Explore First, Exploit Next: The True Shape of Regret in Bandit Problems

Aurélien Garivier, Pierre Ménard, Gilles Stoltz

Explore First, Exploit Next: The True Shape of Regret in Bandit Problems

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR
Pierre Ménard PMENARD@MATH.UNIV-TOULOUSE.FR
IMT: Université Paul Sabatier – CNRS, Toulouse, France

Gilles Stoltz STOLTZ@HEC.FR
GREGHEC: HEC Paris – CNRS, Jouy-en-Josas, France

Abstract

We revisit lower bounds on the regret in the case of multi-armed bandit problems. We obtain non-asymptotic bounds and provide straightforward proofs based only on well-known properties of Kullback-Leibler divergences. These bounds show that in an initial phase the regret grows almost linearly, and that the well-known logarithmic growth of the regret only holds in a final phase. The proof techniques come to the essence of the arguments used and they are deprived of all unnecessary complications.

Keywords: multi-armed bandits, cumulative regret, information-theoretic proof techniques, non-asymptotic lower bounds

1. Introduction

After the works of Lai and Robbins (1985) and Burnetas and Katehakis (1996), it is widely admitted that the growth of the cumulative regret in a bandit problem is a logarithmic function of time, multiplied by a sum of terms involving Kullback-Leibler divergences. The asymptotic nature of the lower bounds, however, appears spectacularly in numerical experiments, where the logarithmic shape is not to be observed on small horizons (see Figure 1, left). Even on larger horizons, the second-order terms keeps a large importance, which causes the regret of some algorithms to remain way *below* the "lower bound" on any experimentally visible horizon (see Figure 1, right).

In this paper, we revisit this question by drawing a more precise picture of the behavior of the regret. We derive non-asymptotic bounds showing the existence of three successive phases: an initial linear phase, when all the arms are essentially drawn uniformly; a transition phase, when the number of observations becomes sufficient to perceive differences; and the final phase, when the distributions associated with all the arms are known with high confidence and when the new draws are just confirming the identity of the best arms with higher and higher degree of confidence (this is the famous logarithmic phase). This last phase may often be out of reach in applications, especially when the number of arms is large.

We provide straightforward proofs, based only on well-known properties of Kullback-Leibler divergences (in particular, they avoid explicit changes of measures). These proof techniques come to the essence of the arguments used so far in the literature and they are deprived of all unnecessary complications. (A detailed comparison to the literature is offered below.)

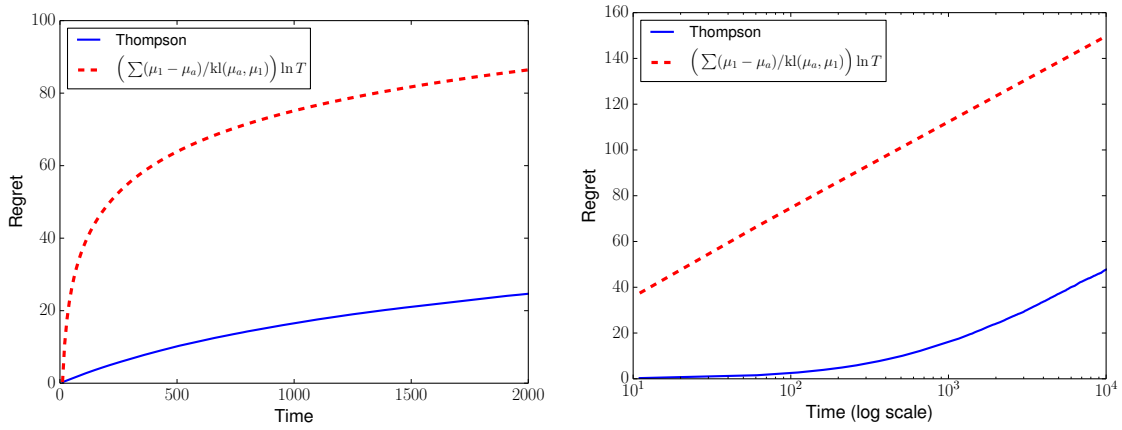

Figure 1: Expected regret of Thompson (1933) Sampling (*blue, solid* line) on a Bernoulli bandit problem with parameters $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05,\, 0.04,\, 0.02,\, 0.015,\, 0.01,\, 0.005)$; expectations are approximated over 500 runs.

Versus the Lai and Robbins (1985) lower bound (*red, dotted* line) for a Bernoulli model; here kl denotes the Kullback-Leibler divergence (3) between Bernoulli distributions.

Left: the shape of regret is not logarithmic at first, rather linear.
Right: the asymptotic lower bound is out of reach unless T is extremely large.

1.1. Setting

We consider the simplest case of a stochastic bandit problem, with finitely many arms indexed by $a \in \{1, \dots, K\}$. Each of these arms is associated with an unknown probability distribution ν_a over $\mathbb{R}$. At each round $t \geqslant 1$, the player pulls the arm A_t and gets a reward Y_t drawn at random according to the distribution ν_{A_t}. This reward is the only piece of information available to the player.

Strategies. A strategy ψ associates an arm with the information gained in the past, possibly based on some auxiliary randomization; without loss of generality, this auxiliary randomization is provided by a sequence $U_0, U_1, U_2, \dots$ of independent and identically distributed random variables, with common distribution the uniform distribution over $[0, 1]$. These variables are also independent of the randomization generating the rewards Y_t. Thus, a strategy is a sequence $\psi = (\psi_t)_{t\geqslant 0}$ of measurable functions, each of which associates with the said past information, namely,

$$I_t = \big(U_0,\, Y_1, U_1,\, \dots,\, Y_t, U_t\big)\,,$$

an arm $\psi_t(I_t) = A_{t+1} \in \{1, \dots, K\}$, where $t \geqslant 0$. The initial information reduces to $I_1 = U_0$ and the first arm is $A_1 = \psi_0(U_0)$.

Probability measures. By Kolmogorov's extension theorem, there exist indeed a measurable space $(\Omega, \mathcal{F})$ such that all probability measures considered above can be defined on the same probability space, a fact we will need later to perform implicit changes of measures. One can take, for instance, $\Omega = [0, 1] \times \big(\mathbb{R} \times [0, 1]\big)^{\mathbb{N}}$.

The probabilistic and strategic setting can thus be formalized as follows. Denoting vector of probability distributions associated with the arms by $\nu = (\nu_a)_{a=1,\dots,K}$, the probability measure $\mathbb{P}_\nu$ over this $(\Omega, \mathcal{F})$ is such that for all $t \geqslant 0$, for all Borel sets $B \subseteq \mathbb{R}$ and $B' \subseteq [0,1]$,

$$\mathbb{P}_\nu\big(Y_{t+1} \in B,\, U_{t+1} \in B' \;\big|\; I_t\big) = \nu_{\psi_t(I_t)}(B)\; \lambda(B')\,, \tag{1}$$

where λ denotes the Lebesgue measure on $[0,1]$.

Regret. A typical measure of the performance of a strategy is given by its regret. To recall its definition, we first denote by $E(\nu_a) = \mu_a$ the expected payoff of arm a and by Δ_a its gap to an optimal arm:

$$\mu^\star = \max_{a=1,\dots,K} \mu_a \qquad \text{and} \qquad \Delta_a = \mu^\star - \mu_a\,.$$

Second, the number of times an arm a is pulled till round T is referred to as

$$N_a(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}\,.$$

Then, the expected regret of the strategy equals, by the tower rule,

$$R_{\nu,T} = T\mu^\star - \mathbb{E}_\nu\left[\sum_{t=1}^{T} Y_t\right] = \mathbb{E}_\nu\left[\sum_{t=1}^{T}\big(\mu^\star - \mu_{A_t}\big)\right] = \sum_{a=1}^{K} \Delta_a\, \mathbb{E}_\nu\big[N_a(T)\big]\,.$$

1.2. Existing lower bounds: a quick literature review

We consider a model $\mathcal{D}$, i.e., a collection of possible distributions ν_a associated with the arms. Lai and Robbins (1985) and later Burnetas and Katehakis (1996) exhibited asymptotic lower bounds and matching asymptotic upper bounds on the normalized regret $R_{\nu,T}/\ln T$, respectively in a one-parameter case and in a general, non-parametric case.

The key quantity $\mathcal{K}_{\inf}$. To state the most general bound, the one of Burnetas and Katehakis (1996), we first denote by KL the Kullback-Leibler divergence between two probability distributions and recall that we denoted by E the expectation operator (that associates with each distribution its expectation). Now, given $\nu_a \in \mathcal{D}$ and a real number x, we introduce

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) :\;\; \nu'_a \in \mathcal{D} \;\text{ and }\; E(\nu'_a) > x\Big\}\,.$$

Burnetas and Katehakis (1996, conditions A1–A3) consider rather mild conditions on the model $\mathcal{D}$ and on the strategy at hand (in particular, its consistency in the sense of Definition 2, but not only). Under these conditions, for any suboptimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,. \tag{2}$$

Previous simplifications of the proof. There were recent attempts to clarify the exposition of the proof of this lower bound, together with the desire of relaxing the stated conditions. The case of Bernoulli models is treated in Bubeck (2010) and Bubeck and Cesa-Bianchi (2012). Only assumptions of consistency of the strategies are required and the associated proof follows the original proof technique, by performing first an explicit change of measure and then applying some Markov–Chernoff bounding.

Recently, Kaufmann et al. (2016) dealt with the case of any model $\mathcal{D}$ but with the restriction that only bandits problems $\nu = (\nu_a)_{a=1,\dots,K}$ with a unique optimal arm should be considered. They still use both an explicit change of measure –to prove the chain-rule equality in (F)– and then apply as well some Markov–Chernoff bounding to the probability of well-chosen events.

Furthermore, Wu et al. (2015) independently developed non-asymptotic problem-dependent lower bounds on the regret of any algorithm, with a focus on more general limited feedback models than just the simplest case of multi-armed bandit problems as in the present article. Their lower bounds can recover as well the asymptotic bounds of Burnetas and Katehakis (1996), and also finite-time minimax lower bounds. These lower bounds are in terms of uniform upper bounds on the regret of the considered strategies, which is in contrast with the lower bounds we develop in Section 3. Therein, we assume extremely mild assumptions on the strategies, if any (some minimal symmetry, for instance) and do not need their regret to be bounded from above.

Concerning distribution-free lower bounds on the regret, for which a special case of the chain-rule equality in (F) is also fundamental, the optimal order is $\sqrt{TK}$ (see Auer et al., 2002b, Cesa-Bianchi and Lugosi, 2006, and for two-armed bandits, Kulkarni and Lugosi, 2000).

1.3. Our contributions

In Section 2, we present Inequality (F), in our opinion the most efficient and most versatile tool for proving lower bounds in bandit models. We carefully detail its remarkably simple proof, together with an elegant derivation of the Burnetas and Katehakis (1996) asympototic lower bound. The power of Inequality (F) is illustrated in Section 3: we study the initial regime when the small number T of draws does not yet permit to unambiguously identify the best arm. We propose three different bounds (each with specific merits). They explain the quasi-linear growth of the regret in this initial phase. We also discuss how the length of the initial phase depends on the number of arms and on the gap between optimal and sub-optimal arms in Kullback-Leibler divergence. Section 4 contains a general non-asymptotic lower bound for the logarithmic (large T) regime. This bound does not only contain the right leading term, but the analysis aims at highlighting what the second-order terms depend on. Results of independent interest on the regularity (upper semi-continuity) of $\mathcal{K}_{\inf}$ are provided in its Subsection 4.2.

2. The fundamental inequality

Our starting point consists of two building blocks, a standard equality and a less standard inequality,

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

where $\mathbb{P}_\nu^{I_T}$ and $\mathbb{P}_{\nu'}^{I_T}$ denote the respective image distributions of I_T under $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$, where kl denotes the Kullback-Leibler divergence for Bernoulli distributions,

$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}\,, \tag{3}$$

and where Z is any $\sigma(I_T)$–measurable random variable with values in $[0,1]$.

Application of this inequality. We will typically consider variables of the form $Z = N_k(T)/T$ for some arm k. That the kl term in (F) then also contains expected numbers of draws of arms will be very handy. Unlike all previous proofs of distribution-dependent lower bounds for bandit problems, we will not have to introduce well-chosen events and control their probability by some Markov–Chernoff boundings.

2.1. Proof of the standard equality in (F)

It can be found, e.g., in the proofs of the distribution-free lower bounds on the bandit regret, in the special case of Bernoulli distributions, see Auer et al. (2002b); Cesa-Bianchi and Lugosi (2006); see also Combes and Proutière (2014). We thus reprove this equality for the sake of completeness only. The chain rule for Kullback-Leibler divergences ensures that for all $t \geqslant 1$,

$$\begin{aligned} \mathrm{KL}\Big(\mathbb{P}_\nu^{I_{t+1}}, \mathbb{P}_{\nu'}^{I_{t+1}}\Big) &= \mathrm{KL}\Big(\mathbb{P}_\nu^{(I_t, Y_{t+1}, U_{t+1})}, \mathbb{P}_{\nu'}^{(I_t, Y_{t+1}, U_{t+1})}\Big) \\ &= \mathrm{KL}\big(\mathbb{P}_\nu^{I_t}, \mathbb{P}_{\nu'}^{I_t}\big) + \mathrm{KL}\Big(\mathbb{P}_\nu^{(Y_{t+1}, U_{t+1})\,|\,I_t}, \mathbb{P}_{\nu'}^{(Y_{t+1}, U_{t+1})\,|\,I_t}\Big) \end{aligned} \tag{4}$$

where the conditional Kullback-Leibler divergence equals, in view of the transition kernel (1),

$$\begin{aligned} \mathrm{KL}\Big(\mathbb{P}_\nu^{(Y_{t+1}, U_{t+1})\,|\,I_t}, \mathbb{P}_{\nu'}^{(Y_{t+1}, U_{t+1})\,|\,I_t}\Big) &= \mathbb{E}_\nu\Big[\mathbb{E}_\nu\Big[\mathrm{KL}\big(\nu_{\psi_t(I_t)} \otimes \lambda,\, \nu'_{\psi_t(I_t)} \otimes \lambda\big) \,\Big|\, I_t\Big]\Big] \\ &= \mathbb{E}_\nu\left[\sum_{a=1}^K \mathrm{KL}(\nu_a, \nu'_a)\, \mathbb{I}_{\{\psi_t(I_t)=a\}}\right]. \end{aligned}$$

Recalling that $A_{t+1} = \psi_t(I_t)$ and iterating the argument in (4) leads to the equality stated in (F).

2.2. Proof of the inequality in (F)

This is our key contribution to a simplified proof of the lower bound (2). It follows from the data-processing inequality (also known as contraction of entropy), i.e., the fact that Kullback-Leibler divergences between image distributions are smaller than the Kullback-Leibler divergences between the original distributions. (The data-processing inequality itself follows, e.g., from a log-sum inequality, i.e., Jensen's inequality applied to $t \mapsto t \ln t$.) Since this result of independent interest we state it in a slightly more general way.

Lemma 1 *Consider a measurable space $(\Gamma, \mathcal{G})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, and any $[0,1]$–valued and $\mathcal{G}$–measurable random variable Z. Then,*

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) \geqslant \mathrm{kl}\big(\mathbb{E}_1[Z], \mathbb{E}_2[Z]\big)\,.$$

Proof We augment the underlying measurable space into $\Gamma \times [0,1]$, where $[0,1]$ is equipped with the Borel σ–algebra and the Lebesgue measure λ, and consider a random variable V independent of Z, with uniform distribution over $[0,1]$. Introduce the event $E = \{Z \geqslant V\}$. By the consideration of product distributions for the first equality and by the data-processing inequality applied to $\mathbb{I}_E$ for the inequality, we have

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) = \mathrm{KL}\big(\mathbb{P}_1 \otimes \lambda,\, \mathbb{P}_2 \otimes \lambda\big) &\geqslant \mathrm{KL}\Big((\mathbb{P}_1 \otimes \lambda)^{\mathbb{I}_E},\, (\mathbb{P}_2 \otimes \lambda)^{\mathbb{I}_E}\Big) \\ &= \mathrm{kl}\big((\mathbb{P}_1 \otimes \lambda)(E),\, (\mathbb{P}_2 \otimes \lambda)(E)\big)\,.\end{aligned}$$

The last equality is by definition of kl as the Kullback-Leibler divergence between Bernoulli distributions. The proof is concluded by noting that for all j,

$$(\mathbb{P}_j \otimes \lambda)(E) = \mathbb{E}_j \otimes \lambda\big[\mathbb{I}_{\{Z \geqslant V\}}\big] = \mathbb{E}_j[Z]$$

by the Fubini-Tonelli theorem. ■

2.3. Application: re-deriving asymptotic distribution-dependent bounds

As a warm-up, we show how the asymptotic distribution-dependent lower bound (2) of Burnetas and Katehakis (1996) can be reobtained, for so-called consistent strategies.

Definition 2 *A strategy ψ is consistent if for all bandits problems ν, for all suboptimal arms a, i.e., for all arms a such that $\Delta_a > 0$, it satisfies $\mathbb{E}_\nu\big[N_a(T)\big] = o(T^\alpha)$ for all $0 < \alpha \leqslant 1$.*

Theorem 3 *For all models $\mathcal{D}$, for all consistent strategies, for all bandits problems ν, for all suboptimal arms a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Proof Given any bandit problem ν and any suboptimal arm a, we consider a modified problem ν' where a is the (unique) optimal arm: $\nu'_k = \nu_k$ for all $k \neq a$ and ν'_a is any distribution in $\mathcal{D}$ such that its expectation μ'_a satisfies $\mu'_a > \mu^\star$ (if such a distribution exists; see the end of the proof otherwise). We apply the fundamental inequality (F) with $Z = N_a(T)/T$. All Kullback-Leibler divergences in its left-hand side are null except the one for arm a, so that we get the lower bound

$$\begin{aligned}\mathbb{E}_\nu\big[N_a(T)\big]\,\mathrm{KL}(\nu_a, \nu'_a) &\geqslant \mathrm{kl}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T,\ \mathbb{E}_{\nu'}\big[N_a(T)\big]/T\Big) \\ &\geqslant \left(1 - \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{T}\right) \ln \frac{T}{T - \mathbb{E}_{\nu'}\big[N_a(T)\big]} - \ln 2\,, \qquad (5)\end{aligned}$$

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

$$\mathrm{kl}(p,q) = \underbrace{p \ln \frac{1}{q}}_{\geqslant 0} + (1-p)\ln\frac{1}{1-q} + \underbrace{\big(p\ln p + (1-p)\ln(1-p)\big)}_{\geqslant -\ln 2}\,.$$

The consistency of ψ together with the fact that all arms $k \neq a$ are suboptimal for ν' entails that

$$\forall\, 0 < \alpha \leqslant 1, \qquad 0 \leqslant T - \mathbb{E}_{\nu'}\big[N_a(T)\big] = \sum_{k \neq a} \mathbb{E}_{\nu'}\big[N_k(T)\big] = o(T^\alpha)\,;$$

in particular, $T - \mathbb{E}_{\nu'}\big[N_a(T)\big] \leqslant T^\alpha$ for T sufficiently large. Therefore, for all $0 < \alpha \leqslant 1$,

$$\liminf_{T\to\infty} \ \frac{1}{\ln T} \ln \frac{T}{T - \mathbb{E}_{\nu'}\big[N_a(T)\big]} \geqslant \liminf_{T\to\infty} \ \frac{1}{\ln T} \ln \frac{T}{T^\alpha} = (1-\alpha)\,.$$

In addition, the consistency of ψ and the suboptimality of a for the bandit problem ν ensure that $\mathbb{E}_\nu\big[N_a(T)\big]/T \to 0$. Substituting these two facts in (5) we proved

$$\liminf_{T\to\infty} \ \frac{\mathbb{E}_\nu\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)}\,.$$

By taking the supremum in the right-hand side over all distributions $\nu'_a \in \mathcal{D}$ with $\mu'_a > \mu^\star$, if at least one such distribution exists, we get the bound of the theorem. Otherwise, $\mathcal{K}_{\inf}(\nu_a, \mu^\star) = +\infty$ by a standard convention on the infimum of an empty set and the bound holds as well. ■

3. Non-asymptotic bounds for small values of T

We prove three such bounds with different merits and drawbacks. Basically, we expect suboptimal arms to be pulled each about T/K of the time when T is small; when T becomes larger, sufficient information was gained for identifying the best arm, and the logarithmic regime can take place.

The first bound shows that $\mathbb{E}_\nu\big[N_a(T)\big]$ is of order T/K as long as T is at most of order $1/\mathcal{K}_{\inf}(\nu_a, \mu^\star)$; we call it an absolute lower bound for a suboptimal arm a. Its drawback is that the times T for which it is valid are independent of the number of arms K, while (at least in some cases) one may expect the initial phase to last until $T \approx K/\mathcal{K}_{\inf}(\nu_a, \mu^\star)$.

The second lower bound thus addresses the dependency of the initial phase in K by considering a relative lower bound between a suboptimal arm a and an optimal arm $a^\star$. We prove that $\mathbb{E}_\nu\big[N_a(T)/N_{a^\star}(T)\big]$ is not much smaller than 1 whenever T is at most of order $K/\mathrm{KL}(\nu_a, \nu_{a^\star})$. Here, the number of arms K plays the expected effect on the length of the initial exploration phase, which should be proportional to K.

The third lower bound is a collective lower bound on all suboptimal arms, i.e., a lower bound on $\sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_\nu\big[N_a(T)\big]$ where $\mathcal{A}^\star(\nu)$ denotes the set of the $A^\star_\nu$ optimal arms of ν. It is of the desired order $T(1 - A^\star_\nu/K)$ for times T of the desired order $K/\mathcal{K}^{\max}_\nu$, where $\mathcal{K}^{\max}_\nu$ is some Kullback-Leibler divergence.

We prove these lower bounds under minimal assumptions on the considered strategies: some mild symmetry and the fact that for suboptimal arms a, the number of pulls $\mathbb{E}_\nu\big[N_a(T)\big]$ should decrease as μ_a decreases, all other distributions of arms being fixed.

3.1. Absolute lower bound for a suboptimal arm

The uniform strategy is the one that pulls an arm uniformly at random at each round.

Definition 4 *A strategy ψ is smarter than the uniform strategy if for all bandit problems ν, for all optimal arms $a^\star$, for all $T \geqslant 1$,*

$$\mathbb{E}_\nu\big[N_{a^\star}(T)\big] \geqslant \frac{T}{K}\,.$$

Theorem 5 *For all strategies ψ that are smarter than the uniform strategy, for all bandit problems ν, for all arms a, for all $T \geqslant 1$,*

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{K}\Big(1 - \sqrt{2T\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\Big)\,.$$

In particular,

$$\forall\, T \leqslant \frac{1}{8\mathcal{K}_{\inf}(\nu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{2K}\,.$$

Proof It suffices to consider suboptimal arms a. As in the proof of Theorem 3, we consider a modified bandit problem ν' with $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in \mathcal{D}$ such that $\mu'_a > \mu^\star$, if such a distribution ν'_a exists (otherwise, the first claimed lower bounds equals $-\infty$). From (F), we get

$$\mathbb{E}_\nu\big[N_a(T)\big]\,\mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T,\ \mathbb{E}_{\nu'}\big[N_a(T)\big]/T\Big)\,.$$

We may assume that $\mathbb{E}_\nu\big[N_a(T)\big]/T \leqslant 1/K$; otherwise, the first claimed bound holds. Since a is the optimal arm under ν' and since the considered strategy is smarter than the uniform strategy, $\mathbb{E}_{\nu'}\big[N_a(T)\big]/T \geqslant 1/K$. Using that $q \mapsto \mathrm{kl}(p,q)$ is increasing on $[p,1]$, we thus get

$$\mathrm{kl}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T,\ \mathbb{E}_{\nu'}\big[N_a(T)\big]/T\Big) \geqslant \mathrm{kl}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T,\ 1/K\Big)\,.$$

Lemmas 6 yields

$$\mathbb{E}_\nu\big[N_a(T)\big]\,\mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T,\ 1/K\Big) \geqslant \frac{K}{2}\Big(\mathbb{E}_\nu\big[N_a(T)\big]/T - 1/K\Big)^2,$$

from which follows, after substitution of the above assumption $\mathbb{E}_\nu\big[N_a(T)\big]/T \leqslant 1/K$ in the left-hand side,

$$\frac{\mathbb{E}_\nu\big[N_a(T)\big]}{T} \geqslant \frac{1}{K} - \sqrt{\frac{2T}{K^2}\mathrm{KL}(\nu_a, \nu'_a)}\,.$$

Taking the infimum over all possible ν'_a and rearranging concludes the proof. ■

The following lemma offers a local Pinsker's inequality; see also Cappé et al. (2013, Lemma 3 in Appendix A.2.1) for a more general version. Of course, the classical Pinsker's inequality,

$$\forall (p,q) \in [0,1]^2, \qquad \mathrm{kl}(p,q) \geqslant 2(p-q)^2\,, \tag{6}$$

is a consequence of the first inequality of this local version.

Lemma 6 *For $0 \leqslant p < q \leqslant 1$, we have* $\displaystyle \mathrm{kl}(p,q) \geqslant \frac{1}{2\max\limits_{x\in[p,q]} x(1-x)}(p-q)^2 \geqslant \frac{1}{2q}(p-q)^2\,.$

Proof We may assume that $p > 0$ and $q < 1$, since for $p = 0$, the result follows by continuity, and for $q = 1$, the inequality is void, as $\mathrm{kl}(p, 1) = +\infty$ when $p < 1$. The first and second derivative of kl equal

$$\frac{\partial}{\partial p}\mathrm{kl}(p,q) = \ln p - \ln(1-p) - \ln q + \ln(1-q) \quad \text{and} \quad \frac{\partial^2}{\partial^2 p}\mathrm{kl}(p,q) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}\,.$$

By Taylor's equality, there exists $r \in [p, q]$ such that

$$\mathrm{kl}(p,q) = \underbrace{kl(q,q)}_{=0} + (p-q)\,\underbrace{\frac{\partial}{\partial p}\mathrm{kl}(q,q)}_{=0} + \frac{(p-q)^2}{2}\,\underbrace{\frac{\partial^2}{\partial^2 p}\mathrm{kl}(r,q)}_{=1/\left(r(1-r)\right)}\,.$$

The proof of the first inequality is concluded by upper bounding $r(1-r)$ by $\max\limits_{x\in[p,q]} x(1-x)$. The second inequality follows from $\max\limits_{x\in[p,q]} x(1-x) \leqslant \max\limits_{x\in[p,q]} x \leqslant q$. ■

3.2. Relative lower bound

Our proof will be based on an assumption of symmetry.

Definition 7 *A strategy ψ is pairwise symmetric for optimal arms if for all bandit problems ν, for each pair of optimal arms $a^\star$ and $a_\star$, the equality $\nu_{a^\star} = \nu_{a_\star}$ entails that, for all $T \geqslant 1$,*

$$\big(N_{a^\star}(T),\, N_{a_\star}(T)\big) \qquad \textit{and} \qquad \big(N_{a_\star}(T),\, N_{a^\star}(T)\big)$$

have the same distribution.

Note that the required symmetry is extremely mild as only pairs of *optimal* arms with the *same* distribution are to be considered. What the equality of distributions means is that the strategy should be based only on payoffs and not on the values of the indexes of the arms.

Theorem 8 *For all strategies ψ that are pairwise symmetric for optimal arms, for all bandit problems ν, for all suboptimal arms a and all optimal arms $a^\star$, for all $T \geqslant 1$,*

$$\textit{either} \quad \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{T}{K} \qquad \textit{or} \quad \mathbb{E}_\nu\left[\frac{\max\big\{N_a(T),\,1\big\}}{\max\big\{N_{a^\star}(T),\,1\big\}}\right] \geqslant 1 - 2\sqrt{\frac{2T\,\mathrm{KL}(\nu_a,\nu_{a^\star})}{K}}\,.$$

Proof For all arms k, we denote by $N_k^+(T) = \max\big\{N_k(T),\,1\big\}$. Given a bandit problem ν and a suboptimal arm a, we form an alternative bandit problem ν' given by $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a = \nu_{a^\star}$, where $a^\star$ is an optimal arm of ν. In particular, arms a and $a^\star$ are both optimal arms under ν'. By the assumption of pairwise symmetry for optimal arms, we have in particular that

$$\mathbb{E}_{\nu'}\left[\frac{N_a^+(T)}{N_a^+(T)+N_{a^\star}^+(T)}\right] = \mathbb{E}_{\nu'}\left[\frac{N_{a^\star}^+(T)}{N_{a^\star}^+(T)+N_a^+(T)}\right] = \frac{1}{2}\,.$$

The latter equality and the fundamental inequality (F) yield in the present case, through the choice of $Z = N_a^+(T)/\big(N_a^+(T) + N_{a^\star}^+(T)\big)$,

$$\mathbb{E}_\nu\big[N_a(T)\big]\ \mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\left(\mathbb{E}_\nu\left[\frac{N_a^+(T)}{N_a^+(T) + N_{a^\star}^+(T)}\right],\ \frac{1}{2}\right). \tag{7}$$

The concavity of the function $x \mapsto x/(1+x)$ and Jensen's inequality show that

$$\mathbb{E}_\nu\left[\frac{N_a^+(T)}{N_a^+(T) + N_{a^\star}^+(T)}\right] = \mathbb{E}_\nu\left[\frac{N_a^+(T)/N_{a^\star}^+(T)}{1 + N_a^+(T)/N_{a^\star}^+(T)}\right] \leqslant \frac{\mathbb{E}_\nu\big[N_a^+(T)/N_{a^\star}^+(T)\big]}{1 + \mathbb{E}_\nu\big[N_a^+(T)/N_{a^\star}^+(T)\big]}\,.$$

We can assume that $\mathbb{E}_\nu\big[N_a^+(T)/N_{a^\star}^+(T)\big] \leqslant 1$, otherwise, the result of the theorem is obtained. In this case, the latter upper bound is smaller than $1/2$. Using in addition that $p \mapsto \mathrm{kl}(p, 1/2)$ is decreasing on $[0, 1/2]$, and assuming that $\mathbb{E}_\nu\big[N_a(T)\big] \leqslant T/K$ (otherwise, the result of the theorem is obtained as well), we get from (7)

$$\frac{T}{K}\ \mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_\nu\big[N_a^+(T)/N_{a^\star}^+(T)\big]}{1 + \mathbb{E}_\nu\big[N_a^+(T)/N_{a^\star}^+(T)\big]},\ \frac{1}{2}\right).$$

Pinsker's inequality (6) entails the inequality

$$\frac{T}{K}\ \mathrm{KL}(\nu_a, \nu'_a) \geqslant 2\left(\frac{1}{2} - \frac{r}{1+r}\right)^2 \qquad \text{where} \quad r = \mathbb{E}_\nu\left[\frac{N_a^+(T)}{N_{a^\star}^+(T)}\right].$$

In particular,

$$\frac{r}{1+r} \geqslant \frac{1}{2} - \sqrt{\frac{T\,\mathrm{KL}(\nu_a, \nu'_a)}{2K}}\,.$$

Applying the increasing function $x \mapsto x/(1-x)$ to both sides, we get

$$r \geqslant \frac{1 - \sqrt{2T\,\mathrm{KL}(\nu_a, \nu'_a)/K}}{1 + \sqrt{2T\,\mathrm{KL}(\nu_a, \nu'_a)/K}} \geqslant \left(1 - \sqrt{\frac{2T\,\mathrm{KL}(\nu_a, \nu'_a)}{K}}\right)^2,$$

where we used $1/(1+x) \geqslant 1 - x$ for the last inequality and where we assumed that T is small enough to ensure $1 - \sqrt{2T\,\mathrm{KL}(\nu_a, \nu'_a)/K} \geqslant 0$. Whether this condition is satisfied or not, we have the (possibly void) lower bound

$$r \geqslant 1 - 2\sqrt{\frac{2T\,\mathrm{KL}(\nu_a, \nu'_a)}{K}}\,.$$

The proof is concluded by noting that by definition $\nu'_a = \nu_{a^\star}$. ■

3.3. Collective lower bound

In this section, for any given bandit problem ν, we denote by $\mathcal{A}^\star(\nu)$ the set of its optimal arms and by $\mathcal{W}(\nu)$ the set of its worse arms, i.e., the ones associated with the distributions with the smaller expectation among all distributions for the arms. We also let $A_\nu^\star$ be the cardinality of $\mathcal{A}^\star(\nu)$.

We define the following partial order $\preccurlyeq$ on bandit problems: $\nu' \preccurlyeq \nu$ if

$$\forall a \in \mathcal{A}^\star(\nu), \quad \nu_a = \nu'_a \qquad \text{and} \qquad \forall a \notin \mathcal{A}^\star(\nu), \quad E(\nu'_a) \leqslant E(\nu_a)\,.$$

In particular, $\mathcal{A}^\star(\nu) = \mathcal{A}^\star(\nu')$ in this case. The definition models the fact that the bandit problem ν' should be easier than ν, as non-optimal arms in ν' are farther away from the optimal arms (in expectation) that in ν. Any reasonable strategy should perform better on ν' than on ν, which leads to the following definition, where we measure performance in the expected number of times optimal arms are pulled. (Recall that the sets of optimal arms are identical for ν and ν'.)

Definition 9 *A strategy ψ is monotonic if for all bandit problems $\nu' \preccurlyeq \nu$,*

$$\sum_{a^\star \in \mathcal{A}^\star(\nu')} \mathbb{E}_{\nu'}\big[N_{a^\star}(T)\big] \geqslant \sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\nu}\big[N_{a^\star}(T)\big]\,.$$

Theorem 10 *For all strategies ψ that are pairwise symmetric for optimal arms and monotonic, for all bandits problem ν,*

$$\sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_{\nu}\big[N_a(T)\big] \geqslant T\left(1 - \frac{A^\star_\nu}{K} - \frac{A^\star_\nu\sqrt{2T\,\mathcal{K}^{\max}_\nu}}{K} - \frac{2A^\star_\nu T\mathcal{K}^{\max}_\nu}{K}\right),$$

$$\text{where} \qquad \mathcal{K}^{\max}_\nu = \min_{w \in \mathcal{W}(\nu)} \max_{a^\star \in \mathcal{A}^\star(\nu)} \mathrm{KL}(\nu_{a_\star}, \nu_w)\,.$$

In particular, the regret is lower bounded according to

$$R_{\nu,T} \geqslant \left(\min_{a \notin \mathcal{A}^\star(\nu)} \Delta_a\right) T\left(1 - \frac{A^\star_\nu}{K} - \frac{A^\star_\nu\sqrt{2T\,\mathcal{K}^{\max}_\nu}}{K} - \frac{2A^\star_\nu T\mathcal{K}^{\max}_\nu}{K}\right).$$

Proof We denote by $\widetilde{w}$ the $w \in \mathcal{W}(\nu)$ achieving the minimum in the defining equation of $\mathcal{K}^{\max}_\nu$. We construct two bandit models from ν. First, the model $\underline{\nu}$ differs from ν only at suboptimal arms $a \notin \mathcal{A}^\star(\nu)$, which we associate with $\underline{\nu}_a = \nu_{\widetilde{w}}$. By construction, $\underline{\nu} \preccurlyeq \nu$. Second, the model $\underline{\underline{\nu}}$ in which each arm is associated with $\nu_{\widetilde{w}}$, i.e., $\underline{\underline{\nu}}_a = \nu_{\widetilde{w}}$ for all $a \in \{1, \dots, K\}$.

By monotonicity of ψ,

$$\sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_{\nu}\big[N_a(T)\big] \geqslant \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,.$$

We can therefore focus our attention, for the rest of the proof, on the $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]$. The strategy is also pairwise symmetric for optimal arms and all arms of $\underline{\underline{\nu}}$ are optimal. This implies in particular that $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_1(T)\big] = \mathbb{E}_{\underline{\underline{\nu}}}\big[N_a(T)\big]$ for all arms a, thus $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_a(T)\big] = T/K$ for all arms a.

Now, the bound (F) with $Z = \sum_{a^\star \in \mathcal{A}^\star(\nu)} N_{a^\star}(T)/T$ and the bandit models $\underline{\underline{\nu}}$ and $\underline{\nu}$ gives

$$\begin{aligned}\sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{a^\star}(T)\big]\mathrm{KL}(\nu_{a^\star}, \nu_{\widetilde{w}}) &\geqslant \mathrm{kl}\left(\sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\underline{\nu}}}\big[N_{a^\star}(T)\big]/T, \sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\nu}}\big[N_{a^\star}(T)\big]/T\right) \\ &= \mathrm{kl}\left(\frac{A^\star_\nu}{K}, \sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\nu}}\big[N_{a^\star}(T)\big]/T\right).\end{aligned}$$

By definition of $\mathcal{K}_\nu^{\max}$ and $\widetilde{w}$, and because $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] = T/K$, we have

$$\sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\nu}}\big[N_{a^\star}(T)\big] \mathrm{KL}(\nu_{a^\star}, \nu_{\widetilde{w}}) = \frac{T A_\nu^\star \mathcal{K}_\nu^{\max}}{K}\,,$$

which yields the inequality

$$\frac{T A_\nu^\star \mathcal{K}_\nu^{\max}}{K} \geqslant \mathrm{kl}\left(\frac{A_\nu^\star}{K},\, x\right) \qquad \text{where} \qquad x = \frac{1}{T} \sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\underline{\nu}}\big[N_{a^\star}(T)\big]\,.$$

We want to upper bound x, in order to get a lower bound on $1-x$. We assume that $x \geqslant A_\nu^\star/K$, otherwise, the bound (8) stated below is also satisfied. Pinsker's inequality (Lemma 6) then ensures that

$$\frac{T A_\nu^\star \mathcal{K}_\nu^{\max}}{K} \geqslant \frac{1}{2x}\left(\frac{A_\nu^\star}{K} - x\right)^2,$$

Lemma 11 below finally entails that

$$x \leqslant \frac{A_\nu^\star}{K}\left(1 + 2T\mathcal{K}_\nu^{\max} + \sqrt{2T\mathcal{K}_\nu^{\max}}\right). \tag{8}$$

The proof is concluded by putting all elements together thanks to the monotonicity of ψ and the definition of x:

$$\sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_\nu\big[N_a(T)\big] \geqslant \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] = T(1-x)\,.$$

■

Lemma 11 *If $x \in \mathbb{R}$ satisfies $(x-\alpha)^2 \leqslant \beta x$ for some $\alpha \geqslant 0$ and $\beta \geqslant 0$, then $x \leqslant \alpha + \beta + \sqrt{\alpha\beta}$.*

Proof By assumption, $x^2 - (2\alpha+\beta)x + \alpha^2 \leqslant 0$. We have that x is smaller than the larger root of the associated polynom, that is,

$$x \leqslant \frac{2\alpha + \beta + \sqrt{(2\alpha+\beta)^2 - 4\alpha^2}}{2} = \frac{2\alpha + \beta + \sqrt{4\alpha\beta + \beta^2}}{2}\,.$$

We conclude with $\sqrt{4\alpha\beta + \beta^2} \leqslant \sqrt{4\alpha\beta} + \sqrt{\beta^2}$. ■

4. Non-asymptotic bounds for large T

We restrict our attention to well-behaved models and super-consistent strategies.

Definition 12 *A model $\mathcal{D}$ is well behaved if there exists a function ω such that for all bandits problems ν, there exists $\varepsilon_0(\mu^\star)$ such that for all suboptimal arms a,*

$$\forall \varepsilon < \varepsilon_0(\mu^\star), \qquad \mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon) \leqslant \mathcal{K}_{\inf}(\nu_a, \mu^\star) + \varepsilon\, \omega(\nu_a, \mu^\star)\,.$$

We could have considered a more general definition, where the upper bound would have been any vanishing function of ε, not only a linear function of ε. However, all examples considered in this paper (see Section 4.2) can be associated with such a linear difference. Those examples of well-behaved models include parametric families like regular exponential families, as well as more massive classes, like the set of all distributions with bounded support (with or without a constraint on the finiteness of support). Some of these examples, namely, regular exponential families and finitely-supported distributions with common bounded support, were the models studied in Cappé et al. (2013) to get non-asymptotic upper bounds on the regret of the optimal order (2).

Definition 13 *A strategy ψ is super consistent on a model $\mathcal{D}$ if there exists a constant $C_{\psi,\mathcal{D}}$ such that for all bandits problems ν in $\mathcal{D}$, for all suboptimal arms a, for all $T \geqslant 2$,*

$$\mathbb{E}_\nu\big[N_a(T)\big] \leqslant C_{\psi,\mathcal{D}} \frac{\ln T}{\Delta_a^2} \,.$$

Super consistence is a refinement of the notion of consistence based on two considerations. First, that there exist such strategies, for instance, the UCB strategy of Auer et al. (2002a) on the model of all distributions with some common bounded support. Second, that together with Pinsker's inequality, which entails in particular that $\mathcal{K}_{\inf}(\nu_a, \mu) \geqslant 2\Delta_a^2$, the bound stated in the definition of super consistency is still weaker than the aim (2).

4.1. A general non-asymptotic lower bound

Throughout this subsection, we fix a strategy ψ that is super consistent with respect to a model $\mathcal{D}$. We recall that we denote by $\mathcal{A}^\star(\nu)$ the set of optimal arms of the bandit problem ν and let $A_\nu^\star$ be its cardinality. We adapt the bounds (F) and (5) by using this time

$$Z = \frac{1}{T} \sum_{a^\star \in \mathcal{A}^\star(\nu)} N_{a^\star}(T)$$

and $\mathrm{kl}(p,q) \geqslant p \ln(1/q) - \ln 2$. For all bandit problems ν' that only differ from ν as far a suboptimal arm a is concerned, whose distribution of payoffs $\nu'_a \in \mathcal{D}$ is such that $\mu'_a > \mu^\star$, we get

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)} \left(\mathbb{E}_\nu[Z] \, \ln \frac{1}{\mathbb{E}_{\nu'}[Z]} - \ln 2 \right). \tag{9}$$

We restrict our attention to distributions $\nu'_a \in \mathcal{D}$ such that the gaps for ν' associated with optimal arms $a^\star \in \mathcal{A}^\star(\nu)$ of ν satisfy $\underline{\Delta} = \mu'_a - \mu^\star \geqslant \varepsilon$, for some parameter $\varepsilon > 0$ to be defined by the analysis. By super consistency, on the one hand,

$$\mathbb{E}_\nu[Z] = 1 - \frac{1}{T} \sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_\nu\big[N_a(T)\big] \geqslant 1 - \frac{1}{T} \left(C_{\psi,\mathcal{D}} \sum_{a \notin \mathcal{A}^\star(\nu)} \frac{1}{\Delta_a^2} \ln T \right);$$

on the other hand,

$$\mathbb{E}_{\nu'}[Z] = \frac{1}{T} \sum_{a^\star \in \mathcal{A}^\star(\nu)} \mathbb{E}_{\nu'}\big[N_a(T)\big] \leqslant \frac{A_\nu^\star \, C_{\psi,\mathcal{D}}}{\underline{\Delta}^2} \frac{\ln T}{T} \,.$$

Denoting

$$H(\nu) = \sum_{a \notin \mathcal{A}^\star(\nu)} \frac{1}{\Delta_a^2} \tag{10}$$

and using that $\underline{\Delta} \geqslant \varepsilon$, a substitution of the two super-consistency inequalities into (9) and an optimization over the considered distributions ν'_a leads to

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon)} \left(1 - C_{\psi,\mathcal{D}} H(\nu) \frac{\ln T}{T}\right) \ln \frac{T\varepsilon^2}{A_\nu^\star \, C_{\psi,\mathcal{D}} \ln T} - \frac{\ln 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon)} \,. \tag{11}$$

The obtained bound holds for all $T \geqslant 2$ (as in the definition of super consistency); however, for small values of T, it might be negative, thus useless.

To proceed, we use the fact that the model $\mathcal{D}$ is well-behaved to relate $\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon)$ to $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$. Since $1/(1+x) \geqslant 1 - x$ for all $x \geqslant 0$, we get by Definition 12

$$\forall \varepsilon < \varepsilon_0(\mu^\star), \qquad \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon)} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \left(1 - \varepsilon \frac{\omega(\nu_a, \mu^\star)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\right).$$

Now, we set $\varepsilon = \varepsilon_T = (\ln T)^{-4}$. Many other choices would have been possible, but this one is such that $\varepsilon_T \leqslant 0.0005$ already for $T \geqslant 1\,000$. Putting all things together, from (11), from the fact that $(1-a)(1-b)(1-c) \geqslant 1 - (a+b+c)$ when $0 \leqslant a, b, c \leqslant 1$, and from the bound $A_\nu^\star \leqslant K$, we get the following theorem.

Theorem 14 *For all super-consistent strategies ψ on well-behaved models $\mathcal{D}$, for all bandit problems ν in $\mathcal{D}$, for all suboptimal arms a,*

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - (a_T + b_T + c_T) \ln T - \frac{\ln 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \,, \tag{12}$$

for all $T \geqslant 2$ large enough so that

$$a_T = \frac{\omega(\nu_a, \mu^\star)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} (\ln T)^{-4} \,, \qquad b_T = C_{\psi,\mathcal{D}} H(\nu) \frac{\ln T}{T} \,, \qquad c_T = \frac{\ln\big(K \, C_{\psi,\mathcal{D}} (\ln T)^9\big)}{\ln T} \,,$$

are all smaller than 1*, where $H(\nu)$ was defined in* (10)*.*

Remark 15 *We have $(a_T + b_T + c_T) \ln T = \mathrm{O}\big(\ln(\ln T)\big)$. The non-asymptotic bound* (12) *is therefore of the form*

$$\mathbb{E}_\nu\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - \mathrm{O}\big(\ln(\ln T)\big) \,.$$

4.2. Two examples (and a half) of well-behaved models

We consider first distributions with common bounded support (and the subclass of such distributions with finite support); and then, regular exponential families. The latter and the subclass of distributions with finite and bounded support are the two models for which Cappé et al. (2013) could prove non-aymptotic upper bounds matching the lower bound (2).

Distributions with common bounded support. We denote by $\mathcal{M}\big([0,M]\big)$ the set of all probability distributions over $[0,M]$, equipped with its Borel σ–algebra, and restrict our model to such distributions with expectation not equal to M.

Lemma 16 *In the model* $\mathcal{D} = \Big\{ \mathrm{m} \in \mathcal{M}\big([0,M]\big) : E(\mathrm{m}) < M \Big\}$*, we have*

$$\forall \mathrm{m} \in \mathcal{D}, \quad \forall \mu^\star \in [0,M), \quad \forall \varepsilon \in \big(0, (M-\mu^\star)/2\big), \\ \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star + \varepsilon) \leqslant \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star) - \ln\left(1 - \frac{2\varepsilon}{M-\mu^\star}\right).$$

In particular, for all $\mathrm{m} \in \mathcal{D}$ *and* $\mu^\star \in [0,M)$*,*

$$\forall \varepsilon \in \big(0, (M-\mu^\star)/4\big), \qquad \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star + \varepsilon) \leqslant \mathcal{K}_{\inf}(\mathrm{m}, \mu^\star) + \frac{4\varepsilon}{M-\mu^\star}\,.$$

Proof We fix m, $\mu^\star$ and ε as indicated for the first bound; in particular, $\mu^\star + \varepsilon < M$. Since m is a probability distribution, it has at most countably many atoms; therefore, there exists some $x \in (\mu^\star + \varepsilon, M)$ such that $\mathrm{m}(\{x\}) = 0$ and $x \geqslant (M + \mu^\star)/2$. In particular, m and the Dirac measure δ_x at this point are singular measures.

We consider some $\mathrm{m}' \in \mathcal{D}$ such that $E(\mathrm{m}') > \mu^\star$ and $\mathrm{m} \ll \mathrm{m}'$ (i.e., m is absolutely continuous with respect to m'). Such distributions exist and they are the only interesting ones in the defining infimum of $\mathcal{K}_{\inf}(\mathrm{m}, \mu^\star)$. We associate with m' the distribution

$$\mathrm{m}'_\alpha = (1-\alpha)\mathrm{m}' + \alpha\delta_x \qquad \text{for the value} \qquad \alpha = \frac{\varepsilon}{x - \mu^\star} \in (0,1)\,.$$

The expectation of m'_α satisfies

$$E\big(\mathrm{m}'_\alpha\big) > (1-\alpha)\mu^\star + \alpha x = \mu^\star + \alpha(x - \mu^\star) = \mu^\star + \varepsilon\,. \tag{13}$$

Now, $\mathrm{m} \ll \mathrm{m}'$ entails that $\mathrm{m} \ll \mathrm{m}'_\alpha$ as well, with respective densities satisfying (because m and δ_x are singular)

$$\frac{\mathrm{dm}}{\mathrm{dm}'_\alpha} = \frac{1}{1-\alpha}\,\frac{\mathrm{dm}}{\mathrm{dm}'} \qquad \text{and} \qquad \frac{\mathrm{dm}}{\mathrm{dm}'_\alpha}(x) = 0\,.$$

Therefore,

$$\mathrm{KL}(\mathrm{m}, \mathrm{m}'_\alpha) = \int \left(\ln \frac{\mathrm{dm}}{\mathrm{dm}'_\alpha}\right) \mathrm{dm} = \ln \frac{1}{1-\alpha} + \int \left(\ln \frac{\mathrm{dm}}{\mathrm{dm}'}\right) \mathrm{dm} = \ln \frac{1}{1-\alpha} + \mathrm{KL}(\mathrm{m}, \mathrm{m}')\,.$$

Since α decreases with x and $x \geqslant (M + \mu^\star)/2$, we get $\alpha \leqslant 2\varepsilon/(M - \mu^\star)$. We substitute this bound in the inequality above and take the infimum in both sides, considering (13), to get the first claimed bound. The second bound follows from the inequality $-\ln(1-x) \leqslant 2x$ for $x \in [0, 1/2]$. ■

Remark 17 *We denote by* $\mathcal{M}_{\mathrm{fin}}\big([0,M]\big)$ *the subset of* $\mathcal{M}\big([0,M]\big)$ *formed by probability distributions with finite support. The proof above shows that the bound of Lemma 16 also holds for the model*

$$\mathcal{D} = \Big\{ \mathrm{m} \in \mathcal{M}_{\mathrm{fin}}\big([0,M]\big) : E(\mathrm{m}) < M \Big\}\,.$$

Regular exponential families. Another example of well-behaved models is given by regular exponential families, see Lehmann and Casella (1998) for a thorough exposition or Cappé et al. (2013) for an alternative exposition focused on multi-armed bandit problems.

Such a family $\mathcal{D}$ is indexed by an open set $I = (m, M)$, where for each $\mu \in I$ there exists a unique distribution $\nu_\mu \in \mathcal{D}$ with expectation μ. (The bounds m and M can be equal to $\pm\infty$.) A key property of such a family is that the Kullback-Leibler divergence between two of its elements can be represented[1] by a twice differentiable and strictly convex function $g : I \to \mathbb{R}$, with increasing first derivative $\dot{g}$ and continuous second derivative $\ddot{g} \geqslant 0$, in the sense that

$$\forall\, (\mu, \mu') \in I^2, \qquad \mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = g(\mu) - g(\mu') - (\mu - \mu')\, \dot{g}(\mu')\,. \tag{14}$$

In particular, $\mu' \mapsto \mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big)$ is strictly convex on I, thus is increasing on $[\mu, M)$. This entails that

$$\forall\, (\mu, \mu^\star) \in I^2 \ \text{ s.t } \ \mu < \mu^\star, \qquad \mathcal{K}_{\inf}(\nu_\mu, \mu^\star) = \mathrm{KL}\big(\nu_\mu, \nu_{\mu^\star}\big)\,. \tag{15}$$

In the lemma below, we restrict our attention to $\varepsilon > 0$ such that $\mu^\star + \varepsilon \in I$, e.g., to $\varepsilon < B_{\mu^\star}$ where

$$B_{\mu^\star} = \min\left\{ \frac{M - \mu^\star}{2},\, 1 \right\}. \tag{16}$$

Lemma 18 *In a model $\mathcal{D}$ given by a regular exponential family indexed by $I = (m, M)$ and whose Kullback-Leibler divergence (14) is represented by a function g, we have, with the notation (16),*

$$\forall\, (\mu, \mu^\star) \in I^2, \quad \forall\, 0 < \varepsilon < B_{\mu^\star}, \qquad \mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon) \leqslant \mathcal{K}_{\inf}(\nu_\mu, \mu^\star) + \varepsilon \left(\mu^\star + B_{\mu^\star} - \mu\right) G_{\mu^\star}$$

where $G_{\mu^\star} = \max\big\{\ddot{g}(x) : \ \mu^\star \leqslant x \leqslant \mu^\star + B_{\mu^\star}\big\}$.

Proof We may assume that $\mu < \mu^\star$, otherwise $\mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon) = \mathcal{K}_{\inf}(\nu_\mu, \mu^\star) = 0$ and the stated bound holds. When $\mu < \mu^\star$, we get by (14) and (15)

$$\begin{aligned}
&\mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon) - \mathcal{K}_{\inf}(\nu_\mu, \mu^\star) \\
&\quad = \ g(\mu^\star) - g(\mu^\star + \varepsilon) - \big(\mu - (\mu^\star + \varepsilon)\big)\, \dot{g}(\mu^\star + \varepsilon) + (\mu - \mu^\star)\, \dot{g}(\mu^\star) \\
&\quad = \ \underbrace{g(\mu^\star) - g(\mu^\star + \varepsilon) + \varepsilon\, \dot{g}(\mu^\star)}_{\leqslant 0} + \big((\mu^\star + \varepsilon) - \mu\big)\big(\dot{g}(\mu^\star + \varepsilon) - \dot{g}(\mu^\star)\big)\,,
\end{aligned}$$

where the inequality is obtained by convexity of g. The proof is concluded by an application of the mean-value theorem,

$$\dot{g}(\mu^\star + \varepsilon) - \dot{g}(\mu^\star) \leqslant \varepsilon \max_{(\mu^\star, \mu^\star + \varepsilon)} \ddot{g}\,,$$

and the bound $\varepsilon \leqslant B_{\mu^\star}$. ∎

The upper bound obtained on $\mathcal{K}_{\inf}(\nu_\mu, \mu^\star + \varepsilon) - \mathcal{K}_{\inf}(\nu_\mu, \mu^\star)$ equals $\varepsilon \left(\mu^\star + B_{\mu^\star} - \mu\right) G_{\mu^\star}$. The examples below propose concrete upper bounds for $G_{\mu^\star}$ in different exponential families. None of these upper bounds involves $B_{\mu^\star}$ as various monotonicity arguments can be invoked.

1. This function g has an intrinsic definition as the convex conjugate of the log-normalization function b in the natural parameter space Θ, where b can also be seen as a primitive of the expectation function $\Theta \to I$. But these properties are unimportant here.

Example 1 *For Poisson distributions, we have* $I = (0, +\infty)$ *and*

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \mu' - \mu + \mu \ln \frac{\mu}{\mu'} \,.$$

We may take $g(\mu) = \mu \ln \mu - \mu$*, so that* $\ddot{g}(\mu) = 1/\mu$ *and* $G_{\mu^\star} = 1/\mu^\star$*.*

Example 2 *For Gamma distributions with known shape parameter* $\alpha > 0$ *(e.g., the exponential distributions when* $\alpha = 1$*), we have* $I = (0, +\infty)$ *and*

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \alpha \left(\frac{\mu}{\mu'} - 1 - \ln \frac{\mu}{\mu'} \right) .$$

We may take $g(\mu) = -\alpha \ln \mu$*, so that* $\ddot{g}(\mu) = \alpha/\mu^2$ *and* $G_{\mu^\star} = \alpha/(\mu^\star)^2$*.*

Example 3 *For Gaussian distributions with known variance* $\sigma^2 > 0$*, we have* $I = (0, +\infty)$ *and*

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \frac{(\mu - \mu')^2}{2\sigma^2} \,.$$

We may take $g(\mu) = \mu^2/(2\sigma^2)$*, so that* $\ddot{g}(\mu) = 1/\sigma^2$ *and* $G_{\mu^\star} = 1/\sigma^2$*.*

Example 4 *For binomial distributions for* n *samples (e.g., Bernoulli distributions when* $n = 1$*), we have* $I = (0, n)$ *and*

$$\mathrm{KL}\big(\nu_\mu, \nu_{\mu'}\big) = \mu \ln \frac{\mu}{\mu'} + (n - \mu) \ln \frac{n - \mu}{n - \mu'} \,.$$

We may take $g(\mu) = \mu \ln \mu + (n - \mu) \ln(n - \mu)$*, so that* $\ddot{g}(\mu) = n/\big(\mu(n - \mu)\big)$*. A possible upper bound is*

$$G_{\mu^\star} \leqslant \frac{2}{\max\{\mu^\star, n - \mu^\star\}} \,.$$

This can be seen by noting that $B_{\mu^\star} \leqslant (n - \mu^\star)/2$ *so that any* $\mu \in [\mu^\star,\, \mu^\star + B_{\mu^\star}]$ *is such that* $\mu \geqslant \mu^\star$ *and* $n - \mu \geqslant n - \mu^\star - B_{\mu^\star} \geqslant (n - \mu^\star)/2$*.*

Acknowledgments

This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program while Gilles Stoltz visited Toulouse in November 2015. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grants ANR-13-BS01-0005 (project SPADRO) and ANR-13-CORD-0020 (project ALICIA). Gilles Stoltz would like to thank Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) for financial support.

Finally, the authors thank Sébastien Gerchinovitz for stimulating discussions and comments.

References

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

S. Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, France, 2010.

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

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

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.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

R. Combes and A. Proutière. Unimodal bandits without smoothness, 2014. arXiv:1406.7447.

E. Kaufmann, O. Capp, and A. Garivier. On the complexity of best arm identication in multi-armed bandit models. *Journal of Machine Learning Research*, 2016. To appear.

S. Kulkarni and G. Lugosi. Minimax lower bounds for the two-armed bandit problem. *IEEE Transactions on Automatic Control*, 45:711–714, 2000.

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

E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.

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.

Y. Wu, A. György, and C. Szepesvari. Online learning with Gaussian payoffs and side observations. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1360–1368. Curran Associates, Inc., 2015.