UM-DAE CEBS

Reading Project

Course-MPr-701

On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models

Author:
Atthaluri Shashank

Advisor:
Prof. Manjesh K. Hanawal

Contents

Abstract

The stochastic multi-armed bandit model is a simple abstraction that has proven useful in many different contexts in statistics and machine learning. our aim is to contribute to a better understanding of the performance in terms of identifying the m best arms. We introduce generic notion of complexity for fixed-confidence settings. In fixed-confidence setting, we provide the first known distribution-dependent lower bound on the complexity that involves information-theoretic quantities and holds when m $\geq$ 1 under general assumptions. In the specific case of two armed-bandits, we derive refined lower bounds in both the fixed-confidence along with matching algorithms for Gaussian and Bernoulli bandit models.

1 The Introduction

What is meant by Bandit? A bandit is a simple slot machine wherein you insert a coin into the machine, pull a lever, and get an immediate reward. But why is it called a bandit? It turns out all casinos configure these slot machines in such a way that all gamblers end up losing money! A multi-armed bandit is a complicated slot machine wherein instead of 1, there are several levers which a gambler can pull, with each lever giving a different return. The probability distribution for the reward corresponding to each lever is different and is unknown to the gambler. In classical stochastic bandit problem a gambler tries to maximize his revenue by sequentially playing one of a finite number of slot machine that are associated with initially unknown distributions. Assuming old-fashioned slot machines, the gambler pulls the arms of the machine one by one in a sequential manner, simultaneously learning about the machine probability distributions and gaining actual monetary reward. Here we are going to find the complexity of m best arms which give the maximum chance of winning the gift. We take bandit model ν which is a collection of K arms, where each arm ν_a (1 $\leq$ a $\leq$ K) is a probability distribution on $\mathbb{R}$ with expectation μ_a. At each time t = 1, 2, . . . , an agent chooses an option $A_t \in \{1, ..., K\}$ and receives an independent draw Z_t from the corresponding arm ν_{A_t}. We denote by $\mathbb{P}_\nu$ (resp.$\mathbb{E}_\nu$) the probability law (resp.expectation) of the process (Z_t). The agent's goal is to identify the m best arms, that is, the set S^*_m of indices of the m arms with highest expectation. Letting $(\mu_{[1]}, ..., \mu_{[K]})$ be the K-tuple of expectations $(\mu_1, ..., \mu_K)$ sorted in decreasing order, we assume that the bandit model ν belongs to a class $\mathcal{M}_m$ such that for every $\nu \in \mathcal{M}_m, \mu_{[m]} > \mu_{[m+1]}$, so that S^*_m is unambiguously defined. In order to identify S^*_m, the agent must use a strategy defining which arms to sample from, when to stop sampling, and which set ˆ to choose. More precisely, its strategy consists in a triple $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$ in which :

the sampling rule determines, based on past observations, which arm A_t is chosen at time t; in other words, At is $\mathcal{F}_{t-1} - measurable, with \mathcal{F}_t = \sigma(A_1, Z_1, ..., A_t, Z_t)$;

the stopping rule τ controls the end of the data acquisition phase and is a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$ satisfying $\mathbb{P}(\tau < +\infty) = 1$;

the recommendation rule provides the arm selection and is a $\mathcal{F}_\tau$- measurable random subset $\hat{S}_m of 1, ..., K$ of size m.

In the bandit literature, Here we consider a setting. i.e. fixedconfidence setting, In this a risk parameter δ is fixed, and a strategy $\mathcal{A}(\delta)$ is called δ-PAC if, for every choice of $\nu \in \mathcal{M}_m$, $\mathbb{P}_\nu(\hat{S}_m = S_m^*) \geq 1 - \delta$. The goal is to obtain δ-PAC strategies that require a number of draws τ_δ that is as small as possible. More precisely, we focus on strategies minimizing the expected number of draws $\mathbb{E}_\nu[\tau_\delta]$, which is also called the sample complexity. The subscript δ in τ_δ will be omitted when there is no ambiguity. We call a family of strategies $A = (\mathcal{A}(\delta))_{\delta\in(0,1)}$ PAC if for every δ, $\mathcal{A}(\delta) is \delta - PAC$. we define the complexity $\kappa_C(\nu)$ of best-arm identification in the fixed-confidence setting as follows:

$$\kappa_C(\nu) = \inf_{A\ PAC} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau_\delta]}{\log\frac{1}{\delta}} \tag{1.1}$$

For a given bandit model ν, and a small value of δ, a fixed-confidence optimal strategy needs an average number of samples of order $\kappa_C(\nu)\log\frac{1}{\delta}$ to identify the m best arms with probability at least 1-δ. Hera we are going to achieve two things 1) a lower bound on the sample complexity of any δ-PAC algorithm, 2) a δ-PAC strategy whose sample complexity attains the lower bound. We present below new lower bounds on $\kappa_C(\nu)$ that feature information-theoretic quantities as well as strategies that match these lower bounds in two-armed bandit models. using a uniform sampling strategy, that sample the arms in a round-robin fashion. Whereas it is well known that when $K > 2$ uniform sampling is not desirable, In this case, an algorithm using uniform sampling can be regarded as a statistical test of the hypothesis $H_0 : (\mu_1 \leq \mu_2) against H_1 : (\mu_1 > \mu_2)$ based on paired samples (X_s, Y_s) of ν_1, ν_2; namely a sequential test in the fixedconfidence setting, in which a (random) stopping rule determines when the experiment is to be terminated.

Consider for instance the case where $\nu_1 and \nu_2$ are Gaussian laws with the same known variance σ^2 , the means $\mu_1 and \mu_2$ 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 $H_0 : P = \mathcal{N}(\mu_1, \sigma^2) \bigotimes \mathcal{N}(\mu_2, \sigma^2) and H_1 : P = \mathcal{N}(\mu_2, \sigma^2) \bigotimes \mathcal{N}(\mu_1, \sigma^2)$. It is known since Wald (1945) that among the sequential tests such that type I and type II error probabilities are both smaller than δ, the Sequential Probability Ratio Test (SPRT) minimizes the expected number of required samples, and is such that $\mathcal{E}_\nu[\tau] \simeq 2\sigma^2/(\mu_1 - \mu_2)^2 \log(1/\delta)$. 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 In using randomized stopping strategies. We will show below that this conclusion is not valid anymore when the values of $\mu_1 and \mu_2$ are not assumed to be known.

2 Related Works

Bandit models have received a considerable interest since their introduction by Thompson in the context of medical trails. Bechhofer(1968) who consider the fixed-confidence setting and strategies based on uniform sampling. In the fixed confidence setting, Paulson (1964) first introduces a sampling strategy based on eliminations for single best arm identification: the arms are successively discarded, the remaining arms being sampled uniformly. Kalyanakrishnan later proposed an algorithm (2012) that is no longer based on eliminations, called LUCB (for Lower and Upper Confidence Bounds) and still designed for bounded bandit models. Bounded distributions are in fact particular examples of distributions with subgaussian tails, to which the proposed algorithms can be easily generalized.

$$H(\nu) = \sum_{a \ \in \ 1,2,\ldots K} \frac{1}{\triangle_a^2} \ with \tag{2.1}$$

$$\triangle_a = \begin{cases} \mu_a - \mu_{[m+1]} & for \ a \ \in \ S_m^*, \\ \mu_{[m]} - \mu_a & for \ a \ \in \ (S_m^*)^c. \end{cases}$$

Bubeck introduce the SAR (for Successive Accepts and Rejects) algorithm. An upper bound on the failure probability of the SAR algorithm yields $\kappa_B \ (\nu) \ \leq \ 8 \ \log(K)H \ (\nu)$. Gabillon(2012) propose the UGapE algorithm for m best-arm identification for m ¿ 1. By changing only one parameter in some confidence regions, this algorithm can be adapted either to the fixed-budget or to the fixed-confidence setting. However, a careful inspection shows that UGapE cannot be used in the fixed-budget setting without the knowledge of the complexity term H (ν).

3 Problem Setup

We first propose a distribution independent lower bound on $\kappa_C(\nu)$ that holds for m > 1 and for general classes of bandit models (Theorem 4). This information-theoretic lower bound permits to interpret the quantity $H(\nu)$ defined in (2.1) as a subgaussian approximation. Theorem 6 in Section 5 proposes a tighter lower bound on $\kappa_C(\nu)$ for general classes of two-armed bandit models, as well as a lower bound on the sample complexity of δ-PAC algorithms using uniform sampling. In Section 6 we propose, for Gaussian bandits with known—but possibly different—variances, an algorithm exactly matching this bound. We also consider the case of Bernoulli distributed arms, for which we show that uniform sampling is nearly optimal in most cases. We propose a new algorithm using uniform sampling and a non-trivial stopping strategy that is close to matching the lower bound. Lemma 1 provides a general relation between the expected number of draws and KullbackLeibler divergences of the arms' distributions, which is the key element to derive the lower bounds . Lemma 7 is a tight deviation inequality for martingales with sub-Gaussian increments, in the spirit of the Law of Iterated Logarithm, that permits here to derive efficient matching algorithms for two-armed bandits.

4 Generic Lower Bound in the Fixed-Confidence Setting

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

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

we make the assumption that there exists a set P of probability measures such that for all $\nu = (\nu_1, ..., \nu_K) \in \mathcal{M}_m, for \ a \in \{1, ..., K\}, \nu_a \in \mathcal{P}$ and that $\mathcal{P}$ satisfies

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

A class $\mathcal{M}_m$ of bandit models satisfying this property is called identifiable.

Let $N_a(t) = \sum_{s=1} t \infty_{As=a}$ be the number of draws of arm a between the instants 1 and t and $N_a = N_a(\tau)$ be the total number of draws of arm a by some algorithm $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$.

Lemma 1(dervies lower bound directly): Let ν and ν' be two bandit models with K arms such that for all a, the distributions ν_a and ν'_a are mutually absolutely continuous. For any almost-surely finite stopping time σ with respect to $(\mathcal{F}_t)$,

$$\sum_{a=1}^{k} \mathbb{E}_\nu[N_a(\sigma) KL(\nu_a, \nu'_a) \ \geq \ \sup_{\mathcal{E} \epsilon \mathcal{F}_\sigma} d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \tag{4.1}$$

where d(x,y) := x log(x/y) + (1 - x) log((1 - x)/(1 - y)) is the binary relative entropy, with the convention that d(0,0) = d(1,1) = 0.
proof : Can refer to Appendix.

Remark 2 This result can be considered as a generalization of Pinsker's inequality to bandit models: in combination with the inequality $d(p,q) \geq 2(p-q)^2$, it yields:

$$\sup_{\mathcal{E} \,\in\, \mathcal{F}_\sigma} |\mathbb{P}_\nu(\mathcal{E}) \ - \ \mathbb{P}_{\nu'}(\mathcal{E})| \ \leq \ \sqrt{\frac{\sum_{a=1}^{k} \ \mathbb{E}_\nu[N_a(\sigma)KL(\nu_a, \nu'_a)}{2}} \tag{4.2}$$

However, it is important not to use this weaker form of the statement, as we will consider events $\mathcal{E}$ of probability very close to 0 or 1. Here, we will make use of the following inequality:

$$\forall \ x \ \in \ [0,1], d(x, 1-x) \ \geq \ log\frac{1}{2.4x}, \tag{4.3}$$

which can be checked easily.

4.1 Lower Bound on the Sample Complexity of a δ-PAC Algorithm

We now propose a non-asymptotic lower bound on the expected number of samples which helps us to identify the m best arms in the fixed confidence setting, which straightforwardly yields a lower bound on $\kappa_C(\nu)$. Theorem 4 holds for an identifiable class of bandit models of the form:

$$\mathcal{M}_m = \{\nu = (\nu_1, ..., \nu_K) : \nu_i \ \epsilon \ \mathcal{P}, \mu_{[m]} \ > \ \mu_{[m+1]}\} \tag{4.4}$$

such that the set of probability measures $\mathcal{P}$ satisfies Assumption 3 below.

Assumption 3 For all p, q ϵ $\mathcal{P}^2$ such that p $\neq$ q, for all $\alpha \ > 0$,

$thereexistsq_1 \in \mathcal{P} \ : KL(p,q) < KL(p,q_1) < KL(p,q){+}\alpha and \mathbb{E}_{X \sim q_1}[X] \ > \ \mathbb{E}_{X \sim q}[X]$,

$thereexists \ q_2 \ \in \ \mathcal{P} \ : KL(p,q) < KL(p,q_2) < KL(p,q){+}\alpha and \mathbb{E}_{X \sim q_2}[X] \ < \ \mathbb{E}_{X \sim q}[X]$.

These continuity conditions of the assumptions that they include families of parametric bandits continuously parameterized by their means (e.g., Bernoulli, Poisson, exponential distributions).

Theorem 4 Let $\nu \ \epsilon \ \mathcal{M}_m$, where $\mathcal{M}_m$ is defined by (4.4), and assume that $\mathcal{P}$ satisfies Assumption 3; any algorithm that is δ-PAC on $\mathcal{M}_m$ satisfies, for δ $\leq$ 0.15,

$$\mathbb{E}_\nu[\tau] \ \geq \ \left[\sum_{a \in S_m^*} \frac{1}{KL(\nu_a, \nu_{[m+1]})} + \sum_{a \notin (S_m^*)} \frac{1}{KL(\nu_a, \nu_{[m]})}\right] \log(\frac{1}{2.4\delta}). \tag{4.5}$$

. Proof. Without loss of generality, one may assume that the arms are ordered such that $\mu_1 \geq ..... \geq \mu_K$. Thus $S_m^* = \{1, ..., m\}$.

Let $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$ be a δ-PAC algorithm and fix $\alpha \ > \ 0$.

For all a $\in \{1, ..., K\}$, from Assumption 3 there exists an alternative model

$$\nu' = (\nu_1, ..., \nu_{a-1}, \nu'_a, \nu_{a+1}, ..., \nu_K)$$

in which the only arm modified is arm a, and ν_a is such that:

$$KL(\nu_a, \nu_{m+1}) < KL(\nu_a, \nu_a) < KL(\nu_a, \nu_{m+1}) + \alpha and \mu'_a < \mu_{m+1} if a \in 1, ..., m,$$

$$KL(\nu_a, \nu_m) < KL(\nu_a, \nu'_a) < KL(\nu_a, \nu_m) + \alpha and \mu'_a > \mu_m if a \in m+1, ..., K.$$

In particular, on the bandit model ν' the set of optimal arms is no longer $\{1, ..., m\}$. Thus, introducing the event $\mathcal{E} = (\hat{S}_m = 1, ..., m) \in \mathcal{F}_\tau$, any δ-PAC algorithm satisfies $\mathbb{P}_\nu(\mathcal{E}) \ \geq 1-\delta$ and $\mathbb{P}'_\nu(\mathcal{E}) \ \geq \ \delta$. Lemma 1 applied to the stopping time τ (such that $N_a(\tau) = N_a$ is the total number of draws of arm a) and the monotonicity properties of d(x, y) (x $\mapsto$ d(x, y) is increasing when x > y and decreasing when x < y) yield

$$KL(\nu_a, \nu'_a)\mathbb{E}_\nu[N_a] \geq d(1-\delta, \delta) \geq \log\left(\frac{1}{2.4\delta}\right)$$

where the last inequality follows from (4.3). From the definition of the alternative model, one obtains for a $\in \{1, ..., m\} \, orb \in \{m+1, ..., K\}$ respectively, for every $\alpha \ > 0$,

$$\mathbb{E}_\nu[N_a] \geq \frac{log\left(\frac{1}{2.4\delta}\right)}{KL(\nu_a, \nu'_{m+1}) + \alpha} \; and \; \mathbb{E}_\nu[N_b] \geq \frac{log\left(\frac{1}{2.4\delta}\right)}{KL(\nu_b, \nu'_m) + \alpha}$$

Letting α tend to zero and summing over the arms yields the bound on $\mathbb{E}_\nu[\tau]$ $= \sum_{a=1}^{k} \mathbb{E}_\nu[N_a]$.

4.2 Bounds on the Complexity for Exponential Bandit Models

Theorem 4 yields the following lower bound on the complexity term:

$$\kappa_C(\nu) \geq \sum_{a \in S_m^*} \frac{1}{KL(\nu_a, \nu_{[m+1]})} + \sum_{a \notin S_m^*} \frac{1}{KL(\nu_a, \nu_{[m]})}$$

Thus, one may want to obtain strategies whose sample complexity can be proved to be of the same magnitude. The only algorithm that has been analyzed so far with an informationtheoretic perspective is the KL-LUCB algorithm of Kaufmann and Kalyanakrishnan (2013), designed for exponential bandit models: that is

$$\mathcal{M}_m = \left\{\nu = (\nu_{\theta_1}, ..., \nu_{\theta_K}) : (\theta_1, ...\theta_K) \in \Theta^K, \theta_{[m]} > \theta_{[m+1]}\right\}$$

where ν_θ belongs to a canonical one-parameter exponential family. This means that there exists a twice differentiable strictly convex function b such that ν_θ has a density with respect to some reference measure given by

$$f_\theta(x) = \exp(\theta x - b(\theta)), \;\; for\ \theta \in \Theta \subset \mathbb{R}. \tag{4.6}$$

Distributions from a canonical one-parameter exponential family can be parameterized either by their natural parameter θ or by their mean. Indeed $\dot{b}(\theta) = \mu(\theta)$, the mean of the distribution ν_θ and $\ddot{b}(\theta) = Var_{[\nu_\theta]} > 0$. The mapping $\theta \mapsto \mu(\theta)$ is strictly increasing, and the means are ordered in the same way as the natural parameters. Exponential families include in particular Bernoulli distributions, or Gaussian distributions with common variances (see Cappé et al. (2013) for more details about exponential families). We introduce the following shorthand to denote the Kullback-Leibler divergence in exponential families: $K(\theta,\theta') = KL(\nu_\theta,\nu_{\theta'})\ for(\theta,\theta') \in \Theta^2$. Combining the upper bound on the sample complexity of the KL-LUCB algorithm obtained by Kaufmann and Kalyanakrishnan (2013) and the lower bound of Theorem 4, the complexity $\kappa_C(\nu)$ can be bounded as

$$\sum_{a \in S_m^*} \frac{1}{K(\theta_a,\theta_{[m+1]})} + \sum_{a \notin S_m^*} \frac{1}{K(\theta_a,\theta_{[m]})} \leq \kappa_C(\nu) \leq 24 \min_{\theta \in [\theta_{[m+1]},\theta_{[m]}]} \sum_{a=1}^{K} \frac{1}{K^*(\theta_a,\theta)}, \tag{4.7}$$

where $K^*(\theta,\theta')$is the Chernoff information between the distributions $\nu_\theta\ and\ \nu_{\theta'}$ (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). Chernoff information is defined as follows and illustrated in Figure 1:

$$K^*(\theta,\theta') = K(\theta^*,\theta), \;\; where \theta * is such that K(\theta^*,\theta) = K(\theta *,\theta')$$

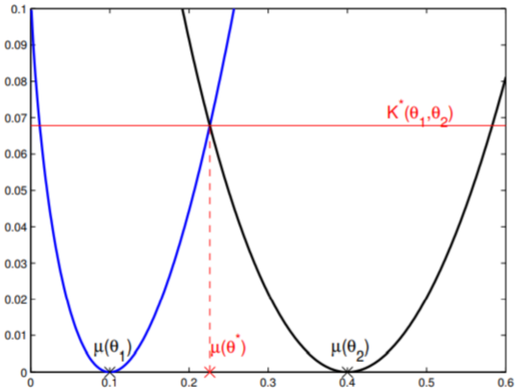


Figure 1: For Bernoulli distributions, the blue and black curves represent respectively $KL(\mathcal{B}(\mu),\mathcal{B}(\mu_1)$ and $KL(\mathcal{B}(\mu),\mathcal{B}(\mu_2)$ as a function of

μ. Their intersection gives the value of the Chernoff information between $\mathcal{B}(\mu_1)\ and \mathcal{B}(\mu_2)$, two distributions alternatively parameterized by their natural parameter $\theta_1\ and\ \theta_2$.

5 Improved Lower Bounds for Two-Armed Bandits

Two armed-bandits are of particular interest as they offer a theoretical framework for sequential A/B Testing(also known as split testing or bucket testing). It is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. How this test works? In an A/B test, you take a webpage and modify it to create a second version of the same page. This change can be a complete redesign of the page. Then, half of your traffic is shown the original version of the page (known as the control) and half are shown the modified version of the page (the variation).As visitors are served either the control or variation, their engagement with each experience is measured and collected, from that we determine the change in experience. Here In our case $\nu_1\ and\ \nu_2$ are probability distributions and $A_t\ \{1, 2\}$ as number of arms.
For two-armed bandits, the upper and lower bounds on the complexity $\kappa_C(\nu)$ given in (6) do not match. We propose in this section a refined lower bound on $\kappa_C(\nu)$ based on a different change of distribution. This lower bound features a quantity reminiscent of Chernoff information, and we will exhibit algorithms matching (or approximately matching) this new bound in Section 6. Theorem 6 provides a non-asymptotic lower bound on the sample complexity $\mathbb{E}_\nu[\tau]$ of any δ-PAC algorithm. It also provides a lower bound on the performance of algorithms using a uniform sampling strategy, which will turn out to be efficient in some cases.
Theorem 6 Let $\mathcal{M}$ be an identifiable class of two-armed bandit models and

let $\nu = (\nu_1, \nu_2) \in \mathcal{M}$ be such that $\mu_1 > \mu_2$. Any algorithm that is δ-PAC on $\mathcal{M}$ satisfies, for all $\delta \in (0, 1]$,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{c_*(\nu)} \log(\frac{1}{2.4\delta}), where c_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \max\{KL(\nu_1,\nu_1'), KL(\nu_2,\nu_2')\} \tag{5.1}$$

Moreover, any δ-PAC algorithm using a uniform sampling strategy satisfies,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{I_*(\nu)} \log(\frac{1}{2.4\delta}), where I_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \frac{KL(\nu_1,\nu_1'), KL(\nu_2,\nu_2')}{2} \tag{5.2}$$

Obviously, one has $I_*(\nu) \leq c_*(\nu)$. Theorem 6 implies in particular that $\kappa_C(\nu) \geq \frac{1}{c_*(\nu)}$. It is possible to give explicit expressions for the quantities $c_*(\nu) and I_*(\nu)$ for important classes of parametric bandit models that will be considered in the next section. The class of Gaussian bandits with known variances σ_1^2 and σ_2^2, further considered in Section 6.1, is

$$\mathcal{M} = \left\{\nu = (\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) : (\mu_1, \mu_2) \ \in \ \mathbb{R}^2, \mu_1 \neq \mu_2\right\} \tag{5.3}$$

For this class,

$$KL(\mathcal{N}(\mu_1, \sigma_1), \mathcal{N}(\mu_2, \sigma_2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}\right] \tag{5.4}$$

and direct computations yield

$$c_*(\nu) = \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2} \ \mathit{and} \ I_*(\nu) = \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}$$

. The observation that, when the variances are different $c_*(\nu) > I_*(\nu)$, will be shown to imply that strategies based on uniform sampling are suboptimalimal (by a factor $1 \leq \frac{2(\sigma_1^2+\sigma_2^2)}{(\sigma_1+\sigma_2)^2} \leq 2$).

The more general class of two-armed exponential bandit models, further considered in Section 6.2,

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

where ν_{θ_0} has density $\{_{\theta_a}$ given by(4.6). There

$$c_*(\nu) = \inf_{\theta\in\Theta} \max(K(\theta_1, \theta), K(\theta_2, \theta)) = K_*(\theta_1, \theta_2)$$

where $K_*(\theta_1, \theta_2) = K(\theta_1, \theta_*)$, with θ_* is defined by $\mathrm{K}(\theta_1,\ \theta_*) = \mathrm{K}(\theta_2,\ \theta_*)$. This quantity is analogous to the Chernoff information $K_*(\theta_1, \theta_2)$ introduced in Section 4 but with ‘reversed’roles for the arguments. $I_*(\nu)$ may also be expressed more explicitly as

$$I_*(\nu) = \frac{K(\theta_1, \bar{\theta}) + K(\theta_2, \bar{\theta})}{2}, \ \mathit{where} \ \mu(\bar{\theta}) = \frac{\mu_1 + \mu_2}{2}$$

We use the property that for two-armed exponential bandit models, the lower bound on $\kappa_C(\nu)$ provided by Theorem 6,

$$\kappa_C(\nu) \geq \left(\frac{1}{K_*(\theta_1, \theta_2)}\right) \tag{5.5}$$

is indeed always tighter than the lower bound of Theorem 4,

$$\kappa_C(\nu) \geq \left(\frac{1}{K_*(\theta_1, \theta_2)} + \frac{1}{K_*(\theta_2, \theta_1)}\right) \tag{5.6}$$

Interestingly, the changes of distribution used to derive the two results are not the same. On the one hand, for inequality (5.6), the changes of distribution involved modify a single arm at a time: one of the arms is moved just below (or just above) the other (see Figure 2, left). This is the idea also used, for example, to obtain the lower bound of Lai and Robbins (1985) on the cumulative regret. On the other hand, for inequality (5.5), both arms are modified at the same time: they are moved close to the common intermediate value θ_* but with a reversed ordering (see Figure 2, right).

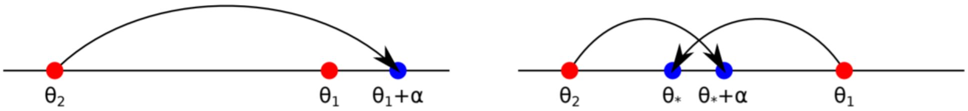


Figure2:Alternative bandit models considered to obtain the lower bounds of Theorem 4 (left) and Theorem 6 (right).

We now give the proof of Theorem 6, in order to show how easily it follows from Lemma 1.

Proof of Theorem 6. Without loss of generality, one may assume that the bandit model $\nu = (\nu_1, \nu_2)$ is such that the best arm is $a^* = 1$. Consider any alternative bandit model $\nu' = (\nu'_1, \nu'_2)$ in which $a^* = 2$. Let $\mathcal{E}$ be the event $\mathcal{E} = (\hat{S}_1 = 1)$, which belongs to $\mathcal{F}_\tau$. Let $\mathcal{A} = ((A_t), \tau, \hat{S}_1)$ be a δ-PAC algorithm: by assumptions, $\mathbb{P}_\nu(\mathcal{E}) \geq 1 - \delta$ and $\mathbb{P}_{\nu'}(\mathcal{E}) \leq \delta$ Applying Lemma 1 (with the stopping time τ) and using again the monotonicity properties of d(x, y) and inequality (3)

$$\mathbb{E}_\nu[N_1]KL(\nu_1, \nu'_1) + \mathbb{E}_\nu[N_2]KL(\nu_2, \nu'_2) \geq \log\left(\frac{1}{2.4\delta}\right) \tag{5.7}$$

Using moreover that $\tau = N_1 + N_2$, one has

$$\mathbb{E}_\nu[\tau] \geq \frac{\log\left(\frac{1}{2.4\delta}\right)}{\max_{a=1,2,} KL(\nu_a, \nu'_a)} \tag{5.8}$$

The result follows by optimizing over the possible model ν' satisfying $\mu'_1 < \mu'_2$ to make the right side of the inequality as large as possible. More precisely, for every $\alpha > 0$, from the definition of $c_*(\nu)$, there exists $\nu'_\alpha = (\nu'_1, \nu'_2)$ for which

$$\max_{a=1,2,} KL(\nu_a, \nu'_a) < c_*(\nu) + \alpha$$

Inequality (5.8) for the particular choice $\nu' = \nu'_\alpha$ *yields* $\mathbb{E}_\nu[\tau] \geq (c_*(\nu) + \alpha)^{-1} \log(1/(2.4\delta))$, and the first statement of Theorem 6 follows by letting α go to zero. In the particular case of exponential bandit models, the alternative model consists in choosing $\nu'_1 = \nu_{\theta_*}$ *and* $\nu'_1 = \nu_{\theta_*} + \epsilon$ for some ϵ, as illustrated on Figure 2, so that $\max_{a=1,2,} KL(\nu_a, \nu'_a)$ is of order $K_*(\theta_1, \theta_2)$. When $\mathcal{A}$ uses uniform sampling, using the fact that $\mathbb{E}_\nu[N_1] = \mathbb{E}_\nu[N_2] = \mathbb{E}_\nu[\tau]/2$ in Equation (5.7) similarly gives the second statement of Theorem 6.

6 Matching Algorithms for Two-Armed Bandits

For specific instances of two-armed bandit models, we now present algorithms with performance guarantees that closely match the lower bounds of

Theorem 6. For Gaussian bandits with known (and possibly different) variances, we describe in Section 6.1 an algorithm termed α-Elimination that is optimal and thus makes it possible to determine the complexity $\kappa_C(\nu)$. For Bernoulli bandit models, we present in Section 6.2 the SGLRT algorithm that uses uniform sampling and is close to optimal.

6.1 Gaussian Bandit Models

We focus here on the class of two-armed Gaussian bandit models with known variances presented in (5.3), where σ_1 and σ_2 are fixed. We prove that

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

by exhibiting a strategy that reaches the performance bound of Theorem 6. This strategy uses non-uniform sampling in case where σ_1 and σ_2 differ. When $\sigma_1 = \sigma_2$, we provide in Theorem 8 an improved stopping rule that is δ-PAC and results in a significant reduction of the expected number of samples used. The α-Elimination algorithm introduced in this Section can also be used in more general two-armed bandit models, where the distribution ν_a is σ_a^2- subgaussian. This means that the probability distribution ν_a satisfies

$$\forall \ \lambda \ \in \ \mathbb{R}, \mathbb{E}_{X \sim \nu_a}\left[e^{\lambda X}\right] \leq \frac{\lambda^2 \sigma_a^2}{2}$$

This covers in particular the cases of bounded distributions with support in [0, 1] (that are 1/4-subgaussian). In these more general cases, the algorithm enjoys the same theoretical properties: it is δ-PAC and its sample complexity is bounded as in Theorem 9 below.

6.1.1 Equal Variances

We start with the simpler case $\sigma_1 = \sigma_2 = \sigma$. Thus, the quantity $I_*(\nu)$ introduced in Theorem 6 coincides with $c_*(\nu)$, which suggests that uniform sampling could be optimal.A uniform sampling strategy equivalently collects paired samples (X_s, Y_s) from both arms. The difference $X_s - Y_s$ is normally distributed 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 both type I and type II error probabilities are bounded by δ. Robbins (1970) proposes the stopping rule

$$\tau = \inf\left\{t \ \in \ 2\mathbb{N}^* : |\textstyle\sum_s = 1^t/2(X_s - Y_s)| > \sqrt{2\sigma^2 t\beta(t,\delta)}\right\} .with\beta(t,\delta) = \tfrac{t+1}{t}\log\left(\tfrac{t+1}{2\delta}\right) \quad (6.1)$$

The recommendation rule chooses the empirically best arm at time τ . This procedure can be seen as an elimination strategy, in the sense of Jennison (1982). The authors of this paper derive a lower bound on the sample complexity of any δ-PAC elimination strategy (whereas our lower bound applies to any δ-PAC algorithm) which is matched by Robbins' algorithm: the above stopping rule τ satisfies

$$\lim_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} = \frac{8\sigma^2}{(\mu_1 - \mu_2)^2}$$

This value coincide with the lower bound on $\kappa_C(\nu)$ of Theorem 6 in the case of two-armed Gaussian distributions with similar known variance σ_2. This proves that in this case, Robbins' rule (6.1) is not only optimal among the class of elimination strategies, but also among the class of δ-PAC algorithm. Any δ-PAC elimination strategy that uses a threshold function (or exploration rate) $\beta(t, \delta)$ smaller than Robbins' also matches our asymptotic lower bound, while stopping earlier than the latter. From a practical point of view, it is therefore interesting to exhibit smaller exploration rates that preserve the δ-PAC property. The failure probability of such an algorithm is upper bounded, for example when $\mu_1 < \mu_2$, by

$$\mathbb{P}_\nu \left(\exists k \in \mathbb{N} : \textstyle\sum_{s=1^k} \frac{X_s - Y_s - (\mu_1 - \mu_2)}{\sqrt{2\sigma^2}} > \sqrt{2k\beta(2k, \delta)} \right) = \mathbb{P} \left(\exists k \in \mathbb{N} : S_k > \sqrt{2k\beta(2k, \delta)} \right) \quad (6.2)$$

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. In this paper, we propose a new deviation inequality for a martingale with sub-Gaussian increments, stated as Lemma 7, that permits to build an explicit confidence region reminiscent of the LIL. A related result was recently derived independently by Jamieson (2014).

Lemma 7 Let $\zeta(u) = \sum_{k \geq 1} k^{-u}$. Let $X_1, X_2, ...$ be independent random variables such that, for all $\lambda \in \mathbb{R}, \phi(\lambda) := \log \mathbb{E}[\exp(\lambda X1)] \leq \lambda^2\sigma^2/2$. For every positive integer t let $S_t = X_1 + ... + X_t$. Then, for all $\eta > 1 and x \geq \frac{8}{(e-1)2}$,

$$\mathbb{P} \left(\exists t \in \mathbb{N} : S_t > \sqrt{2\sigma^2 t (x + \eta \log\log(et))} \right) \leq \sqrt{e}\zeta \left(\eta \left(1 - \tfrac{1}{2x} \right) \right) \left(\tfrac{\sqrt{x}}{2\sqrt{2}} + 1 \right)^\eta \exp(-x)$$

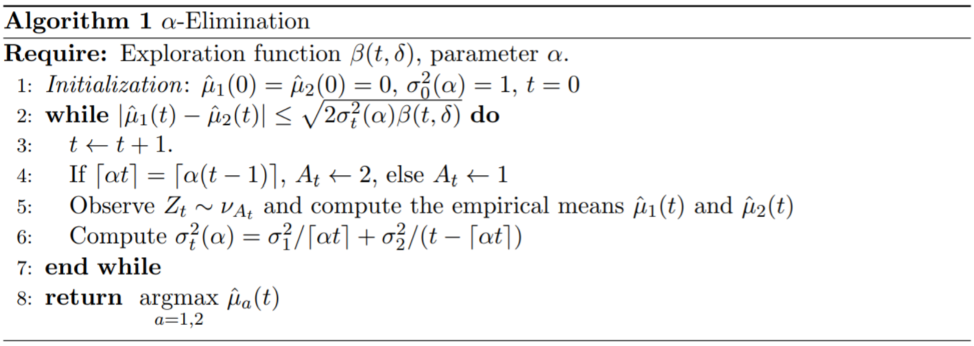
Lemma 7 allows to prove Theorem 8 below, as detailed in Appendix E, where we also provide a proof of Lemma 7 Theorem 8 For $\delta \leq 0.1$, with

$$\beta(t, \delta) = \log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log(\log(et/2)). \quad (6.3)$$

the elimination strategy is δ-PAC. We refer to Section 6 for numerical simulations that illustrate the significant savings (in the average number of samples needed to reach a decision) resulting from the use of the less conservative exploration rate allowed by Theorem 8.

6.2 Mismatched Variances

In the case where $\sigma_1 \neq \sigma_2$, we rely on the α - Elimination strategy, described in Algorithm 1 below. For a = 1, 2, $\hat{\sigma}_a(t)$ 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 $\alpha \in (0, 1)$, that maintains the proportion of draws of arm 1 close to α. At the end of every round t, $N_1(t) = [\alpha t], N_2(t) = t - [\alpha t] and \hat{\sigma}_1(t) - \hat{\sigma}_2(t) \sim \mathcal{N}(\sigma_1 - \sigma_2, \sigma_t^2(\alpha))$ (where $\sigma_t^2(\alpha)$ is defined at line 6 of Algorithm 1). The sampling schedule used here is thus deterministic



```
Algorithm 1 α-Elimination
Require: Exploration function $\beta(t,\delta)$, parameter $\alpha$.
1: Initialization: $\hat{\mu}_1(0) = \hat{\mu}_2(0) = 0$, $\sigma_0^2(\alpha) = 1$, $t = 0$
2: while $|\hat{\mu}_1(t) - \hat{\mu}_2(t)| \le \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}$ do
3:     $t \leftarrow t + 1$.
4:     If $\lceil \alpha t \rceil = \lceil \alpha(t-1) \rceil$, $A_t \leftarrow 2$, else $A_t \leftarrow 1$
5:     Observe $Z_t \sim \nu_{A_t}$ and compute the empirical means $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$
6:     Compute $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$
7: end while
8: return $\operatorname{argmax}_{a=1,2} \hat{\mu}_a(t)$
```

Theorem 9 shows that an optimal allocation of samples between the two arms consists in maintaining the proportion of draws of arm 1 close to $\sigma_1/(\sigma_1 + \sigma_2)$ (which is also the case in the fixed-budget setting, see Section 5.1). Indeed, for $\alpha = \sigma_1/(\sigma_1+\sigma_2)$, the α-elimination algorithm is δ-PAC with a suitable exploration rate and (almost) matches the lower bound on $\mathbb{E}_\nu[\tau]$, at least asymptotically when $\delta \to 0$. Its proof can be found in Appendix.

Theorem 9 If $\alpha = \sigma_1/(\sigma_1+\sigma_2) = \sigma_1/(\sigma_1+\sigma_2)$, 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 $\nu \,\epsilon\, \mathcal{M}$, for every $\in > 0$.

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

Remark 10 When $\sigma_1 = \sigma_2$, 1/2-elimination reduces, up to rounding effects, to the elimination procedure described in Section 6.1.1, for which Theorem 8 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 6.

6.3 Bernoulli Bandit Models

We consider in this section the class of Bernoulli bandit model

$$\mathcal{M} = \left\{\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1, \mu_2)\epsilon(0;1)^2, \mu_1 \neq \mu_2\right\}$$

where each arm can be alternatively parameterized by the natural parameter of the exponential family, $\theta_a = \log(\mu_a/(1-\mu_a))$. Observing that in this particular case little can be gained by departing from uniform sampling, we consider the SGLRT algorithm (to be defined below) that uses uniform sampling together with a stopping rule that is not based on the mere difference of the empirical means. For Bernoulli bandit models, the quantities $I_*(\nu) and c_*(\nu)$ introduced in Theorem 6 happen to be practically very close. There is thus a strong incentive to use uniform sampling and in the rest of this section we consider algorithms that aim at matching the bound (5.2) of Theorem 6—that is, $\mathbb{E}_\nu[\tau] \leq \log(1/\delta)/I_*(\nu)$, at least for small values of δ, which provides an upper bound on $\kappa_C(\nu)$ that is very close to $1/c_*(\nu)$. For simplicity, as $I_*(\nu)$ is here a function of the means of the arms only, we will denote $I_*(\nu)$ by $I_*(\mu_1, \mu_2)$. When the arms are sampled uniformly, finding an algorithm that matches the bound of (5.2) boils down to determining a proper 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 the 1/2 - Elimination procedure described in Algorithm 1 can be used, as each distribution ν_a is bounded and therefore 1/4 - subgaussian. More precisely, with $\beta(t, \delta)$ as in Theorem 8, the algorithm stopping at the first time t such that

$$\hat{\mu}_1(t) - \hat{\mu}_2(t) > \sqrt{2\beta(t,\delta)/t}$$

has its sample complexity bounded by $2/(\mu_1-\mu_2)^2 \log(1/\delta) + o(\log(1/\delta))$. 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 the bound (5.2) of Theorem 6. The approximation $I_*(\mu_1, \mu_2) = (\mu_1-\mu_2)^2/(8\mu_1(1-\mu_1)) + o((\mu_1-\mu_2)^2)$ suggests that the loss with respect to the optimal error exponent is particularly significant when both means are close to 0 or 1. To circumvent this drawback, we propose the SGLRT (for Sequential Generalized Likelihood Ratio Test) stopping rule, described in Algorithm 2. The appearance of I_* in the stopping criterion of Algorithm 2 is a consequence of the observation that it is related to the generalized likelihood ratio statistic for testing the equality of two Bernoulli proportions. To test $H_0 : (\mu_1 = \mu_2) against H_1 : (\mu_1 \neq \mu_2)$ based on t/2 paired samples of the arms $W_s = (X_s, Y_s)$, the Generalized Likelihood Ratio Test (GLRT) rejects H_0 when

$$\frac{\max_{\mu_1,\mu_2:\mu_1=\mu_2} L(W_1, ..., W_{t/2}; \mu_1, \mu_2)}{\max_{\mu_1,\mu_2} L(W_1, ..., W_{t/2}; \mu_1, \mu_2)} < z_\delta$$

where $L(W_1, ..., W_{t/2}; \mu_1, \mu_2)$ denotes the likelihood of the observations given parameters $\mu_1 and \mu_2$. It can be checked that the ratio that appears in the last display is equal to $\exp(-tI_*(\hat{\sigma}_1, t/2, \hat{\sigma}_2, t/2))$. This equality is a consequence of the rewriting

$$I_*(x, y) = H\left(\frac{x+y}{2}\right) - \frac{1}{2}[H(x) + H(y)],$$

where $H(x) = -x\log(x) - (1-x)\log(1-x)$ denotes the binary entropy function. Hence, Algorithm (2) can be interpreted as a sequential version of the GLRT with (varying) threshold $z_{t,\delta} = \exp(-\beta(t, \delta))$.

```
**Algorithm 2** Sequential Generalized Likelihood Ratio Test (SGLRT)
**Require:** Exploration function $\beta(t, \delta)$.
*Initialization*: $\hat{\mu}_1(0) = \hat{\mu}_2(0) = 0.\ t = 0.$
**while** $(tI_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) \leq \beta(t, \delta)) \bigcup (t = 1\ (mod.\ 2))$ **do**
    $t = t + 1.\ A_t = t\ (mod.\ 2).$
    Observe $Z_t \sim \nu_{A_t}$ and compute the empirical means $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$.
**end while**
**return** $a = \underset{a=1,2}{\text{argmax}}\ \hat{\mu}_a(t).$
```

Elements of analysis of the SGLRT. The SGLRT algorithm is also related to the KLLUCB algorithm of Kaufmann and Kalyanakrishnan (2013). A closer examination of the KL-LUCB stopping criterion reveals that, in the specific case of two-armed bandits, it is equivalent to stopping when $tKL_*(\mathcal{B}(\hat{\mu}_1(t)), \mathcal{B}(\hat{\mu}_2(t)))$ gets larger than some threshold We also mentioned the fact that $KL_*(\mathcal{B}(x), \mathcal{B}(y)) and I_*(x, y)$ are very close. Using results from Kaufmann and Kalyanakrishnan (2013), one can thus prove (see Appendix) the following lemma. Lemma 11 With the exploration rate

$$\beta(t, \delta) = 2\log\left(\frac{t(\log(3t))^2}{2}\right)$$

the SGLRT algorithm is δ-PAC. For this exploration rate, we were able to obtain the following asymptotic guarantee on the stopping time τ of Algorithm 2:

$$\forall\epsilon > 0, \limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \leq \frac{(1+\epsilon)}{I_*(\mu_1, \mu_2)}\ a.s$$

(see Lemma 26 in Appendix F for the proof of this result). By analogy with the result of Theorem 8 we conjecture that the analysis of Kaufmann and Kalyanakrishnan (2013)—on which the result of Lemma 11 is based—is too conservative and that the use of an exploration rate of order $\log(\log(t)/\delta)$ should also lead to a δ-PAC algorithm. This conjecture is supported by the numerical experiments reported in Section 6 below. Besides, for this choice of exploration rate, Lemma 26 also shows that

$$\forall \epsilon > 0, \lim_{\delta \to 0} \sup \frac{\tau}{\log(1/\delta)} \leq \frac{(1+\epsilon)}{I_*(\mu_1, \mu_2)} \; a.s$$

7 Conclusion

For two-armed bandits, we obtained rather complete results, identifying the complexity of fixed confidence settings in important parametric families of distributions. In doing so, we observed that standard testing strategies based on uniform sampling are optimal or close to optimal for Gaussian distributions with matched variance or Bernoulli distributions but can be improved (by non-uniform sampling) for Gaussian distributions with distinct variances. This latter observation can certainly be generalized to other models, starting with the case of Gaussian distributions whose variances are a priori unknown. In the case of Bernoulli distributions, we have also shown that fixed-confidence algorithms that use the difference of the empirical means as a stopping criterion are bound to be sub-optimal. For models with more than two arms, we obtained the first generic (i.e. not based on the sub-Gaussian tail assumption) distribution-dependent lower bound on the complexity of m best-arms identification in the fixed-confidence setting (Theorem 4). Currently available performance bounds for algorithms performing m best-arms identification—those of Kaufmann and Kalyanakrishnan (2013) notably—show a small gap with this result and it is certainly of interest to investigate whether those analyses and/or the bound of Theorem 4 may be improved to bridge the gap. Where do we use these bandit algorithms? Bandit algorithms are being used in a lot of research projects in the industry, clinical trials, network routing, online advertising, game designing etc..

8 Acknowledgement

I would like to thank Prof. Manjesh K. Hanawal for his fruitful discussion, despite his busy schedule, he took the time to teach me all the important concepts needed to go through with this project. I also thank Prof. K.S. Malikarjun Rao for suggesting me such an experienced guide.

9 Appendix

Changes of Distributions

Proof of Lemma 1 To prove Lemma 1, we state a first inequality on the expected log-likelihood ratio in Lemma 19, which is of independent interest.

Lemma 19 Let σ be any almost surely finite stopping time with respect to $\mathcal{F}_\tau$. For every event $\mathcal{E} \in \mathcal{F}_\sigma$,

$$\mathbb{E}_\nu[L_\sigma] \geq d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \tag{9.1}$$

Lemma 1 easily follows: introducing (Y_a, s), the sequence of i.i.d. samples successively observed from arm a, the log-likelihood ratio L_t can be rewritten

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

Wald's Lemma (see e.g., Siegmund (1985)) applied to $L_\sigma = \sum_{a=1}^{K} \sum_{s=1}^{N_a(\sigma)} \log\left(\frac{f_a(Y_a,s)}{f'_a(Y_a,s)}\right)$ yields

$$\mathbb{E}_\nu[L_\sigma] = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\sigma)] KL(\nu_a, \nu'_a) \tag{9.2}$$

Combining this equality with the inequality in Lemma 19 completes the proof.

Appendix. A Refined Exploration Rate for α-Elimination

Proof of Theorem 8

According to (6.2), to prove Theorem 8 it is enough to show that for

$$\beta(t,\delta) = \log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log(\log(et/2)).$$

if $S_t = \sum_{s=1}^{t} X_s$ is a sum of i.i.d $\mathcal{N}(0,1)$ random variables, one has

$$\mathbb{P}(\exists t \in \mathbb{N}^* \ : S_t > \sqrt{2t\beta(t,\delta)} \leq \ \delta \tag{9.3}$$

Let $z = \log(1/\delta)$. Using Lemma 7, one can write, choosing $x = z + 3\log z and \beta = 3/2$,

$$\mathbb{P}(\exists t \in \mathbb{N}^* \ : S_t > \sqrt{2t\beta(t,\delta)} \leq \frac{\sqrt{e}}{8}\zeta\left(\frac{3}{2} - \frac{3}{4(z+3\log z)}\right)\frac{(\sqrt{z+3\log z}+\sqrt{8})^{3/2}}{z^3}\delta.$$

It can be shown numerically that for $z \geq 2.03$,

$$\frac{\sqrt{e}}{8}\zeta\left(\frac{3}{2} - \frac{3}{4(z+3\log z)}\right)\frac{(\sqrt{z+3\log z}+\sqrt{8})^{3/2}}{z^3} \leq 1$$

Thus for $\delta \leq \exp(2.03) \leq 0.1$, inequality (9.3) holds Proof of Lemma 7.

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

Lemma 23 For every ¿ 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 24 For every $\eta > 0$,

$$A(\eta) := \frac{4}{((1+\eta)^{1/4} + (1+\eta)^{-1/4})^2} \geq 1 - \frac{\eta^2}{16} \tag{9.4}$$

Lemma 25 Let t be such that $(1+\eta)^{k-1} \leq t \leq (1+\eta)^k$. Then

$$\sigma\sqrt{2}z \geq \frac{A(\eta)z}{\lambda\sqrt{t}} + \frac{\lambda\sigma^2\sqrt{t}}{2} with\lambda = \sigma^{-1}\sqrt{2zA(\eta)/(1+\eta)^{k-1/2}}.$$

Proof of Lemma 25

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

according to Lemma 23

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

$$\mathbb{P}\left(\bigcup_{t\geq 1}\left\{\lambda S_t - t\frac{\lambda\sigma^2}{2} > u\right\}\right) \leq \exp(-u). \tag{9.5}$$

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

$$\mathbb{P}\left(\bigcup_{t1}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right)\sum_{k=1}^{\infty}\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right)$$
$$\leq \sum_{k=1}^{\infty}\mathbb{P}\left(\bigcup_{t\in T_k}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log(k\log(1+\eta))}\right\}\right)$$

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 $k \geq 1$, let $z_k = x + \beta\log(k\log(1+\eta)) and\lambda_k = \sigma^{-1}\sqrt{2z_kA(\eta)/(1+\eta)^{k-1/2}}$ Lemma 25 shows that for every $t \in T_k$,

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

Thus, by inequality (9.5),

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

One chooses $\eta^2=8/\text{x}$ for x such that $x\geq\frac{8}{(e-1)/2}$ (which ensures $\eta\leq e-1$). Using Lemma 24, one obtains that $\exp(A(\eta)x)\leq\sqrt{e}\exp(x)$. Moreover,

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

Thus

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

using the lower bound on $A(\eta)$ given in Lemma 24 and the fact that $A(\eta)$ is upper bounded by 1.

Appendix . Bernoulli Bandit Models

Proof of Lemma 11 Assume that $\mu_1<\mu_2$. Recall the KL-LUCB algorithm of Kaufmann and Kalyanakrishnan (2013). For two-armed bandit models, this algorithm samples the arms uniformly and builds for both arms a confidence interval based on KL-divergence $\mathcal{I}_a(t)=[l_{a,t/2},u_{a,t/2}]$, with

$$u_{a,s}=\sup\left\{q>\hat{\mu}_{a,s}:sd(\hat{\mu}_{a,s},q)\leq\tilde{\beta}(s,\delta)\right\},\ where\ d(x,y)=KL(\mathcal{B}(x),\mathcal{B}(y))$$
$$l_{a,s}=\sup\left\{q<\hat{\mu}_{a,s}:sd(\hat{\mu}_{a,s},q)\leq\tilde{\beta}(s,\delta)\right\}$$

for some exploration rate that we denote by $\tilde{\beta}(t,\delta)$. The algorithm stops when the confidence intervals are separated; that is either $l_{1,t/2}>u_{2,t/2}orl_{2,t/2}>u_{1,t/2}$, and recommends the empirical best arm. A picture helps to convince oneself that

$$(l_{1,s}\ >\ u_{2,s})\Leftrightarrow(\hat{\mu}_{1,s}>\hat{\mu}_{2,s})\cap sd_*(\hat{\mu}_{1,s},\hat{\mu}_{2,s})>\beta(s,\delta)$$

Additionally, as mentioned before, $I_*(x,y)$ is very close to the quantity $d_*(x,y)$ and one has more precisely $I_*(x,y)<d_*(x,y)$. Using all this, we

can upper bound the probability of error of Algorithm 2 in the following way.

$$\begin{aligned}
&\mathbb{P}_\nu(\exists t \in 2\mathbb{N}^* : \hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}, tI_*(\hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}) > \beta(t,\delta)\\
&\leq \mathbb{P}_\nu(\exists t \in 2\mathbb{N}^* : \hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}, (t/2)d_*(\hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}) > (\beta(t,\delta)/2)\\
&= \mathbb{P}_\nu(\exists s \in 2\mathbb{N}^* : \hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}, sd_*(\hat{\mu}_{1,s} > \hat{\mu}_{2,s}) > (\beta(2s,\delta)/2)\\
&= \mathbb{P}_\nu(\exists s \in 2\mathbb{N}^* : l_{1,s} > u_{2,s}) \leq \mathbb{P}_\nu(\exists s \in 2\mathbb{N}^* : (\mu_1 < l_{1,s}) \cup (\mu_2 > u_{2,s}))\\
&\leq 2\sum_{s=1}^{\infty} \exp(-\beta(2s,\delta)/2)
\end{aligned}$$

where the last inequality follows from an union bound and for example Lemma 4 of Kaufmann and Kalyanakrishnan (2013). Note that the indices $l_{1,s}$ *and* $u_{2,s}$ involved here use the exploration rate $\widetilde{\beta}(s,\delta) = \beta(2s,\delta)/2$. The choice $\beta(t,\delta)$ in the statement of the Lemma shows the last series is upper bounded by δ, which concludes the proof

References

[1] Emilie Kaufmann, Olivier Cappe, Aurelien Garivier, On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models, Journal of Machine Learning Research 17 (2016)

[2] Victor Gabbillon, Mohammad Ghavamzadeh, Alessandro Lazaric, Sebastien Bubeck, Multi-Bandit Best Arm Identification

[3] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, Peter Stone, PAC Subset Selection in Stochastic Multi-armed Bandits, In langford, Pineau, editors, Proceedings of the 29th International Conference on Machine Learning. pp. 655–662, Omnipress, New York,NY, USA, 2012.

[4] Jean-Yves Audibert, Sebastien Bubeck, Remi Munos, Best Arm Identification in Multi-Armed Bandits.