the elimination strategy (3) is δ-PAC with*

$$\beta(t,\delta) = \log\frac{1}{\delta} + \frac{3}{4}\log\log\frac{1}{\delta} + \frac{3}{2}\log(1+\log(t/2)). \quad (5)$$

Algorithm 1 α-Elimination

Require: Exploration function $\beta(t,\delta)$, parameter α.

Initialization: $\hat{\mu}_1(0) = \hat{\mu}_2(0) = 0$, $\sigma_0^2(\alpha) = 1$. $t = 0$.
while $|\hat{\mu}_1(t) - \hat{\mu}_2(t)| \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}$ **do**
  $t = t + 1$.
  If $\lceil \alpha t \rceil = \lceil \alpha(t-1) \rceil$, $A_t = 2$, else $A_t = 1$.
  Observe $Z_t \sim \nu_{A_t}$ and compute the empirical means $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$.
  Compute $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$.
end while
return $a = \underset{a=1,2}{\operatorname{argmax}}\, \hat{\mu}_a(t)$

3.2.2. MISMATCHED VARIANCES

In the case where $\sigma_1 \neq \sigma_2$, we rely on an α-elimination strategy, described in Algorithm 1. For $a = 1, 2$, $\hat{\mu}_a(t)$ denotes the empirical mean of the samples gathered from arm a up to time t. The algorithm is based on a non-uniform sampling strategy governed by the parameter $\alpha \in (0,1)$ which ensures that, at the end of every round t, $N_1(t) = \lceil \alpha t \rceil$, $N_2(t) = t - \lceil \alpha t \rceil$ and $\hat{\mu}_1(t) - \hat{\mu}_2(t) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma_t^2(\alpha)\right)$. The sampling schedule used here is thus deterministic.

Theorem 4 shows that the $\sigma_1/(\sigma_1+\sigma_2)$-elimination algorithm, with a suitable exploration rate, is δ-PAC and matches the lower bound on $\mathbb{E}_\nu[\tau]$, at least asymptotically when $\delta \to 0$. Its proof can be found in Appendix C.

Theorem 4 *If $\alpha = \sigma_1/(\sigma_1+\sigma_2)$, the α-elimination strategy using the exploration rate $\beta(t,\delta) = \log \frac{t}{\delta} + 2\log\log(6t)$ is δ-PAC on $\mathcal{M}$ and satisfies, for every $\nu \in \mathcal{M}$, for every $\epsilon > 0$,*

$$\mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\left(\frac{1}{\delta}\right)\right).$$

Remark 5 *When $\sigma_1 = \sigma_2$, $1/2$-elimination reduces, up to rounding effects, to the elimination procedure described in Section 3.2.1, for which Theorem 3 suggests an exploration rate of order $\log(\log(t)/\delta)$. As the feasibility of this exploration rate when $\sigma_1 \neq \sigma_2$ is yet to be established, we focus on Gaussian bandits with equal variances in the numerical experiments of Section 5.*

4. The Bernoulli Case

We consider in this section the class of Bernoulli bandit models defined by

$$\mathcal{M} = \{\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1,\mu_2) \in]0;1[^2, \mu_1 \neq \mu_2\},$$

where each arm can be equivalently parametrized by the natural parameter of the exponential family, $\theta_a = \log(\mu_a/(1-\mu_a))$. Following the notation of Section 2, the Kullback-Leibler divergence between two Bernoulli distributions can be either expressed as a function of the means, $\mathrm{KL}(\mathcal{B}(\mu_1), \mathcal{B}(\mu_2))$, or of the natural parameters, $\mathrm{K}(\theta_1, \theta_2)$.

In this Section, we prove that $\kappa_C(\nu) > \kappa_B(\nu)$ for Bernoulli bandit models (Proposition 7). To do so, we first introduce a static strategy matching the lower bound of Theorem 1 in the fixed-budget case (Proposition 6). This strategy is reminiscent of the algorithm exhibited for Gaussian

bandits in Section 3 and uses parameter-dependent non uniform sampling. This strategy is not directly helpful in practice but we show that it can be closely approximated by an algorithm using the uniform sampling strategy. In the fixed-confidence setting we similarly conjecture that little can be gained from using a non-uniform sampling strategy and propose an algorithm based on a non-trivial stopping strategy that is believed to match the bound of Theorem 2.

Proposition 6 *Let $\alpha(\theta_1, \theta_2)$ be defined by*

$$\alpha(\theta_1, \theta_2) = \frac{\theta^* - \theta_1}{\theta_2 - \theta_1} \quad \textit{where} \quad \mathrm{K}(\theta^*, \theta_1) = \mathrm{K}(\theta^*, \theta_2).$$

For all t, the static strategy that allocates $\lceil \alpha(\theta_1, \theta_2) t \rceil$ samples to arm 1 , and recommends the empirical best arm, satisfies $p_t(\nu) \leq \exp(-K^*(\theta_1, \theta_2)t)$.

This result, proved in Appendix D, shows in particular that for every $\nu \in \mathcal{M}$ there exists a consistent static strategy such that

$$\liminf_{t\to\infty} -\frac{1}{t} \log p_t \geq \mathrm{K}^*(\theta_1, \theta_2), \quad \text{and hence that } \kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\theta_1, \theta_2)}.$$

However, as $\alpha(\theta_1, \theta_2)$ depends in the Bernoulli case on the unknown means of the arms, this optimal static strategy is not useful in practice. So far, it is not known whether there exists a *universal* strategy such that $p_t(\nu) \leq \exp(-\mathrm{K}^*(\theta_1, \theta_2)t)$ for all bandit instance $\nu \in \mathcal{M}$.

For Bernoulli bandit models it can be checked that for all $\nu \in \mathcal{M}$, $\mathrm{K}_*(\theta_1, \theta_2) < \mathrm{K}^*(\theta_1, \theta_2)$. This fact together with Proposition 6 and Theorem 1 yields the following inequality.

Proposition 7 *For all $\nu \in \mathcal{M}$, $\kappa_C(\nu) > \kappa_B(\nu)$.*

In the specific case of Bernoulli distributions, there is a strong incentive to use uniform sampling: the quantities $I^*(\nu)$ and $I_*(\nu)$ introduced in Theorem 2 appear to be very close to $c^*(\nu)$ and $c_*(\nu)$ respectively. This fact is illustrated in Figure 1, on which we represent these different quantities, that are functions of the means μ_1, μ_2 of the arms, as a function of μ_1, for two fixed values of μ_2. Therefore, algorithms matching the bounds of Theorem 2 provide upper bounds on $\kappa_B(\nu)$ (resp. $\kappa_C(\nu)$) very close to $1/c^*(\nu)$ (resp. $1/c_*(\nu)$). In the fixed-budget setting, Lemma 15 shows that the strategy with uniform sampling that recommends the empirical best arm, satisfies $p_t(\nu) \leq e^{-tI^*(\nu)}$, and matches the bound of Theorem 2 (see Remark 16 in Appendix D). Hence, problem-dependent optimal strategy described above can be approximated by a very simple, universal algorithm. Similarly, finding an algorithm for the fixed-confidence setting sampling the arms uniformly and matching the bound of Theorem 2 is a crucial matter. This boils down to finding a good stopping rule. In all the algorithms studied so far, the stopping rule was based on the difference of the empirical means of the arms. For Bernoulli arms, such a strategy can be analyzed with the tools provided in this paper: the algorithm stopping for t such that $\hat{\mu}_{1,t/2} - \hat{\mu}_{2,t/2} > \sqrt{2\beta(t,\delta)/t}$ with $\beta(t,\delta)$ as in Theorem 3 is δ-PAC and its expected running time bounded by $2/(\mu_1 - \mu_2)^2 \log \frac{1}{\delta} + o\left(\log \frac{1}{\delta}\right)$. Yet, Pinsker's inequality implies that $I_*(\mu_1, \mu_2) > (\mu_1 - \mu_2)^2/2$ and this algorithm is thus not optimal with respect to Theorem 2. The approximation $I_*(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2/(8\mu_1(1-\mu_1)) + o\left((\mu_1 - \mu_2)^2\right)$ suggests that the loss with respect to the

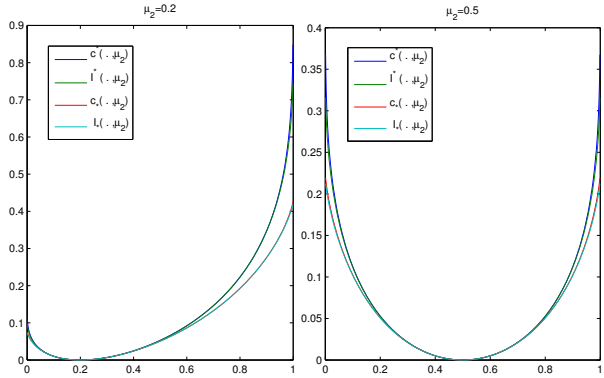

Figure 1: Comparison of different information terms in Bernoulli bandit models.

optimal error exponent is particularly significant when both means are close to 0 or 1. The stopping rule we propose to circumvent this drawback is the following:

$$\tau = \inf\left\{t \in 2\mathbb{N}^* : tI_*(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) > \log\left((\log(t)+1)/\delta\right)\right\}. \tag{6}$$

This algorithm is related to the KL-LUCB algorithm of Kaufmann and Kalyanakrishnan (2013). Indeed, $I_*(x,y)$ mostly coincides with $\mathrm{K}_*(\mathcal{B}(x), \mathcal{B}(y))$ (Figure 1) and a closer examination shows that the stopping criterion in KL-LUCB for two arms is exactly of the form $tK_*(\mathcal{B}(\hat{\mu}_{1,t/2}), \mathcal{B}(\hat{\mu}_{2,t/2})) > \beta(t,\delta)$. The results of Kaufmann and Kalyanakrishnan (2013) show in particular that the algorithm based on (6) is provably δ-PAC for appropriate choices of $\beta(t,\delta)$. However, by analogy with the result of Theorem 3 we believe that the analysis of Kaufmann and Kalyanakrishnan (2013) is too conservative and that the proposed approach should be δ-PAC for exploration rates $\beta(t,\delta)$ that grow as a function of t only at rate $\log\log t$.

5. Numerical Experiments and Discussion

The goal of this Section is twofold: to compare results obtained in the fixed-budget and fixed-confidence settings and to illustrate the improvement resulting from the adoption of the reduced exploration rate of Theorem 3.

In Figure 2, we consider two bandit models: the 'easy' one is $\mathcal{N}(0.5, 0.25) \otimes \mathcal{N}(0, 0.25)$, $\kappa = 8$ (left) and the 'difficult' one is $\mathcal{N}(0.01, 0.25) \otimes \mathcal{N}(0, 0.25)$, $\kappa = 20000$ (right). In the fixed-budget setting, stars ('*') report the probability of error $p_n(\nu)$ as a function of n. In the fixed-confidence setting, we plot both the empirical probability of error by circles ('O') and the specified maximal error probability δ by crosses ('X') as a function of the empirical average of the running times. Note the logarithmic scale used for the probabilities on the y-axis. All results are averaged on $N = 10^6$ independent Monte Carlo replications. For comparison purposes, a plain line represents the theoretical rate $x \mapsto \exp(-x/\kappa)$ which is a straight line on the log scale.

In the fixed-confidence setting, we report results for algorithms of the form (3) with $g(t,\delta) = \sqrt{2\sigma^2 t\beta(t,\delta)}$ for three different exploration rates $\beta(t,\delta)$. The exploration rate we consider are: the provably-PAC rate of Robbins' algorithm $\log(t/\delta)$ (large blue symbols), the conjectured 'optimal' exploration rate $\log((\log(t)+1)/\delta)$, almost provably δ-PAC according to Theorem 3 (bold green symbols), and the rate $\log(1/\delta)$, which would be appropriate if we were to perform the stopping test

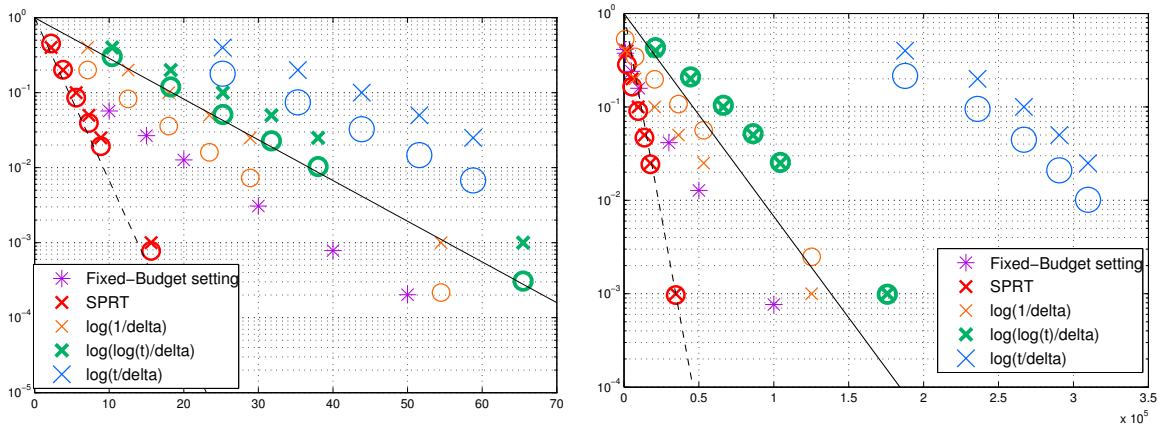

Figure 2: Experimental results (descriptions in text; plots best viewed in color).

only at a single pre-specified time (orange symbols). For each algorithm, the log probability of error is approximately a linear function of the number of samples, with a slope close to $-1/\kappa$, where κ is the complexity. We can visualize the gain in sample complexity achieved by smaller exploration rates, but while the rate $\log((\log(t)+1)/\delta)$ appears to guarantee the desired probability of error across all problems, the use of $\log(1/\delta)$ seems too risky, as one can see that the probability of error becomes larger than δ on difficult problems. To illustrate the gain in sample complexity when the means of the arms are known, we add in red the SPRT algorithm mentioned in the introduction along with the theoretical relation between the probability of error and the expected number of samples, materialized as a dashed line. The SPRT stops for t such that $|(\mu_1-\mu_2)(S_{1,t/2}-S_{2,t/2})| > \log(1/\delta)$.

Robbins' algorithm is δ-PAC and matches the complexity (which is illustrated by the slope of the measures), though in practice the use of the exploration rate $\log((\log(t)+1)/\delta)$ leads to huge gain in terms of number of samples used. It is important to keep in mind that running times play the same role as error exponents and hence the threefold increase of average running times observed on the rightmost plot of Figure 2 when using $\beta(t,\delta) = \log(t/\delta)$ is really prohibitive. This illustrates the asymptotic nature of our notion of complexity: the leading term in $\mathbb{E}_\nu[\tau]$ is indeed $\kappa \log(1/\delta)$ but there is a second-order constant term which is not negligible for fixed value of δ. Jamieson et al. (2013) implicitly consider an alternative complexity for Gaussian bandit models with common known variance: for a fixed value of δ, if $\Delta = \mu_1 - \mu_2$, they show that when the gap Δ goes to zeros, the sample complexity if of order some constant—that depends on δ—multiplied by $\Delta^{-2} \log\log \Delta^{-2}$.

If one compares on each problem the results for the fixed-budget setting to those for the best δ-PAC algorithm (in green), one can see that to obtain the same probability of error, the fixed-confidence algorithm needs an average number of samples of order at least twice larger than the deterministic number of samples required by the fixed-budget setting algorithm. This remark should be related to the fact that a δ-PAC algorithm is designed to be uniformly good across all problems, whereas consistency is a weak requirement in the fixed-budget setting: any strategy that draws both arm infinitely often and recommends the empirical best is consistent. Figure 2 shows that when the values of μ_1 and μ_2 are unknown, the sequential version of the test is no more preferable to its batch counterpart and can even become much worse if the exploration rate $\beta(t,\delta)$ is chosen too conservatively. This observation should be mitigated by the fact that the sequential (or fixed-confidence)

approach is adaptive with respect to the difficulty of the problem whereas it is impossible to predict the efficiency of a batch (or fixed-budget) experiment without some prior knowledge regarding the problem under consideration.

6. Elements of Proof

6.1. Proof of Theorem 1

The cornerstone of the proof of all the lower bounds given in this paper is Lemma 8 which relates the probabilities of the same event under two different models to the expected number of draws of each arm. Its proof, which may be found in Appendix A, encapsulates the technical aspects of the change of distributions. $N_a(t)$ denotes the number of draws of arm a up to round t and $N_a = N_a(\tau)$ is the total number of draws of arm a by some algorithm $\mathcal{A}$.

Lemma 8 *Let ν and ν' be two bandit models. For any $A \in \mathcal{F}_\tau$ such that $0 < \mathbb{P}_\nu(A) < 1$*

$$\mathbb{E}_\nu[N_1]\mathrm{KL}(\nu_1, \nu_1') + \mathbb{E}_\nu[N_2]\mathrm{KL}(\nu_2, \nu_2') \geq d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)), \tag{7}$$

where $d(x,y) := \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)) = x\log(x/y) + (1-x)\log\big((1-x)/(1-y)\big)$.

Without loss of generality, assume that the bandit model $\nu = (\nu_1, \nu_2)$ is such that $a^* = 1$. Consider any alternative bandit model $\nu' = (\nu_1', \nu_2')$ in which $a^* = 2$ and the event $A = (\hat{a} = 1)$ where $\hat{a}$ is the arm chosen by algorithm $\mathcal{A}$. Clearly $A \in \mathcal{F}_\tau$.

Fixed-Confidence Setting. Let $\mathcal{A}$ be a δ-PAC algorithm. Then it is correct on both ν and ν' and satisfies $\mathbb{P}_\nu(A) \geq 1-\delta$ and $\mathbb{P}_{\nu'}(A) \leq \delta$. Using monotonicity properties of d (for example $x \mapsto d(x,y)$ is increasing when $x > y$ and decreasing when $x < y$) and inequality (7) in Lemma 8 yields $\mathbb{E}_\nu[N_1]\mathrm{KL}(\nu_1, \nu_1') + \mathbb{E}_\nu[N_2]\mathrm{KL}(\nu_2, \nu_2') \geq d(\delta, 1-\delta)$, and hence

$$\mathbb{E}_\nu[\tau] \geq \frac{d(\delta, 1-\delta)}{\max_{a=1,2} \mathrm{KL}(\nu_a, \nu_a')},$$

using that $\tau = N_1 + N_2$. Optimizing over the possible model ν' satisfying $\mu_1' < \mu_2'$ to make the right hand side of the inequality as large as possible gives the result, using moreover that for $\delta \leq 0.15$, it can be shown that $d(1-\delta, \delta) \geq \log(1/(2\delta))$.

Fixed-Budget Setting. Inequality (7) in Lemma 8 applied to A yields

$$\mathbb{E}_{\nu'}[N_1(t)]\mathrm{KL}(\nu_1', \nu_1) + \mathbb{E}_{\nu'}[N_2(t)]\mathrm{KL}(\nu_2', \nu_2) \geq d(\mathbb{P}_{\nu'}(A), \mathbb{P}_\nu(A)).$$

Note that $p_t(\nu) = 1 - \mathbb{P}_\nu(A)$ and $p_t(\nu') = \mathbb{P}_{\nu'}(A)$. As algorithm $\mathcal{A}$ is correct on both ν and ν', for every $\epsilon > 0$ there exists $t_0(\epsilon)$ such that for all $t \geq t_0(\epsilon)$, $\mathbb{P}_{\nu'}(A) \leq \epsilon \leq \mathbb{P}_\nu(A)$. For $t \geq t_0(\epsilon)$,

$$\mathbb{E}_{\nu'}[N_1(t)]\mathrm{KL}(\nu_1', \nu_1) + \mathbb{E}_{\nu'}[N_2(t)]\mathrm{KL}(\nu_2', \nu_2) \geq d(\epsilon, 1-p_t(\nu)) \geq (1-\epsilon)\log\frac{1-\epsilon}{p_t(\nu)} + \epsilon\log\epsilon.$$

Taking the limsup (denoted by $\overline{\lim}$) and letting ϵ go to zero, one can show that

$$\overline{\lim} - \frac{1}{t}\log p_t(\nu) \quad \leq \quad \overline{\lim}\Big[\frac{\mathbb{E}_{\nu'}[N_1(t)]}{t}\mathrm{KL}(\nu_1', \nu_1) + \frac{\mathbb{E}_{\nu'}[N_2(t)]}{t}\mathrm{KL}(\nu_2', \nu_2)\Big] \leq \max_{a=1,2}\mathrm{KL}(\nu_a', \nu_a).$$

Optimizing over the possible model ν' satisfying $\mu_1' < \mu_2'$ to make the right hand side of the inequality as small as possible gives the result.

6.2. Proof of Theorem 3

According to (4), the proof of Theorem 3 boils down to finding an exploration rate such that $\mathbb{P}(\exists t \in \mathbb{N}^* : S_t > \sqrt{2\sigma^2 t\beta(t,\delta)}) \leq \delta$, where $S_t = X_1 + \cdots + X_t$ is a sum of i.i.d. normal random variable. Lemma 9 provides such a confidence region. Its proof can be found in Appendix B.

Lemma 9 *Let $\zeta(u) = \sum_{k\geq 1} k^{-u}$. Let $X_1, X_2, \ldots$ be independent random variables such that, for all $\lambda \in \mathbb{R}$, $\phi(\lambda) := \log \mathbb{E}[\exp(\lambda X_1)] \leq \lambda^2\sigma^2/2$. For every positive integer t let $S_t = X_1 + \cdots + X_t$. Then, for all $\beta > 1$ and $x \geq \frac{8}{(e-1)^2}$,*

$$\mathbb{P}\Big(\exists t \in \mathbb{N}^* : S_t > \sqrt{2\sigma^2 t(x + \beta \log\log(et))}\Big) \leq \sqrt{e}\,\zeta\Big(\beta\big(1 - \frac{1}{2x}\big)\Big)\Big(\frac{\sqrt{x}}{2\sqrt{2}} + 1\Big)^{\beta} \exp(-x).$$

Let $\beta(t,\delta)$ be of the form $\beta(t,\delta) = \log\frac{1}{\delta} + c\log\log\frac{1}{\delta} + d\log\log(et)$, for some constants $c > 0$ and $d > 1$. Lemma 9 yields

$$\mathbb{P}\Big(\exists t \in \mathbb{N} : S_t > \sqrt{2\sigma^2 t\beta(t,\delta)}\Big) \leq \zeta\Big(d\big(1 - \frac{1}{2(z + c\log z)}\big)\Big)\frac{\sqrt{e}}{(2\sqrt{2})^d}\frac{(\sqrt{z + c\log z} + \sqrt{8})^d}{z^c}\delta,$$

where $z := \log\frac{1}{\delta} > 0$. To upper bound the above probability by δ, at least for large values of z (which corresponds to small values of δ), it suffices to choose the parameters c and d such that

$$\sqrt{e}\,\zeta\Big(d\big(1 - \frac{1}{2(z + c\log z)}\big)\Big)\frac{1}{(2\sqrt{2})^d}\frac{(\sqrt{z + c\log z} + 2\sqrt{2})^d}{z^c} \leq 1.$$

For $c = d/2$, the left hand side tends to $\sqrt{e}\zeta\,(d)/(2\sqrt{2})^d$ when z goes to infinity, which is smaller than 1 for $d \geq 1.47$. Thus, for δ small enough, the desired inequality holds for $d = 3/2$ and $c = 3/4$, which corresponds to the exploration rate of Theorem 3.

7. Conclusion

We provide distribution-dependent lower bounds for best-arm identification in the context of two-armed bandit models. These bounds involve information-theoretic quantities that reflect the typical causes of failure, which are different from those appearing in regret analysis. For Gaussian and Bernoulli bandit models, we exhibit matching algorithms showing that these bounds are (mostly) tight, highlighting relationships between the complexities of the fixed-budget and fixed-confidence settings. Our numerical experiments illustrate the significance of using appropriate exploration rates in the context of best arm(s) identification and we believe that Lemma 8 can be adapted to deal with more general K-armed bandit scenarios.

These results suggest three practical implications for A/B testing. First, for Binary and Gaussian-like responses with matched variances it is reasonable to consider only tests—i.e., strategies using uniform sampling—rather than general sequential sampling strategies. Second, using a sequential stopping rule in this context is mostly of interest because it does not requires prior knowledge of the complexity of the problem. It should however not be expected to reduce the (average) running time of the experiment for a given probability of error. This leads to the third message regarding the utmost importance of using proper (i.e., provably δ-PAC but not too conservative) exploration rates when using a sequential stopping rule.

Acknowledgments

We thank Sébastien Bubeck for fruitful discussions during the visit of the first author at Princeton University. This work was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

References

S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

J-Y. Audibert, S. Bubeck, and R. Munos. Best Arm Identification in Multi-armed Bandits. In *Conference on Learning Theory (COLT)*, 2010.

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

Robert Bechhofer, Jack Kiefer, and Milton Sobel. *Sequential identification and ranking procedures*. The university of Chicago Press, 1968.

S. Bubeck, R. Munos, and G. Stoltz. Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science*, 412:1832–1852, 2011.

S. Bubeck, T. Wang, and N. Viswanathan. Multiple Identifications in multi-armed bandits. In *International Conference on Machine Learning (ICML)*, 2013.

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.

T. Cover and J. Thomas. *Elements of Information Theory (2nd Edition)*. Wiley, 2006.

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 *Neural Information and Signal Processing (NIPS)*, 2012.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an optimal exploration algorithm for multi-armed bandits. *arXiv:1312.7308*, 2013.

C. Jennison, I.M. Johnstone, and B.W. Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. *Statistical Decision Theory and Related Topics III*, 2:55–86, 1982.

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 *Conference On Learning Theory (COLT)*, 2013.

E. Kaufmann, A. Garivier, and O. Cappé. On Bayesian Upper-Confidence Bounds for Bandit Problems. In *AISTATS*, 2012a.

E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Algorithmic Learning Theory (ALT)*, 2012b.

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

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.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

H. Robbins. Statistical Methods Related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

D. Siegmund. *Sequential Analysis*. Springer-Verlag, 1985.

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

A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Appendix A. Proof of Lemma 8: Changes of Distributions

Under the identifiability assumption, there exists a common measure λ such that for all $\nu = (\nu_1, \nu_2)$, for all $a \in \{1, 2\}$ ν_a has a density f_a with respect to λ.

Let $\nu \in \mathcal{M}$ be a bandit model, and consider an alternative bandit model $\nu' \in \mathcal{M}$. f_a, f'_a are the densities of ν_a, ν'_a respectively and one can introduce the log-likelihood ratio of the observations up to time t under an algorithm $\mathcal{A}$:

$$L_t := \sum_{s=1}^{t} \mathbb{1}_{(A_s=1)} \log\left(\frac{f_1(Z_s)}{f'_1(Z_s)}\right) + \sum_{s=1}^{t} \mathbb{1}_{(A_s=2)} \log\left(\frac{f_2(Z_s)}{f'_2(Z_s)}\right).$$

The key element in a change of distribution is the following classical lemma (whose proof is omitted) that relates the probabilities of an event under $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$ through the log-likelihood ratio of the observations. Such a result has often been used in the bandit literature for ν and ν' that differ just from one arm (either $\nu_1 = \nu'_1$ or $\nu_2 = \nu'_2$), for which the expression of the log-likelihood ratio is simpler. As we will see, here we consider more general changes of distributions.

Lemma 10 *Let σ be any stopping time with respect to $\mathcal{F}_t$. For every event $A \in \mathcal{F}_\sigma$ (i.e. A such that $A \cap (\sigma = t) \in \mathcal{F}_t$),*

$$\mathbb{P}_{\nu'}(A) = \mathbb{E}_\nu[\mathbb{1}_A \exp(-L_\sigma)]$$

Let τ be the stopping rule of algorithm $\mathcal{A}$. We start by showing that for all $A \in \mathcal{F}_\tau$, $\mathbb{P}_\nu(A) = 0$ if and only if $\mathbb{P}_{\nu'}(A) = 0$. Thus, if $0 < \mathbb{P}_\nu(A) < 1$ one also has $0 < \mathbb{P}_{\nu'}(A) < 1$ and the quantity $d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A))$ in the statement of Lemma 8 is well defined. Let $A \in \mathcal{F}_\tau$. Lemma 10 yields $\mathbb{P}_{\nu'}(A) = \mathbb{E}_\nu[\mathbb{1}_A \exp(-L_\tau)]$. Thus $\mathbb{P}_{\nu'}(A) = 0$ implies $\mathbb{1}_A \exp(-L_\tau) = 0$ $\mathbb{P}_\nu - a.s.$ As $\mathbb{P}_\nu(\tau < +\infty) = 1$, $\mathbb{P}_\nu(\exp(L_\tau) > 0) = 1$ and $\mathbb{P}_{\nu'}(A) = 0 \Rightarrow \mathbb{P}_\nu(A) = 0$. A similar reasoning yields $\mathbb{P}_\nu(A) = 0 \Rightarrow \mathbb{P}_{\nu'}(A) = 0$.

We now prove Lemma 8. Let $A \in \mathcal{F}_\tau$ be such that $0 < \mathbb{P}_\nu(A) < 1$. Then $0 < \mathbb{P}_{\nu'}(A) < 1$. Lemma 10 and the conditional Jensen inequality leads to

$$\begin{aligned}
\mathbb{P}_{\nu'}(A) &= \mathbb{E}_\nu[\exp(-L_\tau)\mathbb{1}_A] = \mathbb{E}_\nu[\mathbb{E}_\nu[\exp(-L_\tau)|\mathbb{1}_A]\mathbb{1}_A] \\
&\geq \mathbb{E}_\nu[\exp\left(-\mathbb{E}_\nu[L_\tau|\mathbb{1}_A]\right)\mathbb{1}_A] = \mathbb{E}_\nu[\exp\left(-\mathbb{E}_\nu[L_\tau|A]\right)\mathbb{1}_A] \\
&= \exp\left(-\mathbb{E}_\nu[L_\tau|A]\right)\mathbb{P}_\nu(A).
\end{aligned}$$

Writing the same for the event $\overline{A}$ yields $\mathbb{P}_{\nu'}(\overline{A}) \geq \exp\left(-\mathbb{E}_\nu[L_\tau|\overline{A}]\right)\mathbb{P}_\nu(\overline{A})$ and finally

$$\mathbb{E}_\nu[L_\tau|A] \geq \log\frac{\mathbb{P}_\nu(A)}{\mathbb{P}_{\nu'}(A)} \quad \text{and} \quad \mathbb{E}_\nu[L_\tau|\overline{A}] \geq \log\frac{\mathbb{P}_\nu(\overline{A})}{\mathbb{P}_{\nu'}(\overline{A})}.$$

Therefore one can write

$$\begin{aligned}
\mathbb{E}_\nu[L_\tau] &= \mathbb{E}_\nu[L_\tau|A]\mathbb{P}_\nu(A) + \mathbb{E}_\nu[L_\tau|\overline{A}]\mathbb{P}_\nu(\overline{A}) \\
&\geq \mathbb{P}_\nu(A)\log\frac{\mathbb{P}_\nu(A)}{\mathbb{P}_{\nu'}(A)} + \mathbb{P}_\nu(\overline{A})\log\frac{\mathbb{P}_\nu(\overline{A})}{\mathbb{P}_{\nu'}(\overline{A})} = d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)). \qquad (8)
\end{aligned}$$

Introducing $(Y_{a,t})$, the sequence of i.i.d. samples successively observed from arm a, the log-likelihood ratio L_t can be rewritten

$$L_t = \sum_{a=1}^{2}\sum_{t=1}^{N_a(t)} \log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right); \quad \text{and} \quad \mathbb{E}_\nu\left[\log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right)\right] = \mathrm{KL}(\nu_a, \nu'_a).$$

Applying Wald's Lemma (see for example Siegmund (1985)) to $L_\tau = \sum_{a=1}^{2}\sum_{t=1}^{N_a}\log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right)$, where $N_a = N_a(\tau)$ is the total number of draws of arm a, yields

$$\mathbb{E}_\nu[L_\tau] = \mathbb{E}_\nu[N_1]\mathrm{KL}(\nu_1, \nu'_1) + \mathbb{E}_\nu[N_2]\mathrm{KL}(\nu_2, \nu'_2),$$

which concludes the proof together with inequality (8).

Appendix B. Proof of Lemma 9: An Optimal Confidence Region

We start by stating three technical lemmas, whose proofs are partly omitted.

Lemma 11 *For every $\eta > 0$, every positive integer k, and every integer t such that $(1+\eta)^{k-1} \leq t \leq (1+\eta)^k$,*

$$\sqrt{\frac{(1+\eta)^{k-1/2}}{t}} + \sqrt{\frac{t}{(1+\eta)^{k-1/2}}} \leq (1+\eta)^{1/4} + (1+\eta)^{-1/4}.$$

Lemma 12 *For every $\eta > 0$,*

$$A(\eta) := \frac{4}{\left((1+\eta)^{1/4} + (1+\eta)^{-1/4}\right)^2} \geq 1 - \frac{\eta^2}{16}.$$

Lemma 13 *Let t be such that $(1+\eta)^{k-1} \leq t \leq (1+\eta)^k$. Then, if $\lambda = \sigma^{-1}\sqrt{2zA(\eta)/(1+\eta)^{k-1/2}}$,*

$$\sigma\sqrt{2z} \geq \frac{A(\eta)z}{\lambda\sqrt{t}} + \frac{\lambda\sigma^2\sqrt{t}}{2}\ .$$

Proof:

$$\frac{A(\eta)z}{\lambda\sqrt{t}} + \frac{\lambda\sigma^2\sqrt{t}}{2} = \frac{\sigma\sqrt{2zA(\eta)}}{2}\left(\sqrt{\frac{(1+\eta)^{k-1/2}}{t}} + \sqrt{\frac{t}{(1+\eta)^{k-1/2}}}\right) \leq \sigma\sqrt{2z}$$

according to Lemma 11.

□

An important fact is that for every $\lambda \in \mathbb{R}$, because the X_i are σ-subgaussian, $W_t = \exp(\lambda S_t - t\frac{\lambda^2\sigma^2}{2}))$ is a super-martingale, and thus, for every positive u,

$$\mathbb{P}\left(\bigcup_{t\geq 1}\left\{\lambda S_t - t\frac{\lambda^2\sigma^2}{2} > u\right\}\right) \leq \exp(-u). \tag{9}$$

Let $\eta \in]0, e-1]$ to be defined later, and let $T_k = \mathbb{N} \cap \left[(1+\eta)^{k-1}, (1+\eta)^k\right[$.

$$\mathbb{P}\left(\bigcup_{t\geq 1}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right) \leq \sum_{k=1}^{\infty}\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right)$$
$$\leq \sum_{k=1}^{\infty}\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\left(k\log(1+\eta)\right)}\right\}\right)\ .$$

We use that $\eta \leq e-1$ to obtain the last inequality since this condition implies

$$\log(\log(e(1+\eta)^{k-1}) \geq \log(k\log(1+\eta)).$$

For a positive integer k, let $z_k = x + \beta\log\left(k\log(1+\eta)\right)$ and $\lambda_k = \sigma^{-1}\sqrt{2z_kA(\eta)/(1+\eta)^{k-1/2}}$. Lemma 13 shows that for every $t \in T_k$,

$$\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\} \subset \left\{\frac{S_t}{\sqrt{t}} > \frac{A(\eta)z_k}{\lambda_k\sqrt{t}} + \frac{\sigma^2\lambda_k\sqrt{t}}{2}\right\}\ .$$

Thus, by inequality (9),

$$\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\}\right) \le \mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sqrt{t}} > \frac{A(\eta)z_k}{\lambda_k\sqrt{t}} + \frac{\sigma^2\lambda_k\sqrt{t}}{2}\right\}\right)$$
$$= \mathbb{P}\left(\bigcup_{t\in T_k}\left\{\lambda_k S_t - \frac{\sigma^2\lambda_k^2 t}{2} > A(\eta)z_k\right\}\right)$$
$$\le \exp\left(-A(\eta)z_k\right) = \frac{\exp(-A(\eta)x)}{(k\log(1+\eta))^{\beta A(\eta)}} .$$

One chooses $\eta^2 = 8/x$ for x such that $x \ge \frac{8}{(e-1)^2}$ (which ensures $\eta \le e-1$). Using Lemma 12, one obtains that $\exp(-A(\eta)x) \le \sqrt{e}\exp(-x)$. Moreover,

$$\frac{1}{\log(1+\eta)} \le \frac{1+\eta}{\eta} = \frac{\sqrt{x}}{2\sqrt{2}} + 1 .$$

Thus,

$$\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\}\right) \le \frac{\sqrt{e}}{k^{\beta A(\eta)}}\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta A(\eta)}\exp(-x) \le \frac{\sqrt{e}}{k^{\beta A(\eta)}}\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta}\exp(-x)$$

and hence,

$$\mathbb{P}\left(\bigcup_{t\ge 1}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x+\beta\log\log(et)}\right\}\right) \le \sqrt{e}\zeta\left(\beta A(\eta)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta A(\eta)}\exp\left(-x\right)$$
$$\le \sqrt{e}\zeta\left(\beta\left(1-\frac{1}{2x}\right)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta}\exp\left(-x\right) ,$$

using the lower bound on $A(\eta)$ given in Lemma 12 and the fact that $A(\eta)$ is upper bounded by 1.

Appendix C. Proof of Theorem 4

Let $\alpha = \sigma_1/(\sigma_1+\sigma_2)$. We first prove that with the exploration rate $\beta(t,\delta) = \log(t/\delta)+2\log\log(6t)$ the algorithm is δ-PAC. Assume that $\mu_1 > \mu_2$ and recall $\tau = \inf\{t\in\mathbb{N} : |d_t| > \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\}$. The probability of error of the α-elimination strategy is upper bounded by

$$\begin{aligned}
\mathbb{P}_\nu\left(d_\tau \le -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right) &\le \mathbb{P}_\nu\left(d_\tau - (\mu_1-\mu_2) \le -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right)\\
&\le \mathbb{P}_\nu\left(\exists t\in\mathbb{N}^* : d_t - (\mu_1-\mu_2) < -\sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\right)\\
&\le \sum_{t=1}^{\infty}\exp\left(-\beta(t,\delta)\right),
\end{aligned}$$

by an union bound and Chernoff bound applied to $d_t - (\mu_1 - \mu_2) \sim \mathcal{N}\left(0, \sigma_t^2(\alpha)\right)$. The choice of $\beta(t,\delta)$ mentioned above ensures that the series in the right hand side is upper bounded by δ, which shows the algorithm is δ-PAC:

$$\sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \leq \delta \sum_{t=1}^{\infty} \frac{1}{t(\log(6t))^2} \leq \delta \left(\frac{1}{(\log 6)^2} + \int_1^{\infty} \frac{dt}{t(\log(6t))^2} \right) = \delta \left(\frac{1}{(\log 6)^2} + \frac{1}{\log(6)} \right) \leq \delta.$$

To upper bound the expected sample complexity, we start by upper bounding the probability that τ exceeds some deterministic time T:

$$\begin{aligned}
\mathbb{P}_\nu(\tau \geq T) &\leq \mathbb{P}_\nu\left(\forall t = 1 \dots T,\ d_t \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} \right) \leq \mathbb{P}_\nu\left(d_T \leq \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right) \\
&= \mathbb{P}_\nu\left(d_T - (\mu_1 - \mu_2) \leq -\left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right] \right) \\
&\leq \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right]^2 \right).
\end{aligned}$$

The last inequality follows from Chernoff bound and holds for T such that

$$(\mu_1 - \mu_2) > \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}. \tag{10}$$

Now, for $\gamma \in]0,1[$ we introduce

$$T_\gamma^* := \inf\left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, (\mu_1 - \mu_2) - \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} > \gamma(\mu_1 - \mu_2) \right\}.$$

This quantity is well defined as $\sigma_t^2(\alpha)\beta(t,\delta) \underset{t\to\infty}{\to} 0$. We then upper bound the expectation of τ:

$$\begin{aligned}
\mathbb{E}_\nu[\tau] &\leq T_\gamma^* + \sum_{T=T_\gamma^*+1} \mathbb{P}\left(\tau \geq T\right) \\
&\leq T_\gamma^* + \sum_{T=T_\gamma^*+1} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)} \right]^2 \right) \\
&\leq T_\gamma^* + \sum_{T=T_\gamma^*+1}^{\infty} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)} \gamma^2 (\mu_1 - \mu_2)^2 \right).
\end{aligned}$$

For all $t \in \mathbb{N}^*$, it is easy to show that the following upper bound on $\sigma_t^2(\alpha)$ holds:

$$\forall t \in \mathbb{N},\ \sigma_t^2(\alpha) \leq \frac{(\sigma_1 + \sigma_2)^2}{t} \times \frac{t - \frac{\sigma_1}{\sigma_2}}{t - \frac{\sigma_1}{\sigma_2} - 1}. \tag{11}$$

Using the bound (11), one has

$$\begin{aligned}
\mathbb{E}_\nu[\tau] &\leq T_\gamma^* + \int_0^{\infty} \exp\left(-\frac{t}{2(\sigma_1 + \sigma_2)^2} \frac{t - \frac{\sigma_1}{\sigma_2} - 1}{t - \frac{\sigma_1}{\sigma_2}} \gamma^2 (\mu_1 - \mu_2)^2 \right) dt \\
&= T_\gamma^* + \frac{2(\sigma_1 + \sigma_2)^2}{\gamma^2(\mu_1 - \mu_2)^2} \exp\left(\frac{\gamma^2(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2} \right).
\end{aligned}$$

We now give an upper bound on T_γ^*. Let $r \in [0, e/2 - 1]$. There exists $N_0(r)$ such that for $t \geq N_0(r)$, $\beta(t,\delta) \leq \log(t^{1+r}/\delta)$. Using also (11), one gets $T_\gamma^* = \max(N_0(t), \tilde{T}_\gamma)$, where

$$\tilde{T}_\gamma = \inf\left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}(1-\gamma)^2 t > \frac{t - \frac{\sigma_1}{\sigma_2} - 1}{t - \frac{\sigma_1}{\sigma_2}} \log \frac{t^{1+r}}{\delta} \right\}.$$

If $t > (1+\gamma\frac{\sigma_1}{\sigma_2})/\gamma$ one has $(t - \frac{\sigma_1}{\sigma_2} - 1)/(t - \frac{\sigma_1}{\sigma_2}) \leq (1-\gamma)^{-1}$. Thus $\tilde{T}_\gamma = \max((1+\gamma\frac{\sigma_1}{\sigma_2})/\gamma, T'_\gamma)$, with

$$T'_\gamma = \inf\left\{ t_0 \in \mathbb{N} : \forall t \geq t_0, \exp\left(\frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}(1-\gamma)^3 t\right) \geq \frac{t^{1+r}}{\delta} \right\}.$$

The following Lemma, whose proof can be found below, helps us bound this last quantity.

Lemma 14 *For every $\beta, \eta > 0$ and $s \in [0, e/2]$, the following implication is true:*

$$x_0 = \frac{s}{\beta}\log\left(\frac{e\log\left(1/(\beta^s\eta)\right)}{\beta^s\eta}\right) \quad \Rightarrow \quad \forall x \geq x_0, \;\; e^{\beta x} \geq \frac{x^s}{\eta}.$$

Applying Lemma 14 with $\eta = \delta$, $s = 1 + r$ and

$$\beta = \frac{(1-\gamma)^3(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}$$

leads to

$$T'_\gamma \leq \frac{(1+r)}{(1-\gamma)^3} \times \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + R(\mu_1,\mu_2,\sigma_1,\sigma_2,\gamma,r),$$

with

$$R(\mu_1,\mu_2,\sigma_1,\sigma_2,\gamma,r) = \frac{1+r}{(1-\gamma)^3}\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[1 + (1+r)\log\left(\frac{2(\sigma_1+\sigma_2)^2}{(1-\gamma)^3(\mu_1-\mu_2)^2}\right)\right].$$

Now for $\epsilon > 0$ fixed, choosing r and γ small enough leads to

$$\mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + \mathcal{C}(\mu_1,\mu_2,\sigma_1,\sigma_2,\epsilon),$$

where $\mathcal{C}$ is a constant independent of δ. This concludes the proof.

□

Proof of Lemma 14 Lemma 14 easily follows from the fact that for any $s, \eta > 0$,

$$x_0 = s\log\left(\frac{e\log\left(\frac{1}{\eta}\right)}{\eta}\right) \quad \Rightarrow \quad \forall x \geq x_0, \;\; e^x \geq \frac{x^s}{\eta}$$

Indeed, it suffices to apply this statement to $x = x\beta$ and $\eta = \eta\beta^s$. The mapping $x \mapsto e^x - x^s/\eta$ is increasing when $x \geq s$. As $x_0 \geq s$, it suffices to prove that x_0 defined above satisfies $e^{x_0} \geq x_0^s/\eta$.

$$\begin{aligned}
\log\left(\frac{x_0^s}{\eta}\right) &= s\log\left(s\log\left(\frac{e\log\frac{1}{\eta}}{\eta}\right)\right) + \log\frac{1}{\eta} \\
&= s\left(\log(s) + \log\left[\log\frac{1}{\eta} + \log\left(e\log\frac{1}{\eta}\right)\right]\right) + \log\frac{1}{\eta} \\
&\leq s\left(\log(s) + \log\left[2\log\frac{1}{\eta}\right]\right) + \log\frac{1}{\eta}
\end{aligned}$$

where we use that for all y, $\log(y) \leq \frac{1}{e}y$. Then

$$\log\left(\frac{x_0^s}{\eta}\right) \leq s\left(\log(s) + \log(2) + \log\log\frac{1}{\eta} + \log\frac{1}{\eta}\right).$$

For $s \leq \frac{e}{2}$, $\log(s) + \log(2) \leq 1$, hence

$$\log\left(\frac{x_0^s}{\eta}\right) \leq s\left(1 + \log\log\frac{1}{\eta} + \log\frac{1}{\eta}\right) = s\log\left(\frac{e\log\left(\frac{1}{\eta}\right)}{\eta}\right) = x_0,$$

which is equivalent to $e^{x_0} \geq \frac{x_0^s}{\eta}$ and concludes the proof.

Appendix D. An Optimal Static Strategy for Bernoulli Bandit Models

Bounding the probability of error of a static strategy using n_1 samples from arm 1 and n_2 samples from arm 2 relies on the following Lemma, that applies more generally to exponential families.

Lemma 15 *Let $(X_{1,t})_{t\in\mathbb{N}}$ and $(X_{2,t})_{t\in\mathbb{N}}$ be two independent i.i.d sequences, such that $X_{1,1} \sim \nu_{\theta_1}$ and $X_{2,1} \sim \nu_{\theta_2}$ belong to an exponential family. Assume that $\mu(\theta_1) > \mu(\theta_2)$. Then*

$$\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} < \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t}\right) \leq \exp(-(n_1+n_2)g_\alpha(\theta_1,\theta_2)), \tag{12}$$

where $\alpha = \frac{n_1}{n_1+n_2}$ and $g_\alpha(\theta_1,\theta_2) := \alpha\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) + (1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2)$.

The function $\alpha \mapsto g_\alpha(\theta_1,\theta_2)$, can be maximized analytically, and the value α^* that realizes the maximum is given by

$$\begin{aligned}
\mathrm{K}(\alpha^*\theta_1 + (1-\alpha^*)\theta_2, \theta_1) &= \mathrm{K}(\alpha^*\theta_1 + (1-\alpha^*)\theta_2, \theta_2) \\
\alpha^*\theta_1 + (1-\alpha^*)\theta_1 &= \theta^* \\
\alpha^* &= \frac{\theta^* - \theta_2}{\theta_1 - \theta_2}
\end{aligned}$$

where θ^* is defined by $\mathrm{K}(\theta^*, \theta_1) = \mathrm{K}(\theta^*, \theta_2) = \mathrm{K}^*(\theta_1, \theta_2)$. More interestingly, the associated rate is such that

$$g_{\alpha^*}(\theta_1,\theta_2) = \alpha^*\mathrm{K}(\theta^*,\theta_1) + (1-\alpha^*)\mathrm{K}(\theta^*,\theta_2) = \mathrm{K}^*(\theta_1,\theta_2),$$

which leads to Proposition 6.

Remark 16 *When $\mu_1 > \mu_2$, applying Lemma 15 with $n_1 = n_2 = t/2$ yields*

$$\mathbb{P}\left(\hat{\mu}_{1,t/2} < \mu_{2,t/2}\right) \leq \exp\left(-\frac{\mathrm{K}\left(\theta_1, \frac{\theta_1+\theta_2}{2}\right) + \mathrm{K}\left(\theta_2, \frac{\theta_1+\theta_2}{2}\right)}{2} t\right) = \exp\big(-I_*(\nu)t\big),$$

which shows that uniform sampling matches the lower bound of Theorem 2.

Proof of Lemma 15 The i.i.d. sequences $(X_{1,t})_{t\in\mathbb{N}}$ and $(X_{2,t})_{t\in\mathbb{N}}$ have respective densities f_{θ_1} and f_{θ_2} where $f_\theta(x) = A(x)\exp(\theta x - b(\theta)$ and $\mu(\theta_1) = \mu_1, \mu(\theta_2) = \mu_2$. α is such that $n_1 = \alpha n$ and $n_2 = (1-\alpha)n$. One can write

$$\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} - \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t} < 0\right) = \mathbb{P}\left(\alpha\sum_{t=1}^{n_2} X_{2,t} - (1-\alpha)\sum_{t=1}^{n_1} X_{1,t} \geq 0\right).$$

For every $\lambda > 0$, multiplying by λ, taking the exponential of the two sides and using Markov's inequality (this technique is often referred to as Chernoff's method), one gets

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} - \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t} < 0\right) &\leq \left(\mathbb{E}_\nu[e^{\lambda\alpha X_{2,1}}]\right)^{(1-\alpha)n}\left(\mathbb{E}_\nu[e^{\lambda(1-\alpha)X_{1,1}}]\right)^{\alpha n} \\ &= \exp\Bigg(\underbrace{\big[(1-\alpha)\phi_{X_{2,1}}(\lambda\alpha) + \alpha\phi_{X_{1,1}}(-(1-\alpha)\lambda)\big]}_{G_\alpha(\lambda)}\, n\Bigg)\end{aligned}$$

with $\phi_X(\lambda) = \log\mathbb{E}_\nu[e^{\lambda X}]$ for any random variable X. If $X \sim f_\theta$ a direct computation gives $\phi_X(\lambda) = b(\lambda+\theta) - b(\theta)$. Therefore the function $G_\alpha(\lambda)$ introduced above rewrites

$$G_\alpha(\lambda) = (1-\alpha)(b(\lambda\alpha+\theta_2) - b(\theta_2)) + \alpha(b(\theta_1 - (1-\alpha)\lambda) - b(\theta_1)).$$

Using that $b'(x) = \mu(x)$, we can compute the derivative of G and see that this function as a unique minimum in λ^* given by

$$\begin{aligned}\mu(\theta_1 - (1-\alpha)\lambda^*) &= \mu(\theta_2 + \alpha\lambda^*) \\ \theta_1 - (1-\alpha)\lambda^* &= \theta_2 + \alpha\lambda^* \\ \lambda^* &= \theta_1 - \theta_2,\end{aligned}$$

using that $\theta \mapsto \mu(\theta)$ is one-to-one. One can also show that

$$G(\lambda^*) = (1-\alpha)[b(\alpha\theta_1 + (1-\alpha)\theta_2) - b(\theta_2)] + \alpha[b(\alpha\theta_1 + (1-\alpha)\theta_2) - b(\theta_1)].$$

Using the expression of the KL-divergence between ν_{θ_1} and ν_{θ_2} as a function of the natural parameters: $\mathrm{K}(\theta_1,\theta_2) = \mu(\theta_1)(\theta_1-\theta_2) - b(\theta_1) + b(\theta_2)$, one can also show that

$$\begin{aligned}&\alpha\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) \\ &\qquad = -\alpha(1-\alpha)\mu(\alpha\theta_1 + (1-\alpha)\theta_2)(\theta_1-\theta_2) + \alpha[-b(\alpha\theta_1 + (1-\alpha)\theta_2) + b(\theta_1)] \\ &(1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2) \\ &\qquad = \alpha(1-\alpha)\mu(\alpha\theta_1 + (1-\alpha)\theta_2)(\theta_1-\theta_2) + (1-\alpha)[-b(\alpha\theta_1 + (1-\alpha)\theta_2) + b(\theta_2)]\end{aligned}$$

Summing these two equalities leads to

$$G(\lambda^*) = -\left[\alpha \mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) + (1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2)\right] = -g_\alpha(\theta_1, \theta_2).$$

Hence the inequality $\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} < \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t}\right) \leq \exp(G(\lambda^*)n)$ is exactly (12).