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

Aurélien Garivier, Pierre Ménard, Gilles Stoltz

MATHEMATICS OF OPERATIONS RESEARCH
DOI: 10.1287/moor.2017.0928

INFORMS
DOI 10.1287/moor.2017.0928

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

Aurélien Garivier
IMT: Université Paul Sabatier – CNRS, Toulouse, France, aurelien.garivier@math.univ-toulouse.fr, http://www.math.univ-toulouse.fr/ agarivie/

Pierre Ménard
IMT: Université Paul Sabatier – CNRS, Toulouse, France

Gilles Stoltz
GREGHEC: HEC Paris – CNRS, Jouy-en-Josas, France, stoltz@hec.fr, http://stoltz.perso.math.cnrs.fr

We revisit lower bounds on the regret in the case of multi-armed bandit problems. We obtain non-asymptotic, distribution-dependent bounds and provide simple proofs based only on well-known properties of Kullback-Leibler divergences. These bounds show in particular that in the 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 information-theoretic arguments used and they involve no unnecessary complications.

Key words: Multi-armed bandits, cumulative regret, information-theoretic proof techniques, non-asymptotic lower bounds
MSC2000 subject classification: Primary: 68T05, 62L10; Secondary: 62B10, 94A17, 94A20
OR/MS subject classification: Primary: computer science: artificial intelligence (learning and adaptive systems); statistics: sequential methods (sequential analysis); secondary: statistics: sufficiency and information (information-theoretic topics); information and communication, circuits: communication, information (measures of information, entropy; sampling theory)
History: Submitted June 10, 2016; revised March 3, 2017; accepted December 8, 2017

---

1. Introduction. After the works of Lai and Robbins [21] and Burnetas and Katehakis [9], 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 clearly 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 keep 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; see also Garivier et al. [16]).

First contribution: a folk result made rigorous. It seems to be a folk result (or at least, a widely believed result) that the regret should be linear in an initial phase of a bandit problem. However, all references that we were pointed out exhibit such a linear behavior only for limited bandit settings; we discuss them below, in the section about literature review. We are the first to provide linear distribution-dependent lower bounds for small horizons that hold for general bandit

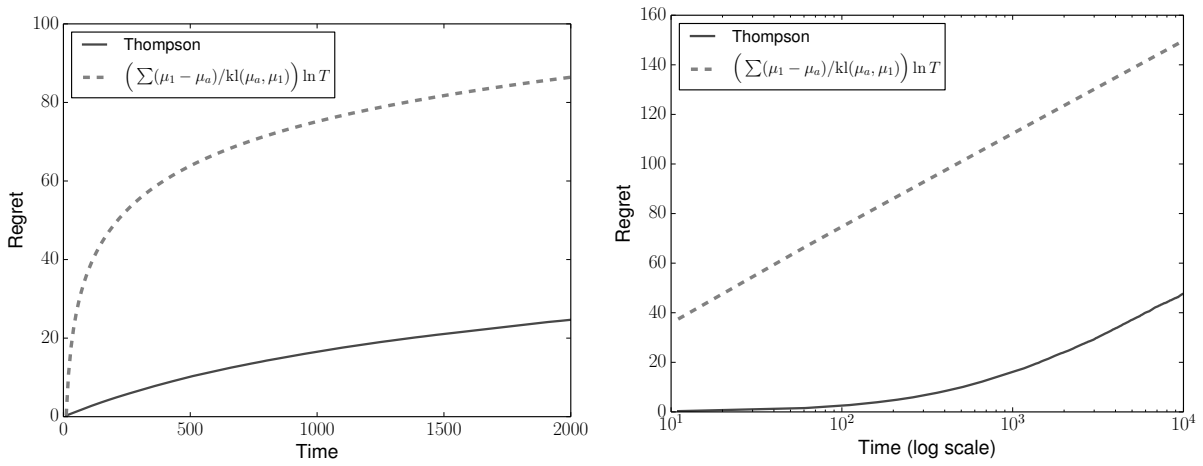

FIGURE 1. Expected regret of Thompson [24] Sampling (*blue, solid* line) on a Bernoulli bandit problem with parameters $(\mu_a)_{1\leqslant a\leqslant 6} = (0.05, 0.04, 0.02, 0.015, 0.01, 0.005)$; expectations are approximated over 500 runs.

Versus the Lai and Robbins [21] lower bound (*red, dotted* line) for a Bernoulli model; here kl denotes the Kullback-Leibler divergence (5) 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.

problems, with no restriction on the shape or on the expectations of the distributions over the arms.

Thus we may draw a more precise picture of the behavior of the regret in any bandit problem. Indeed, our bounds show 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.

Second contribution: a generic tool for proving distribution-dependent bandit lower bounds. On the technical side, we provide simple proofs, based on the fundamental information-theoretic inequality (6) stated in Section 2, which generalizes and simplifies previous approaches based on explicit changes of measures. In particular, we are able to re-derive the asymptotic distribution-dependent lower bounds of Lai and Robbins [21], Burnetas and Katehakis [9] and Cowan and Katehakis [14] in a few lines. This may perhaps be one of the most striking contributions of this paper. As a final set of results, we offer non-asymptotic versions of these lower bounds for large horizons, and exhibit the optimal order of magnitude of the second-order term in the regret bound, namely, $-\ln(\ln T)$.

The proof techniques come to the essence of the arguments used so far in the literature and they involve no unnecessary complications; they only rely on well-known properties of Kullback-Leibler divergences.

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 $\mathbb{R}$. We assume that each ν_a has a well-defined expectation and call $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ a bandit problem.

At each round $t \geqslant 1$, the player pulls the arm A_t and gets a real-valued reward Y_t drawn independently 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 $U_0, U_1, U_2, \ldots$ of independent and identically distributed random variables, with common distribution the uniform distribution over $[0,1]$. Formally, a strategy is a sequence $\psi = (\psi_t)_{t\geqslant 0}$ of measurable functions, each of which associates with the said past information, namely,

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

an arm $\psi_t(I_t) = A_{t+1} \in \{1,\ldots,K\}$, where $t \geqslant 0$. The initial information reduces to $I_0 = U_0$ and the first arm is $A_1 = \psi_0(U_0)$. The auxiliary randomization is conditionally independent of the sequence of rewards in the following sense: for $t \geqslant 1$, the randomization U_t used to pick A_{t+1} is independent of I_{t-1} and Y_t.

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

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

The number of times an arm a is pulled until round T by a strategy ψ is referred to as

$$N_{\psi,a}(T) = \sum_{t=1}^{T} \mathbb{I}_{\{A_t=a\}} = \sum_{t=1}^{T} \mathbb{I}_{\{\psi_{t-1}(I_{t-1})=a\}}\,.$$

The expected regret of a strategy ψ equals, by the tower rule (see details below),

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

In the equation above, the notation $\mathbb{E}_{\underline{\nu}}$ refers to the expectation associated with the bandit problem $\underline{\nu} = (\nu_a)_{a=1,\ldots,K}$; it is made formal in Section 2.

To show (1), we use that by the definition of the bandit setting, the distribution of the obtained payoff Y_t only depends on the chosen arm A_t and is independent from the past random draws of the $Y_1, \ldots, Y_{t-1}$. More precisely, conditionally on A_t, the distribution of Y_t is ν_{A_t} so that

$$\mathbb{E}_{\underline{\nu}}\big[Y_t \,|\, A_t\big] = \mu_{A_t}\,, \qquad \text{thus} \qquad \mathbb{E}_{\underline{\nu}}[Y_t] = \mathbb{E}_{\underline{\nu}}\Big[\mathbb{E}_{\underline{\nu}}\big[Y_t \,|\, A_t\big]\Big] = \mathbb{E}_{\underline{\nu}}\big[\mu_{A_t}\big]\,,$$

where we used the tower rule for the second set of equalities.

1.2. The general asymptotic lower bound: a quick literature review. We consider a bandit model $\mathcal{D}$, i.e., a collection of possible distributions ν_a associated with the arms. (That is, $\mathcal{D}$ is a subset of the set of all possible distributions over $\mathbb{R}$ with an expectation.) Lai and Robbins [21] and later Burnetas and Katehakis [9] exhibited asymptotic lower bounds and matching asymptotic upper bounds on the normalized regret $R_{\psi,\underline{\nu},T}/\ln T$, respectively in a one-parameter case and in a more general, multi-dimensional parameter case, under mild conditions on $\mathcal{D}$. We believe that the extension of these bounds to any, even non-parametric, model was a known or at least conjectured result (see, for instance, the introduction of Cappé et al. [11]). It turns out that recently, Cowan and Katehakis [14] provided a clear non-parametric statement, though under additional mild conditions on the model $\mathcal{D}$, which, as we will see, are not needed.

In the sequel, we denote by KL the Kullback-Leibler divergence between two probability distributions. We also recall that we denoted by E the expectation operator (that associates with each distribution its expectation).

To state the bound for the case of an arbitrary model $\mathcal{D}$, we will use the following key quantity $\mathcal{K}_{\inf}$ introduced by Burnetas and Katehakis [9, quantity (3)–(b) on page 125].

The key quantity $\mathcal{K}_{\inf}$**.** For any given $\nu_a \in \mathcal{D}$ and any real number x,

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

by convention, the infimum of the empty set equals $+\infty$. When the considered strategy is uniformly fast convergent in the sense of Definition 1 (stated later in this paper), then, for any suboptimal arm a,

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

Note that by the convention on the infimum of the empty set, this lower bound is void as soon as there exists no $\nu'_a \in \mathcal{D}$ such that $E(\nu'_a) > \mu^\star$.

Previous partial simplifications of the proof of (2)**.** We re-derive the above bound in a few lines in Section B.1.

There had been recent attempts to clarify the exposition of the proof of this lower bound, together with the desire of dropping the mild conditions that were still needed so far on the model $\mathcal{D}$. We first mention that Cowan and Katehakis [14] provided a more general and streamlined approach than the original expositions by Lai and Robbins [21] and Burnetas and Katehakis [9].

The case of Bernoulli models was discussed in Bubeck [5] and Bubeck and Cesa-Bianchi [6]. Only assumptions of uniform fast convergence of the strategies are required (see Definition 1) and the associated proof follows the original proof technique, by performing first an explicit change of measure and then applying some Markov–Chernoff bounding. More recently, Jiang [18, Section 2.2] presented a proof (only in the Bernoulli case) not relying on any explicit change of measure but with many additional technicalities with respect to our exposition, including some Markov bounding of well-chosen events. We have been referred to this PhD dissertation only recently, after completing the present paper.

As far as general bandit models are concerned, we may cite Kaufmann et al. [19, Appendix B]: they deal with the case of any model $\mathcal{D}$ but with the restriction that only bandit problems $\underline{\nu}$ with a unique optimal arm should be considered. They still use both an explicit change of measure –to prove the chain-rule equality in (6)– and then apply as well some Markov–Chernoff bounding to the probability of well-chosen events. With a different aim, Combes and Proutière [13] presented similar arguments.

We also wish to mention the contribution of Wu et al. [25], though their focus and aim are radically different. With respect to some aspects, their setting and goal is wider or more general: they developed non-asymptotic problem-dependent lower bounds on the regret of any algorithm, in the case of more general limited feedback models than just the simplest case of multi-armed bandit problems. Their lower bounds can recover the asymptotic bounds of Burnetas and Katehakis [9], but only up to a constant factor as they acknowledge in their contribution. 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 need some assumptions on the strategies –extremely mild ones, though: some minimal symmetry– and do not need their regret to be bounded from above. However, the main difference with respect to this reference is that its focus is limited to specific bandit models, namely Gaussian bandits models, while Burnetas and Katehakis [9] and the present paper do not impose such a restriction on the bandit model.

1.3. Other bandit lower bounds: a brief literature review. In this paper, we are mostly interested in general distribution-dependent lower bounds, that hold for all bandit problems, just like (2). We do target generality. This is in contrast with many earlier lower bounds in the multi-armed bandit setting, which are rather of the following form, which we will refer to as (well-chosen):

"*There exists some* well-chosen*, difficult bandit problem such that all strategies suffer a regret larger than* [...]." (well-chosen)

Specific examples and pointers for this kind of bounds are given below. An interesting variation is provided by Mannor and Tsitsiklis [23, Theorem 10], who state that for all strategies, there exists some well-chosen, difficult Bernoulli bandit problem such that the regret is linear at first and then, logarithmic.

On the contrary, we will issue statements of the following form, which we will refer to as (all):

"*For* all *bandit problems, all (reasonable) strategies suffer a regret larger than* (...)." (all)

Sometimes, but not always, we will have to impose some mild restrictions on the considered strategies (like some minimal symmetry, or some notion of uniform fast convergence); this is what we mean by requiring the strategies to be "reasonable".

We discuss briefly below two other sets of regret lower bounds. We are pleased to mention that our fundamental inequality was already used in at least one subsequent article, namely by Garivier et al. [16], to prove in a few lines matching lower bounds for a refined analysis of explore-then-commit strategies.

The distribution-free lower bound. This inequality states that for the model $\mathcal{D} = \mathcal{M}\big([0,1]\big)$ of all probability distributions over $[0,1]$, for all strategies ψ, for all $T \geqslant 1$ and all $K \geqslant 2$,

$$\sup_{\underline{\nu}} R_{\psi,\underline{\nu},T} \geqslant \frac{1}{20} \min\Big\{\sqrt{KT}, T\Big\}; \tag{3}$$

see Auer et al. [3], Cesa-Bianchi and Lugosi [12], and for two-armed bandits, Kulkarni and Lugosi [20]. We re-derive the above bound in Section B.1 of the appendix. This re-derivation follows the very same proof scheme as in the original proof; the only difference is that some steps (e.g., the use of chain-rule equality for Kullback-Leibler divergences) are implemented separately as parts of the proof of our general inequality (6). In particular, the well-chosen difficult bandit problems used to prove this bound are composed of Bernoulli distributions with parameters $1/2$ and $1/2+\varepsilon$, where ε is carefully tuned according to the values of T and K. This bound therefore rather falls under the umbrella (well-chosen).

Lower bounds for sub-Gaussian bandit problems in the case when $\mu^\star$ or the gaps Δ are known. This framework and the exploitation of this knowledge was first studied by Bubeck et al. [7]. They consider a bandit model $\mathcal{D}$ containing only sub-Gaussian distributions with parameter $\sigma^2 \leqslant 1$; that is, distributions ν_a, with expectations $\mu_a \in \mathbb{R}$, such that

$$\forall \lambda \in \mathbb{R}, \qquad \int_{\mathbb{R}} \exp\big(\lambda(y-\mu_a)\big)\, \mathrm{d}\nu_a(y) \leqslant \exp\left(\frac{\lambda^2}{2}\right). \tag{4}$$

Examples of such distributions include Gaussian distributions with variance smaller than 1 and bounded distributions with range smaller than 2.

They study how much smaller the regret bounds can get when either the maximal expected payoff $\mu^\star$ or the gaps Δ_a are known. For the case when the gaps Δ_a are known but not $\mu^\star$, they exhibit a lower bound on the regret matching previously known upper bounds, thus proving their optimality. For the case when $\mu^\star$ is known but not the gaps, they offer an algorithm and its associated regret upper bound, as well as a framework for deriving a lower bound; later work (see Bubeck et al. [8] and Faure et al. [15]) point out that a bounded regret can be achieved in this case.

We (re-)derive these two lower bounds in a few lines in Section B.2 of the appendix. In particular, the well-chosen difficult bandit problems used are composed of Gaussian distributions $\mathcal{N}(\mu_a, 1)$, with expectations $\mu_a \in \{-\Delta, 0, \Delta\}$. Only statements of the form (well-chosen), not of the form (all), are obtained. Put differently, no general distribution-dependent statement like: "For all bandit problems in which the gaps Δ (or the maximal expected payoff $\mu^\star$) are known, all (reasonable) strategies suffer a regret larger than [...]" is proposed by Bubeck et al. [7]; only well-chosen, difficult bandit problems are considered. This is in strong contrast with our general distribution-dependent bounds for the initial linear regime, provided in Section 3.

1.4. Outline of our contributions. In Section 2, we present Inequality (6), 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 re-derivation of the earlier asymptotic lower bounds by Lai and Robbins [21], Burnetas and Katehakis [9] and Cowan and Katehakis [14]. Some other earlier bounds are also re-derived in Appendix B, namely, the distribution-free lower bound by Auer et al. [3] as well as the bounded-regret Gaussian lower bounds by Bubeck et al. [7] in the case when $\mu^\star$ or the gaps Δ are known.

The true power of Inequality (6) 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. These lower bounds are extremely strong as they hold for all possible bandit problems, not just for some well-chosen ones.

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, and re-derivation of earlier lower bounds. We denote by kl the Kullback-Leibler divergence for Bernoulli distributions:

$$\forall p, q \in [0,1]^2, \qquad \mathrm{kl}(p,q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}\,. \tag{5}$$

We show in this section that for all strategies ψ, for all bandit problems $\underline{\nu}$ and $\underline{\nu}'$, for all $\sigma(I_T)$–measurable random variables Z with values in $[0,1]$,

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\, \mathrm{KL}(\nu_a, \nu'_a) \geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z], \mathbb{E}_{\underline{\nu}'}[Z]\big)\,. \tag{6}$$

Inequality (6) will be referred to as the fundamental inequality of this article. We will typically apply it by considering variables of the form $Z = N_{\psi,k}(T)/T$ for some arm k. That the kl term in (6) 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 bounding. Implicit changes of measures will however be performed by considering bandit problems $\underline{\nu}$ and $\underline{\nu}'$ and their associated probability measures $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$.

Underlying probability measures. The proof of (6) will be based, among others, on an application of the chain rule for Kullback-Leibler divergences. For this reason, it is helpful to construct and define the underlying measures, so that the needed stochastic transition kernels appear clearly.

By Kolmogorov's extension theorem, there exists a measurable space $(\Omega, \mathcal{F})$ on which all probability measures $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$ considered above can be defined; e.g., $\Omega = [0,1] \times \big(\mathbb{R} \times [0,1]\big)^{\mathbb{N}}$. Given the probabilistic and strategic setting described in Section 1.1, the probability measure $\mathbb{P}_{\underline{\nu}}$ over this $(\Omega, \mathcal{F})$ is such that for all $t \geqslant 0$, for all Borel sets $B \subseteq \mathbb{R}$ and $B' \subseteq [0,1]$,

$$\mathbb{P}_{\underline{\nu}}\big(Y_{t+1} \in B,\, U_{t+1} \in B' \,\big|\, I_t\big) = \nu_{\psi_t(I_t)}(B)\; \lambda(B')\,, \tag{7}$$

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

Remark 1. Equation (7) actually reveals that the distributions $\mathbb{P}_{\underline{\nu}}$ should be indexed as well by the considered strategy ψ. Because the important element in the proofs will be the dependency on $\underline{\nu}$ (we will replace $\underline{\nu}$ by alternative bandit problems $\underline{\nu}'$), we drop the dependency on ψ in the notation for the underlying probability measures. This will not come at the cost of clarity as virtually all events A_ψ and random variables Z_ψ that will be considered will depend on ψ: we will almost always deal with probabilities of the form $\mathbb{P}_{\underline{\nu}}(A_\psi)$ or expectations of the form $\mathbb{E}_{\underline{\nu}}[Z_\psi]$.

2.1. Proof of the fundamental inequality (6). We let $\mathbb{P}_{\underline{\nu}}^{I_T}$ and $\mathbb{P}_{\underline{\nu}'}^{I_T}$ denote the respective distributions (pushforward measures) of I_T under $\mathbb{P}_{\underline{\nu}}$ and $\mathbb{P}_{\underline{\nu}'}$. We add an intermediate equation in (6),

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\, \mathrm{KL}(\nu_a, \nu_a') = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_T}, \mathbb{P}_{\underline{\nu}'}^{I_T}\big) \geqslant \mathrm{kl}\big(\mathbb{E}_{\underline{\nu}}[Z], \mathbb{E}_{\underline{\nu}'}[Z]\big)\,, \tag{8}$$

and are left with proving a standard equality (via the chain rule for Kullback-Leibler divergences) and a less standard inequality (following from the data-processing inequality for Kullback-Leibler divergences).

Remark 2. Although this possibility is not used in the present article, it is important to note, after Kaufmann et al. [19, Lemma 1], that (8) actually holds not only for deterministic values of T but also for any stopping time with respect to the filtration generated by $(I_t)_{t \geqslant 1}$.

Proof of the equality in (8). This equality 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. [3] and Cesa-Bianchi and Lugosi [12]; see also Combes and Proutière [13]. We thus reprove this equality for the sake of completeness only.

We use the symbol $\otimes$ to denote products of measures. The stochastic transition kernel (7) exactly indicates that the conditional distribution of (Y_{t+1}, U_{t+1}) given I_t equals

$$\mathbb{P}_{\underline{\nu}}^{(Y_{t+1},U_{t+1})\,|\,I_t} = \nu_{\psi_t(I_t)} \otimes \lambda\,.$$

Because the conditional distribution at hand takes such a simple form, the chain rule for Kullback-Leibler divergences applies; it ensures that for all $t \geqslant 0$,

$$\begin{aligned}\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_{t+1}}, \mathbb{P}_{\underline{\nu}'}^{I_{t+1}}\Big) &= \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{(I_t,Y_{t+1},U_{t+1})}, \mathbb{P}_{\underline{\nu}'}^{(I_t,Y_{t+1},U_{t+1})}\Big) \\ &= \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_t}, \mathbb{P}_{\underline{\nu}'}^{I_t}\big) + \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{(Y_{t+1},U_{t+1})\,|\,I_t}, \mathbb{P}_{\underline{\nu}'}^{(Y_{t+1},U_{t+1})\,|\,I_t}\Big)\,,\end{aligned} \tag{9}$$

where

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

Recalling that $A_{t+1}=\psi_t(I_t)$, we proved so far

$$\mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{I_{t+1}},\,\mathbb{P}_{\underline{\nu}'}^{I_{t+1}}\Big)=\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_t},\,\mathbb{P}_{\underline{\nu}'}^{I_t}\big)+\mathbb{E}_{\underline{\nu}}\left[\sum_{a=1}^{K}\mathrm{KL}(\nu_a,\nu'_a)\,\mathbb{I}_{\{A_{t+1}=a\}}\right].$$

Iterating the argument and using that $\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_0},\,\mathbb{P}_{\underline{\nu}'}^{I_0}\big)=\mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{U_0},\,\mathbb{P}_{\underline{\nu}'}^{U_0}\big)=\mathrm{KL}(\lambda,\lambda)=0$ leads to the equality stated in (8).

***Proof of the inequality in* (8).** *This is our key contribution to a simplified proof of the lower bound* (2). It is a consequence of the data-processing inequality (also known as contraction of entropy), i.e., the fact that Kullback-Leibler divergences between pushforward measures are smaller than the Kullback-Leibler divergences between the original probability measures; see Lemma 5 in Appendix A for a statement and elements of proof.

We actually state our inequality in a slightly more general way, as it is of independent interest.

LEMMA 1. *Consider a measurable space $(\Gamma,\mathcal{G})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, and any $\mathcal{G}$–measurable random variable $Z:\Omega\to[0,1]$. We denote respectively by $\mathbb{E}_1$ and $\mathbb{E}_2$ the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$. 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 $\mathcal{B}\big([0,1]\big)$ and the Lebesgue measure λ. We denote by $\mathcal{G}\otimes\mathcal{B}\big([0,1]\big)$ the σ–algebra generated by product sets in $\mathcal{G}\times\mathcal{B}\big([0,1]\big)$. Now, for any event $E\in\mathcal{G}\otimes\mathcal{B}\big([0,1]\big)$, by the consideration of product distributions for the equality and by the data-processing inequality (Lemma 5) applied to $X=\mathbb{I}_E$ for the inequality, we have

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2)=\mathrm{KL}\big(\mathbb{P}_1\otimes\lambda,\,\mathbb{P}_2\otimes\lambda\big)\geqslant\mathrm{KL}\Big((\mathbb{P}_1\otimes\lambda)^{\mathbb{I}_E},\,(\mathbb{P}_2\otimes\lambda)^{\mathbb{I}_E}\Big).$$

The distribution $(\mathbb{P}_j\otimes\lambda)^{\mathbb{I}_E}$ of $\mathbb{I}_E$ under $\mathbb{P}_j\otimes\lambda$ is a Bernoulli distribution, with parameter the probability of E under $\mathbb{P}_j\otimes\lambda$; therefore, using the notation kl, we have got so far

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

We consider $E=\big\{(\gamma,x)\in\Gamma\times[0,1]:x\leqslant Z(\gamma)\big\}$ and note noting that for all j, by the Fubini-Tonelli theorem,

$$(\mathbb{P}_j\otimes\lambda)(E)=\int_\Omega\left(\int_{[0,1]}\mathbb{I}_{\{x\leqslant Z(\gamma)\}}\,\mathrm{d}\lambda(x)\right)\mathrm{d}\mathbb{P}_j(\gamma)=\int_\Omega Z(\gamma)\,\mathrm{d}\mathbb{P}_j(\gamma)=\mathbb{E}_j[Z].$$

This concludes the proof of this lemma. □

2.2. Application: re-derivation of the general asymptotic distribution-dependent bound. As a warm-up, we show how the asymptotic distribution-dependent lower bound (2) of Burnetas and Katehakis [9] can be reobtained, for so-called uniformly fast convergent strategies.

Definition 1. A strategy ψ is uniformly fast convergent on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a, i.e., for all arms a such that $\Delta_a > 0$, for all $0 < \alpha \leqslant 1$, it satisfies $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] = o(T^\alpha)$.

Theorem 1. *For all models $\mathcal{D}$, for all uniformly fast convergent strategies ψ on $\mathcal{D}$, for all bandit problems $\underline{\nu}$, for all suboptimal arms a,*

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

Proof. Given any bandit problem $\underline{\nu}$ and any suboptimal arm a, we consider a modified problem $\underline{\nu}'$ 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 (6) with $Z = N_{\psi,a}(T)/T$. All Kullback-Leibler divergences in its left-hand side are null except the one for arm a, so that we get the lower bound

$$\begin{aligned}\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a, \nu'_a) &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T}\right)\\ &\geqslant \left(1 - \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T}\right)\ln\frac{T}{T - \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]} - \ln 2\,,\end{aligned} \tag{10}$$

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

$$\mathrm{kl}(p,q) = \underbrace{p\ln\frac{1}{q}}_{\geqslant 0} + (1-p)\ln\frac{1}{1-q} + \underbrace{\big(p\ln p + (1-p)\ln(1-p)\big)}_{\geqslant -\ln 2}\,. \tag{11}$$

The uniform fast convergence of ψ together with the fact that all arms $k \neq a$ are suboptimal for $\underline{\nu}'$ entails that

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

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

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

In addition, the uniform fast convergence of ψ and the suboptimality of a for the bandit problem $\underline{\nu}$ ensure that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]/T \to 0$. Substituting these two facts in (10) we proved

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

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, $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}) = +\infty$ 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 $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ is of order T/K as long as T is at most of order $1/\mathcal{K}_{\inf}(\nu_a,\mu^\star,\mathcal{D})$; 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 $T\approx K/\mathcal{K}_{\inf}(\nu_a,\mu^\star,\mathcal{D})$.

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 $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)/N_{\psi,a^\star}(T)\big]$ 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 $\sum_{a\notin\mathcal{A}^\star(\underline{\nu})}\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ where $\mathcal{A}^\star(\underline{\nu})$ denotes the set of the $A^\star_{\underline{\nu}}$ optimal arms of $\underline{\nu}$. It is of the desired order $T(1-A^\star_{\underline{\nu}}/K)$ for times T of the desired order $K/\mathcal{K}^{\max}_{\underline{\nu}}$, where $\mathcal{K}^{\max}_{\underline{\nu}}$ is some Kullback-Leibler divergence.

Minimal restrictions on the considered strategies. We prove these lower bounds under minimal assumptions on the considered strategies: either some mild symmetry (much milder than asking for symmetry under permutation of the arms, see Definition 3); or the fact that for suboptimal arms a, the number of pulls $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]$ should decrease as μ_a decreases, all other distributions of arms being fixed (see Definitions 2 and 4). These assumptions are satisfied by all well-performing strategies we could think of: the UCB strategy of Auer et al. [2], the KL-UCB strategy of Cappé et al. [11], Thompson [24] Sampling, EXP3 of Auer et al. [3], etc.

These mild restrictions on the considered strategies are necessary to rule out the irrelevant strategies (e.g., always pull arm 1) that would perform extremely well for some particular bandit problems $\underline{\nu}$. This is because we aim at proving distribution-dependent lower bounds that are valid for all bandit problems $\underline{\nu}$: we prefer to impose the (mild) constraints on the strategies.

Note that the assumption of uniform fast convergence (Definition 1), though classical and well accepted, is quite strong. Note that it is necessary for a strategy to satisfy some symmetry and to be smarter than the uniform strategy in the limit (not for all T, see Definition 2) to be uniformly fast convergent. Hence, the class of strategies we consider is essentially much larger than the subset of uniformly fast convergent strategies.

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 2. A strategy ψ is smarter than the uniform strategy on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all optimal arms $a^\star$, for all $T\geqslant 1$,

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

Theorem 2. *For all models $\mathcal{D}$, for all strategies ψ that are smarter than the uniform strategy on $\mathcal{D}$, for all bandit problems $\underline{\nu}$, for all arms a, for all $T\geqslant 1$,*

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

In particular,

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

Proof. The definition of being smarter than the uniform strategy takes care of the lower bound for optimal arms a: it thus suffices to consider suboptimal arms a. As in the proof of Theorem 1, we consider a modified bandit problem $\underline{\nu}'$ 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 (6), we get

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T}\right).$$

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

$$\mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big]}{T}\right) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{1}{K}\right).$$

Lemma 6 of Appendix A yields

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T},\ \frac{1}{K}\right) \geqslant \frac{K}{2}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]}{T} - \frac{1}{K}\right)^2,$$

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

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

Taking the supremum of the right-hand side over all $\nu'_a \in \mathcal{D}$ such that $E(\nu'_a) > \mu^\star$ and rearranging concludes the proof. □

3.2. Relative lower bound. Our proof will be based on an assumption of symmetry (milder than requiring that if the arms are permuted in a bandit problem, the algorithm behaves the same way, as in Definition 7).

Definition 3. A strategy ψ is pairwise symmetric for optimal arms on $\mathcal{D}$ if for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for each pair of optimal arms $a^\star$ and $a_\star$, the equality $\nu_{a^\star} = \nu_{a_\star}$ entails that, for all $T \geqslant 1$,

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

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 3. *For all models $\mathcal{D}$, for all strategies ψ that are pairwise symmetric for optimal arms on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a and all optimal arms $a^\star$, for all $T \geqslant 1$,*

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

In particular,

$$\forall T \leqslant \frac{K}{32\,\mathrm{KL}(\nu_a,\nu_{a^\star})}, \qquad \textit{either}\quad \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{K} \qquad \textit{or}\quad \mathbb{E}_{\underline{\nu}}\left[\frac{\max\big\{N_{\psi,a}(T),1\big\}}{\max\big\{N_{\psi,a^\star}(T),1\big\}}\right] \geqslant \frac{1}{2}\,.$$

That is, on average, in the small T regime, each suboptimal arm is played at least half the number of times when an optimal arm was played.

Proof. For all arms k, we denote by $N^+_{\psi,k}(T) = \max\{N_{\psi,k}(T), 1\}$. Given a bandit problem $\underline{\nu}$ and a suboptimal arm a, we form an alternative bandit problem $\underline{\nu}'$ 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 $\underline{\nu}$. In particular, arms a and $a^\star$ are both optimal arms under $\underline{\nu}'$. By the assumption of pairwise symmetry for optimal arms, we have in particular that

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

The latter equality and the fundamental inequality (6) yield in the present case, through the choice of $Z = N^+_{\psi,a}(T)/\big(N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)\big)$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\ \mathrm{KL}(\nu_a,\nu'_a) \geqslant \mathrm{kl}\left(\mathbb{E}_{\underline{\nu}}\left[\frac{N^+_{\psi,a}(T)}{N^+_{\psi,a}(T)+N^+_{\psi,a^\star}(T)}\right],\ \frac{1}{2}\right). \tag{12}$$

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

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

We can assume that $\mathbb{E}_{\underline{\nu}}\big[N^+_{\psi,a}(T)/N^+_{\psi,a^\star}(T)\big] \leqslant 1$, 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 $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \leqslant T/K$ (otherwise, the result of the theorem is obtained as well), we get from (12)

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

Pinsker's inequality (in its classical form, see Appendix A for a statement) entails the inequality

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

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 $\underline{\nu}$, we denote by $\mathcal{A}^\star(\underline{\nu})$ the set of its optimal arms and by $\mathcal{W}(\underline{\nu})$ the set of its worst arms, i.e., the ones associated with the distributions with the smallest expectation among all distributions for the arms. We also let $A^\star_{\underline{\nu}}$ be the cardinality of $\mathcal{A}^\star(\underline{\nu})$.

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

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

In particular, $\mathcal{A}^\star(\underline{\nu}) = \mathcal{A}^\star(\underline{\nu}')$ in this case. The definition models the fact that the bandit problem $\underline{\nu}'$ should be easier than $\underline{\nu}$, as non-optimal arms in $\underline{\nu}'$ are farther away from the optimal arms (in expectation) that in $\underline{\nu}$. Any reasonable strategy should perform better on $\underline{\nu}'$ than on $\underline{\nu}$, 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 $\underline{\nu}$ and $\underline{\nu}'$.)

Definition 4. A strategy ψ is monotonic on a model $\mathcal{D}$ if for all bandit problems $\underline{\nu}' \preccurlyeq \underline{\nu}$ in $\mathcal{D}$,

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

Theorem 4. *For all models $\mathcal{D}$, for all strategies ψ that are pairwise symmetric for optimal arms and monotonic on $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, suboptimal arms are collectively sampled at least*

$$\sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant T\left(1 - \frac{A^\star_{\underline{\nu}}}{K} - \frac{A^\star_{\underline{\nu}}\sqrt{2T\,\mathcal{K}^{\max}_{\underline{\nu}}}}{K} - \frac{2A^\star_{\underline{\nu}} T \mathcal{K}^{\max}_{\underline{\nu}}}{K}\right),$$
$$\text{where} \qquad \mathcal{K}^{\max}_{\underline{\nu}} = \min_{w \in \mathcal{W}(\underline{\nu})} \max_{a^\star \in \mathcal{A}^\star(\underline{\nu})} \mathrm{KL}(\nu_w, \nu_{a^\star})\,.$$

In particular,

$$\forall T \leqslant \frac{K}{8\,A^\star_{\underline{\nu}}\mathcal{K}^{\max}_{\underline{\nu}}}, \qquad \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{T}{2}\left(1 - \frac{A^\star_{\underline{\nu}}}{K}\right).$$

To get a lower bound on the regret from this theorem, we use

$$R_{\psi,\underline{\nu},T} \geqslant \left(\min_{a \notin \mathcal{A}^\star(\underline{\nu})} \Delta_a\right) \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big]\,. \tag{13}$$

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

In the second model $\underset{\sim}{\nu}$, each arm is associated with $\nu_{\widetilde{w}}$, i.e., $\underset{\sim}{\nu}_a = \nu_{\widetilde{w}}$ for all $a \in \{1,\dots,K\}$.

By monotonicity of ψ,

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

We can therefore focus our attention, for the rest of the proof, on the $\mathbb{E}_{\underline{\underline{\nu}}}\big[N_{\psi,a}(T)\big]$. The strategy is also pairwise symmetric for optimal arms and all arms of $\underset{\sim}{\nu}$ are optimal. This implies in particular that $\mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,1}(T)\big] = \mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a}(T)\big]$ for all arms a, thus $\mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a}(T)\big] = T/K$ for all arms a.

Now, the bound (6) with $Z = \sum_{a^\star \in \mathcal{A}^\star(\underset{\sim}{\nu})} \frac{N_{\psi,a^\star}(T)}{T}$ and the bandit models $\underset{\sim}{\nu}$ and $\underline{\underset{\sim}{\nu}}$ gives

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

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

$$\sum_{a^\star \in \mathcal{A}^\star(\underset{\sim}{\nu})} \mathbb{E}_{\underset{\sim}{\nu}}\big[N_{\psi,a^\star}(T)\big]\mathrm{KL}(\nu_{\widetilde{w}}, \nu_{a^\star}) \leqslant \frac{T A^\star_{\underset{\sim}{\nu}} \mathcal{K}^{\max}_{\underset{\sim}{\nu}}}{K},$$

which yields the inequality

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

We want to upper bound x, in order to get a lower bound on $1-x$. We assume that $x \geqslant A^\star_{\underset{\sim}{\nu}}/K$, otherwise, the bound (14) stated below is also satisfied. Pinsker's inequality (actually, its local refinement stated as Lemma 6 in Appendix A) then ensures that

$$\frac{T A^\star_{\underset{\sim}{\nu}} \mathcal{K}^{\max}_{\underset{\sim}{\nu}}}{K} \geqslant \frac{1}{2x}\left(\frac{A^\star_{\underset{\sim}{\nu}}}{K} - x\right)^2,$$

Lemma 2 below finally entails that

$$x \leqslant \frac{A^\star_{\underset{\sim}{\nu}}}{K}\left(1 + 2T\mathcal{K}^{\max}_{\underset{\sim}{\nu}} + \sqrt{2T\mathcal{K}^{\max}_{\underset{\sim}{\nu}}}\right). \tag{14}$$

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

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

□

Lemma 2. *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 polynomial, 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 $\sqrt{4\alpha\beta+\beta^2} \leqslant \sqrt{4\alpha\beta} + \sqrt{\beta^2}$. □

3.4. Numerical illustrations. In this section we illustrate some of the bounds stated above for the initial linear regime, namely, the bounds of Theorems 2 and 4. It turned out that because of the "or" statement in Theorem 3, its bound was less easy to illustrate. We need much more difficult bandit problems than the one of Figure 1 in order to clearly observe the initial linear phase.

Theorem 2 is illustrated in Figure 2. We observe that in the bandit problems contemplated therein, the expected numbers of pulls of the suboptimal arms considered indeed lie between $T/(2K)$ and T/K in the initial phase, as prescribed by the theorem. We see, however, that this initial phase is probably longer than what was quantified.

Theorem 4 is illustrated in Figure 3. For a large number of arms, the regret lower bound (13) deriving as a consequence of the considered theorem is larger than a bound based on the decomposition of the regret (1) and Theorem 2.

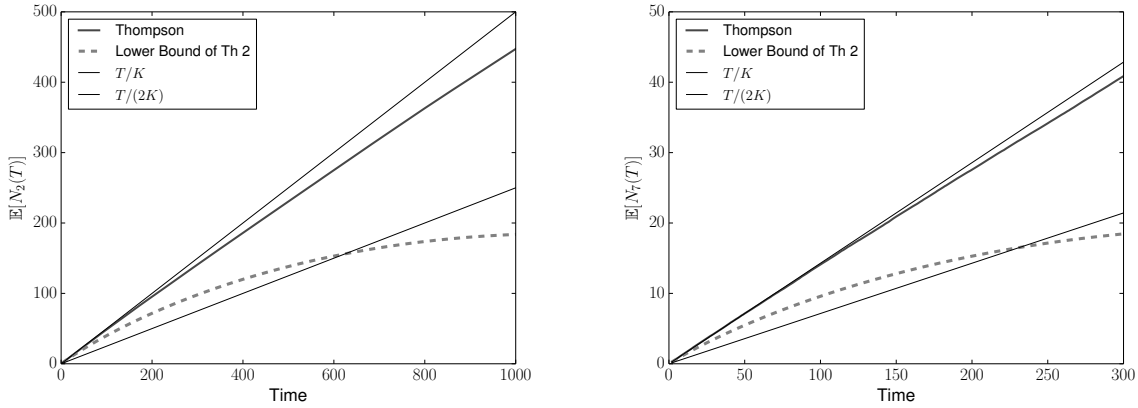

FIGURE 2. Expected number of pulls of the most suboptimal arm for Thompson [24] Sampling (*blue, solid* line) on Bernoulli bandit problems, versus the lower bound (*red, dashed* line) of Theorem 2 for the model $\mathcal{D}$ of all Bernoulli distributions; expectations are approximated over 1,000 runs.

Left: parameters $(\mu_a)_{1\leqslant a\leqslant 2} = (0.5, 0.49)$, with characteristic time $1/\big(8\,\mathcal{K}_{\inf}(\nu_2, \mu^\star, \mathcal{D})\big) \approx 625$.
Right: parameters $(\mu_a)_{1\leqslant a\leqslant 7} = (0.05, 0.048, 0.047, 0.046, 0.045, 0.044, 0.043)$, with $1/\big(8\,\mathcal{K}_{\inf}(\nu_7, \mu^\star, \mathcal{D})\big) \approx 231$.

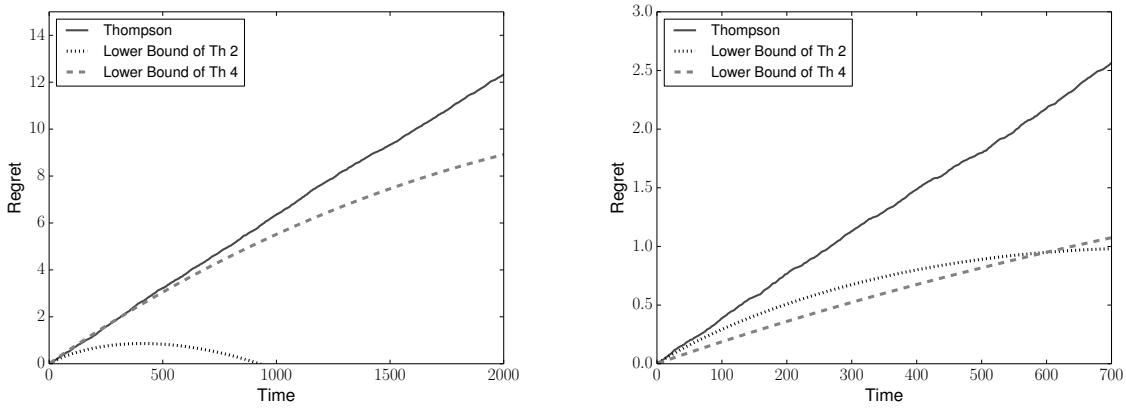

FIGURE 3. Expected regret of Thompson [24] Sampling (*blue, solid* line) on Bernoulli bandit problems, versus the lower bound (*red, dashed* line) of Theorem 4 using (13) and the lower bound (*black, dotted* line) of Theorem 2 using (1), for the model $\mathcal{D}$ of all Bernoulli distributions; expectations are approximated over 3,000 runs.

Left: parameters $(\mu_a)_{1\leqslant a\leqslant 10} = (0.05, 0.043, \ldots, 0.043)$, with characteristic time $K/\big(8\,A^\star_\nu\,\mathcal{K}^{\max}_\nu\big) \approx 1,250$.
Right: parameters $(\mu_a)_{1\leqslant a\leqslant 7} = (0.05, 0.048, 0.047, 0.046, 0.045, 0.044, 0.043)$, with $K/\big(8\,A^\star_\nu\,\mathcal{K}^{\max}_\nu\big) \approx 1,619$.

4. Non-asymptotic bounds for large T. We restrict our attention to well-behaved models and uniformly super-fast convergent strategies. For a given model $\mathcal{D}$, we denote by $E(\mathcal{D})$ the interior of the set of all expectations of distributions in $\mathcal{D}$. That a model is well-behaved means that the function $\mathcal{K}_{\inf}$ is locally Lipschitz continuous in its second variable, as is made formal in the following definition.

DEFINITION 5. A model $\mathcal{D}$ is well behaved if there exist two functions $\varepsilon_{\mathcal{D}} : E(\mathcal{D}) \to (0, +\infty)$ and $\omega_{\mathcal{D}} : \mathcal{D} \times E(\mathcal{D}) \to (0, +\infty)$ such that for all distributions $\nu_a \in \mathcal{D}$ and all $x \in E(\mathcal{D})$ with $x > E(\nu_a)$,

$$\forall \varepsilon < \varepsilon_{\mathcal{D}}(x), \qquad \mathcal{K}_{\inf}(\nu_a, x+\varepsilon, \mathcal{D}) \leqslant \mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}) + \varepsilon\,\omega_{\mathcal{D}}(\nu_a, x)\,.$$

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. [11] to get non-asymptotic upper bounds on the regret of the optimal order (2).

Definition 6. A strategy ψ is uniformly super-fast convergent on a model $\mathcal{D}$ if there exists a constant $C_{\psi,\mathcal{D}}$ such that for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a, for all $T \geqslant 2$,

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

Uniform super-fast convergence is a refinement of the notion of uniform fast convergence based on two considerations. First, that there exist such strategies, for instance, the `UCB` strategy of Auer et al. [2] on any bounded model $\mathcal{D}$, i.e., a model with distributions all supported within a common bounded interval $[m, M]$. Second, Pinsker's inequality (see Appendix A) and Lemma 1 entail in particular that for such bounded models $\mathcal{D}$,

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D}) \geqslant \mathrm{kl}\left(\frac{\mu_a - m}{M-m}, \frac{\mu^\star - m}{M-m}\right) \geqslant \frac{2}{(M-m)^2}\Delta_a^2\,;$$

therefore, the upper bound stated in the definition of uniform super-fast convergence is still weaker than the lower bound (2).

Note that Definition 6 could be relaxed even more: we are mostly interested therein in the logarithmic growth rate $\ln T$. We imposed the $C_{\psi,\mathcal{D}}/\Delta_a^2$ upper bound mostly for simplicity and readability of the calculations that lead to Theorem 5. It would be of course possible to rather consider more abstract problem-dependent constants of the form $C_{\psi,\mathcal{D}}(a,\nu)$, at least as soon as some minimal properties are assumed with respect to the behavior of such constants as functions of the gap $\mu^\star - \mu_a$.

4.1. A general non-asymptotic lower bound. Throughout this subsection, we fix a strategy ψ that is uniformly super-fast convergent with respect to a model $\mathcal{D}$. We recall that we denote by $\mathcal{A}^\star(\underline{\nu})$ the set of optimal arms of the bandit problem $\underline{\nu}$ and let $A^\star_{\underline{\nu}}$ be its cardinality. We adapt the bounds (6) and (10) by using this time

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

and $\mathrm{kl}(p,q) \geqslant p\ln(1/q) - \ln 2$, see (11). For all bandit problems $\underline{\nu}'$ that only differ from $\underline{\nu}$ as far a suboptimal arm a is concerned, whose distribution of payoffs $\nu'_a \in \mathcal{D}$ is such that $\mu'_a = E(\nu'_a) > \mu^\star$, we get

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)}\left(\mathbb{E}_{\underline{\nu}}[Z]\,\ln\frac{1}{\mathbb{E}_{\underline{\nu}'}[Z]} - \ln 2\right). \tag{15}$$

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

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

on the other hand,

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

Denoting

$$H(\underline{\nu}) = \sum_{a \notin \mathcal{A}^\star(\underline{\nu})} \frac{1}{\Delta_a^2} \tag{16}$$

and using that $\underline{\Delta} \geqslant \varepsilon$, a substitution of the two super-fast convergence inequalities into (15) and an optimization over the considered distributions ν'_a leads to

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})} \left(1 - C_{\psi,\mathcal{D}} H(\underline{\nu}) \frac{\ln T}{T}\right) \ln \frac{T\varepsilon^2}{A^\star_{\underline{\nu}}\, C_{\psi,\mathcal{D}} \ln T} - \frac{\ln 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})}\,. \tag{17}$$

The obtained bound holds for all $T \geqslant 2$ (as in the definition of uniform super-fast convergence); 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 $\mathcal{K}_{\inf}(\nu_a, \mu^\star + \varepsilon, \mathcal{D})$ to $\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})$. Since $1/(1+x) \geqslant 1 - x$ for all $x \geqslant 0$, we get by Definition 5

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

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 (17), 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 $A^\star_{\underline{\nu}} \leqslant K$, we get the following theorem.

THEOREM 5. *For all uniformly super-fast convergent strategies ψ on well-behaved models $\mathcal{D}$, for all bandit problems $\underline{\nu}$ in $\mathcal{D}$, for all suboptimal arms a,*

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})} \big(1 - (a_T + b_T + c_T)\big) - \frac{\ln 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star, \mathcal{D})}\,, \tag{18}$$

for all $T \geqslant 2$ large enough so that $(\ln T)^{-4} < \varepsilon_{\mathcal{D}}(\mu^\star)$ and

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

are all smaller than 1, where $H(\underline{\nu})$ was defined in (16).

REMARK 3. We have $(a_T + b_T + c_T) \ln T = \mathrm{O}\big(\ln(\ln T)\big)$. The non-asymptotic bound (18) is therefore of the form

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

Note that the second-order term of typical non-asymptotic upper bounds (e.g., by Cappé et al. [11]) had long been of the form $+(\ln T)^\alpha$ for some $\alpha \in (0,1)$. But recently, Honda and Takemura [17, Theorem 5] showed that at least for models containing distributions that have each a bounded support, the second-order is of order $-\ln(\ln T)$. Our lower bound above thus shows the optimality of the order of magnitude of this second-order term.

4.2. Two (and a half) examples 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. [11] could prove non-asymptotic 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 3. *In the model* $\mathcal{D} = \Big\{\mathrm{m} \in \mathcal{M}\big([0,M]\big) : E(\mathrm{m}) < M\Big\}$, *we have*

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

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

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

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 $\mathcal{K}_{\inf}(\mathrm{m},\mu^\star,\mathcal{D})$. 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

$$E\big(\mathrm{m}'_\alpha\big) > (1-\alpha)\mu^\star + \alpha x = \mu^\star + \alpha(x-\mu^\star) = \mu^\star + \varepsilon\,. \tag{19}$$

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 (19), 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 4. 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 3 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 [22] for a thorough exposition or Cappé et al. [11] 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

$$\forall\,(\mu,\mu')\in I^2,\qquad \mathrm{KL}\big(\nu_\mu,\nu_{\mu'}\big)=g(\mu)-g(\mu')-(\mu-\mu')\,\dot g(\mu')\,. \tag{20}$$

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

$$\forall\,(\mu,\mu^\star)\in I^2\ \text{ s.t }\ \mu<\mu^\star,\qquad \mathcal{K}_{\inf}(\nu_\mu,\mu^\star,\mathcal{D})=\mathrm{KL}\big(\nu_\mu,\nu_{\mu^\star}\big)\,. \tag{21}$$

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

$$B_{\mu^\star}=\min\left\{\frac{M-\mu^\star}{2},\,1\right\}. \tag{22}$$

The minimum with 1 is considered merely for $B_{\mu^\star}$ to always have a finite value; otherwise, the bound in the lemma below would be uninformative.

Lemma 4. *In a model $\mathcal{D}$ given by a regular exponential family indexed by $I=(m,M)$ and whose Kullback-Leibler divergence (20) is represented by a function g, we have, with the notation (22),*

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

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

Proof. Since $\mu<\mu^\star$, we get by (20) and (21)

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

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 $\mathcal{K}_{\inf}(\nu_\mu,\mu^\star+\varepsilon,\mathcal{D})-\mathcal{K}_{\inf}(\nu_\mu,\mu^\star,\mathcal{D})$ 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 actually involves $B_{\mu^\star}$ as various monotonicity arguments can be invoked.

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 $G_{\mu^\star}=1/\mu^\star$.

[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 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

$$G_{\mu^\star} \leqslant \frac{2n}{\mu^\star(n - \mu^\star)}.$$

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 $n - \mu \geqslant n - \mu^\star - B_{\mu^\star} \geqslant (n - \mu^\star)/2$.

Appendix A: Reminder of some elements of information theory. For the sake of self-completeness we recall two selected basic facts pertaining to Kullback-Leibler divergences.

The data-processing inequality. The most elegant proof we are aware of relies on a conditional Jensen's inequality applied to $t \mapsto t \ln t$; see Ali and Silvey [1].

LEMMA 5. *Consider a measurable space $(\Gamma, \mathcal{G})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, any other $(\Gamma', \mathcal{G}')$ measurable space, and any random variable $X : (\Gamma, \mathcal{G}) \to (\Gamma', \mathcal{G}')$. Then,*

$$\mathrm{KL}\big(\mathbb{P}_1^X, \mathbb{P}_2^X\big) \leqslant \mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2),$$

where $\mathbb{P}_1^X$ and $\mathbb{P}_2^X$ denote the respective distributions of X under $\mathbb{P}_1$ and $\mathbb{P}_2$.

On local refinements of Pinsker's inequality. Pinsker's inequality reads, for Bernoulli distributions, in its most classical form:

$$\forall (p, q) \in [0, 1]^2, \qquad \mathrm{kl}(p, q) \geqslant 2(p - q)^2. \tag{23}$$

The lemma below offers a local refinement of Pinsker's inequality for Bernoulli distributions; the classical form (23) follows by noting that $x(1 - x) \leqslant 1/4$ for $x \in [0, 1]$. Cappé et al. [11, Lemma 3 in Appendix A.2.1] offer an extension of this local refinement to any one-parameter regular exponential family.

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

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

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

Appendix B: Re-derivation of other earlier lower bounds In this section, we re-derive the bounds discussed in Section 1.3, based on our fundamental inequality (6). We do so to illustrate the power and the versatility of (6). However, we point out again that the lower bounds discussed here are much weaker than the ones derived in the main body of the paper: in the terminology of Section 1.3, they are of the form (well-chosen) rather than of the form (all).

B.1. Distribution-free lower bound. We consider the bound (3) recalled in Section 1.3. More specifically, we re-prove Theorem A.2 of Auer et al. [3], from which the stated bound (3) follows by optimization over ε.

THEOREM 6. *Consider the bandit model $\mathcal{D}=\mathcal{M}\big([0,1]\big)$ of all probability distributions over $[0,1]$. For all $\varepsilon\in(0,1/2)$, for all strategies ψ, there exists a bandit problem $\underline{\nu}'$ in $\mathcal{M}\big([0,1]\big)$ such that*

$$R_{\psi,\underline{\nu}',T}\geqslant T\varepsilon\left(1-\frac{1}{K}-\frac{1}{2}\sqrt{\frac{T}{K}\ln\frac{1}{1-4\varepsilon^2}}\right).$$

This problem $\underline{\nu}'$ can be given by Bernoulli distributions, with parameters $1/2$ for all arms but one, for which the parameter is $1/2+\varepsilon$.

As a consequence, the worst-case regret of any strategy ψ against all bandit problems $\underline{\nu}$ in $\mathcal{M}\big([0,1]\big)$ is lower bounded as announced in (3):

$$\sup_{\underline{\nu}} R_{\psi,\underline{\nu},T}\geqslant \sup_{\varepsilon\in(0,1/2)} T\varepsilon\left(1-\frac{1}{K}-\frac{1}{2}\sqrt{\frac{T}{K}\ln\frac{1}{1-4\varepsilon^2}}\right)\geqslant\frac{1}{20}\min\Big\{\sqrt{KT},\,T\Big\}.$$

The second inequality above is proved by a simple calculation indicated after the proof of Theorem A.2 of Auer et al. [3]: pick $\varepsilon=\min\big\{\sqrt{K/T},\,1\big\}/4$ and use $-\ln(1-u)\leqslant\big(4\ln(4/3)\big)u$ for $u\in(0,1/4)$. The constant $1/20$ can actually be improved into $1/8$, see Cesa-Bianchi and Lugosi [12, Theorem 6.11].

Proof. We fix a strategy and $\varepsilon\in(0,1/2)$. We denote by $\underline{\nu}$ the bandit problem where all distributions are given by Bernoulli distributions with parameter $1/2$. There exists an arm $k\in\{1,\dots,K\}$ such that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]\leqslant T/K$, as these K numbers of pulls sum up to T. We define the bandit

problem $\underline{\nu}'$ by $\nu'_a = \nu_a$ for $a \neq k$, that is, ν'_a is a symmetric Bernoulli distribution, while ν'_k is the Bernoulli distribution with parameter $1/2+\varepsilon$. By (1), we have

$$R_{\psi,\underline{\nu}',T} = \sum_{a\neq k} \varepsilon\, \mathbb{E}_{\underline{\nu}'}\big[N_{\psi,a}(T)\big] = T\varepsilon\left(1 - \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right). \tag{24}$$

A direct computation of $\mathrm{kl}(1/2, 1/2+\varepsilon)$ and the application of (6) indicate that

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{2} \ln\frac{1}{1-4\varepsilon^2} = \mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]\,\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\varepsilon\right) \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T}, \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right).$$

Now, Pinsker's inequality (in its classical form, see Appendix A) ensures that

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{2} \ln\frac{1}{1-4\varepsilon^2} \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T}, \frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T}\right) \geqslant 2\left(\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T} - \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T}\right)^2.$$

Solving for $\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]/T$, based on whether $\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]/T$ is larger or smaller than $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]/T$, we get, in all cases,

$$\frac{\mathbb{E}_{\underline{\nu}'}\big[N_{\psi,k}(T)\big]}{T} \leqslant \frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T} + \frac{1}{2}\sqrt{\frac{\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big]}{T} \ln\frac{1}{1-4\varepsilon^2}}.$$

The proof is concluded by substituting the fact that $\mathbb{E}_{\underline{\nu}}\big[N_{\psi,k}(T)\big] \leqslant T/K$ by definition of k, and by combining the obtained inequality with (24). □

The short proof above actually re-uses absolutely all the original arguments of Auer et al. [3]: the same Bernoulli distributions, the chain rule for Kullback-Leibler divergences, Pinsker's inequality. It is merely stated in a compact way, that puts under the same umbrella the distribution-dependent and the distribution-free lower bounds for multi-armed bandit problems.

B.2. Lower bounds for the case when $\mu^\star$ or the gaps Δ are known. We consider here the second framework discussed in Section 1.3, with sub-Gaussian bandit problems. For simplicity, and following Bubeck et al. [7], we restrict our attention to lower bounds for two-armed bandit problems (i.e., for $K = 2$).

Known largest expected payoff $\mu^\star$ but unknown gap Δ. The lower bound stated in Theorem 7 below corresponds to Theorem 8 of Bubeck et al. [7], later revisited by the authors, see [8]. It turns out that, as hinted at in, e.g., Faure et al. [15, end of Section 1.4], the initially claimed $\ln T$ dependency is incorrect and a bounded regret can be guaranteed. As shown in Theorem 9 in the next section, this bound on the regret can be as small as $\ln(1/\Delta)/\Delta$. The lower bound we could get using our techniques is of order $1/\Delta$.

To state it, we restrict our attention to strategies ψ symmetric in some sense, e.g., in the sense of Definition 3 stated later on. We actually need very little symmetry here: the considered strategies ψ should just be such that in the bandit problem $\underline{\nu}_0 = \big(\mathcal{N}(0,1), \mathcal{N}(0,1)\big)$, in which the two arms have the same distribution,

$$\mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,1}(T)\big] = \mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,2}(T)\big] = \frac{T}{2}. \tag{25}$$

Of course, all reasonable strategies are usually even more symmetric than that: they are usually stable by permutations over the arms (i.e., they base their decisions only on the payoffs received, not on the labeling of the arms).

THEOREM 7. *For all $\Delta > 0$ we consider $\underline{\nu}_\Delta = \big(\mathcal{N}(0,1), \mathcal{N}(-\Delta,1)\big)$ and $\underline{\nu}_0 = \big(\mathcal{N}(0,1), \mathcal{N}(0,1)\big)$. For all strategies ψ that are symmetric in the sense of (25), for all $\Delta > 0$, for all $T \geqslant 1$,*

$$\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant \frac{1}{\Delta^2 + 1/T} \qquad \textit{and} \qquad R_{\psi,\underline{\nu}_\Delta,T} \geqslant \frac{\Delta}{\Delta^2 + 1/T}\,.$$

In addition, for all strategies ψ and for all T such that $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 1$,

$$\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant \min\left\{\frac{2\ln 2}{\Delta^2 + 2\ln(4T)/T},\ \frac{T}{2}\right\} \qquad \textit{and} \qquad R_{\psi,\underline{\nu}_\Delta,T} \geqslant \min\left\{\frac{2(\ln 2)\Delta}{\Delta^2 + 2\ln(4T)/T},\ \frac{T\Delta}{2}\right\}.$$

Note that the constraint that $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 1$ is satisfied for all $T \geqslant K$ by most of the reasonable strategies, as the latter typically start by playing each arm once (in a random order).

Proof. We first note that $R_{\psi,\underline{\nu}_\Delta,T} = \Delta\, \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]$. Inequality (6) entails that

$$\begin{aligned}\frac{\Delta^2}{2}\,\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] &= \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]\ \mathrm{KL}\big(\mathcal{N}(-\Delta,1), \mathcal{N}(0,1)\big) \\ &\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_0}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T},\ \frac{1}{2}\right), \quad (26)\end{aligned}$$

where we used respectively, for the two equalities, the closed-form expression for the Kullback-Leibler divergences between Gaussian distribution with the same variance and the symmetry assumption on the strategy. Pinsker's inequality (in its classical form, see Appendix A), followed by the inequality

$$\forall\, x \in \mathbb{R}, \qquad 2\left(\frac{1}{2} - x\right)^2 \geqslant \frac{1}{2} - 2x\,,$$

yields

$$\frac{\Delta^2}{2}\,\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big] \geqslant 2\left(\frac{1}{2} - \frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T}\right)^2 \geqslant \frac{1}{2} - 2\frac{\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]}{T}\,.$$

Simple manipulations entail the first claimed bound on $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]$.

For the second one, given the form of the lower bound, which involves a minimum with $T/2$, it suffices to consider the case when $\mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]/T \leqslant 1/2$. We use that

$$\mathrm{kl}(x,\,1/2) = \ln 2 - h(x)\,, \qquad \text{where} \qquad h(x) = -\big(x\ln x + (1-x)\ln(1-x)\big)$$

is the binary entropy function. Now, Calabro [10, page 8] indicates that $h(x) \leqslant x\ln(4/x)$ for all $x \in [0, 1/2]$, so that, restricting our attention to $x \geqslant 1/T$, we get

$$\forall\, x \in [1/T,\,1/2], \qquad \mathrm{kl}\left(x,\,\frac{1}{2}\right) \geqslant \ln 2 - x\ln\left(\frac{4}{x}\right) \geqslant \ln 2 - x\ln(4T)\,.$$

Substituting this inequality into (26), using that $x = \mathbb{E}_{\underline{\nu}_\Delta}\big[N_{\psi,2}(T)\big]/T$ lies in $[1/T,\,1/2]$, concludes the proof. □

The proof above, which is simple and direct, illustrates the interest of Inequality (6) over the standard approaches used so far to prove lower bounds in the same or similar settings.

Known gap Δ but unknown largest expected payoff $\mu^\star$. The lower bound stated in Theorem 8 below corresponds to Theorem 6 of Bubeck et al. [7]. It shows the optimality of the performance bound $\ln(T\Delta^2)/\Delta$ on the regret of the Improved–UCB strategy introduced by [4] and further studied by [16]. The latter improved the constant in the leading term, which equals $\ln(T\Delta^2)/(2\Delta)$ when the gap Δ between the expected payoffs between the two Gaussian arms with variance 1 is known.

We denote by W the Lambert function: for all $u \geqslant 0$, there exists a unique $v \geqslant 0$ such that $u\exp(u) = v$, which is denoted by $v = W(u)$. The Lambert function W is increasing on $[0, +\infty)$. One may easily check that

$$\forall x \geqslant \mathrm{e}, \qquad \ln(x) - \ln\big(\ln(x)\big) \leqslant W(x) \leqslant \ln(x)\,.$$

We state below two lower bounds: one for all strategies ψ, in terms of a maximum between two regrets; and one for strategies that are symmetric and invariant by translation. These properties of symmetry and invariance by translation are most natural requirements. To define them, for all $c \in \mathbb{R}$ and all distributions ν, we denote by $\tau_c(\nu)$ the distribution of $Y + c$ when $Y \sim \nu$.

Definition 7. A strategy ψ for K–armed bandits is symmetric and invariant by translation of the payoffs if for all permutations σ of $\{1, \dots, K\}$, all $c \in \mathbb{R}$, and all $T \geqslant 1$, the distribution of $\big(N_{\psi,1}(T), \dots, N_{\psi,K}(T)\big)$ in the bandit problem $(\nu_1, \dots, \nu_K)$ is equal to the one of $\big(N_{\psi,\sigma^{-1}(1)}(T), \dots, N_{\psi,\sigma^{-1}(K)}(T)\big)$ in the bandit problem $\big(\tau_c(\nu_{\sigma(1)}), \dots, \tau_c(\nu_{\sigma(K)})\big)$.

Theorem 8. *We fix $\Delta > 0$ and consider $\underline{\nu}_1 = \big(\mathcal{N}(0,1), \mathcal{N}(-\Delta,1)\big)$ and $\underline{\nu}_2 = \big(\mathcal{N}(0,1), \mathcal{N}(\Delta,1)\big)$. Then, for all strategies ψ, for all $T \geqslant 1$,*

$$\max\big\{R_{\psi,\underline{\nu}_1,T},\, R_{\psi,\underline{\nu}_2,T}\big\} \geqslant \min\left\{\frac{W\big(T\Delta^2/1.2\big)}{2\Delta},\, \frac{T\Delta}{2}\right\}. \tag{27}$$

Or, alternatively, for all strategies ψ that are symmetric and invariant by translation of the payoffs, for all $T \geqslant 1$,

$$R_{\psi,\underline{\nu}_1,T} = R_{\psi,\underline{\nu}_2,T} \geqslant \frac{W\big(T\Delta^2/1.2\big)}{2\Delta}\,.$$

Remark 5. We compare the obtained bound (27) to Theorem 6 of Bubeck et al. [7]. First, the proof reveals that (27) holds for all distributions $\underline{\nu}_1 = \big(P_0, \mathcal{N}(-\Delta,1)\big)$ and $\underline{\nu}_2 = \big(P_0, \mathcal{N}(\Delta,1)\big)$ where P_0 is a probability distribution with expectation 0. For instance, Bubeck et al. [7] considered the Dirac mass δ_0 at 0.

Second, Theorem 6 of Bubeck et al. [7] offers the bound

$$\max\big\{R_{\psi,\underline{\nu}_1,T},\, R_{\psi,\underline{\nu}_2,T}\big\} \geqslant \frac{\ln\big(T\Delta^2/2\big)}{4\Delta}\,. \tag{28}$$

Asymptotically, as $T \to +\infty$, our bound (27) is smaller by a factor of 2. For small values of T (or small values of Δ), the bound (28) is void as the logarithmic term is non-positive, while our bound is always nonnegative. The second argument of the minimum in (27) is unimportant, as the regret is always bounded by $T\Delta$.

Proof. We have $R_{\psi,\underline{\nu}_1,T} = \Delta\,\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]$ and $R_{\psi,\underline{\nu}_2,T} = \Delta\,\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]$, so that it suffices to lower bound

$$x = \frac{1}{T}\max\Big\{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big],\, \mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]\Big\}\,.$$

We assume below that the maximum is given by the first term; otherwise, the proof below should be adapted by exchanging the roles of $\underline{\nu}_1$ and $\underline{\nu}_2$. Inequality (6) indicates that

$$2T\Delta^2 x = 2\Delta^2 \mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big] = \mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]\ \mathrm{KL}\big(\mathcal{N}(-\Delta,1),\mathcal{N}(\Delta,1)\big)$$
$$\geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}\left(x,\ 1-\frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T}\right).$$

Given the form of the lower bound in the theorem, which involves a minimum with $T\Delta/2$, we may assume, with no loss of generality, that $x \leqslant 1/2$. Since $\mathrm{kl}(x,\cdot)$ is increasing on $[x,1]$ and since

$$1-\frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T} \geqslant 1-x \geqslant \frac{1}{2} \geqslant x\,,$$

by definition of x and the assumption $x \leqslant 1/2$, we get

$$2T\Delta^2 x \geqslant \mathrm{kl}(x,1-x) = (1-2x)\ln\frac{1-x}{x}\,.$$

Note that the case $x=0$ is excluded by the inequality above. A function study shows that

$$\forall\, x\in(0,1), \qquad (1-2x)\ln\frac{1-x}{x} \geqslant \ln\frac{1}{2.4\,x}\,.$$

Substituting this lower bound and taking exponents, we are left with studying the inequality

$$\exp\big(2T\Delta^2 x\big) \geqslant \frac{1}{2.4\,x}\,, \qquad \text{or equivalently,} \qquad 2T\Delta^2 x\ \exp\big(2T\Delta^2 x\big) \geqslant \frac{T\Delta^2}{1.2}\,.$$

By definition of the Lambert function W, we rewrite this inequality as $2T\Delta^2 x \geqslant W\big(T\Delta^2/1.2\big)$, which concludes the proof of the first statement.

For the second statement, we note that the property of invariance by translation of the payoffs ensures that

$$x = \frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T} = \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,1}(T)\big]}{T}\,.$$

Therefore, the fundamental inequality (6) directly gives in this case

$$2T\Delta^2 x \geqslant \mathrm{kl}\left(\frac{\mathbb{E}_{\underline{\nu}_1}\big[N_{\psi,2}(T)\big]}{T},\ \frac{\mathbb{E}_{\underline{\nu}_2}\big[N_{\psi,2}(T)\big]}{T}\right) = \mathrm{kl}(x,1-x)\,,$$

and we do not need to distinguish whether x is larger than $1/2$ or not. The end of the proof of the first statement of the theorem did not use that $x \leqslant 1/2$ and can still safely be followed for the second statement. □

Appendix C: A finite-regret algorithm when $\mu^\star$ is known. In this section, and in this section only, as we are discussing a specific strategy (described below in a box), we will not index the regret, the number of times a given arm is pulled, etc., by the said specific strategy.

We consider the sub-Gaussian framework described in Section 1.3 and restrict our attention to the case when $\mu^\star$ is known. We provide a refinement of the results of Bubeck et al. [7, Section 3], already known by these authors themselves (see, e.g., Faure et al. [15]). The algorithm considered below is inspired by Algorithm 1 of Bubeck et al. [7]. For each $t \geqslant 1$ and $a \in \{1,\dots,K\}$ such that $N_a(t) \geqslant 1$, we denote by

$$\widehat{\mu}_{a,t} = \frac{1}{N_a(t)}\sum_{s=1}^{t} Y_s\,\mathbb{I}_{\{A_s=a\}}$$

the empirical mean of the rewards obtained between rounds 1 and t when playing arm a.

Algorithm 1: An algorithm with bounded regret, thanks to the knowledge of $\mu^\star$

Bandit problem: $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ where each ν_a is sub-Gaussian in the sense of (4)

Parameters: the value of $\mu^\star = \max_{a=1,\dots,K} \mu_a$

For: each $t \in \{1,\dots,K\}$, **do:** play arm t.

For: each round $t \geqslant K+1$,

1. Let $\mathcal{C}_t = \left\{ a \in \{1,\dots,K\} \ : \ \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\dfrac{4\ln N_a(t-1)}{N_a(t-1)}} \right\}$ be the set of candidate arms;
2. If $\mathcal{C}_t \neq \emptyset$, play an arm A_t at random in $\mathcal{C}_t$, update $t := t+1$;
3. If $\mathcal{C}_t = \emptyset$, play $A_t = 1$, $A_{t+1} = 2$, $\dots$, $A_{t+K} = t+K-1$, update $t := t+K$.

We use the notation introduced before (1), but, as indicated above, without the indexations in the considered strategy.

THEOREM 9. *For all bandit problems $\underline{\nu} = (\nu_a)_{a=1,\dots,K}$ where each distribution ν_a is sub-Gaussian in the sense of (4), the regret of the algorithm above is bounded by*

$$R_{\underline{\nu},T} \leqslant \sum_{a:\Delta_a>0} \left(\frac{36\ln(17/\Delta_a)}{\Delta_a} + 3\Delta_a \right).$$

Proof. We fix an optimal arm $a^\star$. In view of (1), it suffices to bound $\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]$ for each suboptimal arm a. Each arm is played once between 1 and K. For all $t \geqslant K+1$, a suboptimal arm a can only be played if $a \in \mathcal{C}_t$ (step 2 of the second for loop) or if we are in a sequence where each arm is played successfully (step 3 of the second for loop). In the latter case, the set of candidate arms at round $t-a+1$ was empty. It did not contain $a^\star$. This optimal arm is played also once in the sequence of pulls corresponding to step 3, at time $t-a+a^\star+1$. At time $t-a+a^\star$ we still had $N_{a^\star}(t-a+a^\star) = N_{a^\star}(t-a+1)$, so that the condition for being a candidate was violated as well:

$$\widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\ln N_a(t-a+a^\star)}{N_a(t-a+a^\star)}}\,.$$

All in all, we proved the inclusion: for $t \geqslant K+1$,

$$\begin{aligned}\{A_t = a\} \subseteq & \left\{ A_t = a \ \text{ and } \ \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\frac{4\ln N_a(t-1)}{N_a(t-1)}} \right\} \\ & \cup \left\{ A_{t-a+a^\star} = a^\star \ \text{ and } \ \widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\ln N_a(t-a+a^\star)}{N_a(t-a+a^\star)}} \right\}.\end{aligned}$$

We now only sketch the next argument, as we proceed similarly to all multi-armed bandit analyses, by resorting to Doob's optional sampling theorem, which asserts that the rewards Y_s obtained at those rounds s when $A_s = a$ are independent and identically distributed according to ν_a. We denote by $\overline{\mu}_{a,n}$ the empirical average of the first n rewards obtained by arm a during the game. Then,

$$\begin{aligned}\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] &\leqslant 1 + \sum_{t=K+1}^{T} \mathbb{P}\left\{ A_t = a \ \text{ and } \ \widehat{\mu}_{a,t-1} - \mu^\star > -\sqrt{\frac{4\ln N_a(t-1)}{N_a(t-1)}} \right\} \\ &\quad + \sum_{t=K+1}^{T} \mathbb{P}\left\{ A_{t-a+a^\star} = a^\star \ \text{ and } \ \widehat{\mu}_{a^\star,t-a+a^\star} - \mu^\star \leqslant -\sqrt{\frac{4\ln N_a(t-a+a^\star)}{N_a(t-a+a^\star)}} \right\} \\ &\leqslant 1 + \sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a,n} - \mu^\star > -\sqrt{\frac{4\ln n}{n}} \right\} + \sum_{n\geqslant 1} \mathbb{P}\left\{ \overline{\mu}_{a^\star,n} - \mu^\star \leqslant -\sqrt{\frac{4\ln n}{n}} \right\}. \qquad (29)\end{aligned}$$

As indicated already in Bubeck et al. [7], for each arm a, the sub-Gaussian assumption on ν_a, together with a Crámer–Chernoff bound, indicates that for all $n \geqslant 1$ and all $\varepsilon > 0$,

$$\max\Big\{\mathbb{P}\big\{\overline{\mu}_{a,n} - \mu_a \geqslant \varepsilon\big\},\ \mathbb{P}\big\{\overline{\mu}_{a,n} - \mu_a \leqslant -\varepsilon\big\}\Big\} \leqslant \exp\left(-\frac{n\varepsilon^2}{2}\right). \tag{30}$$

We substitute this inequality in the bound (29) obtained above. On the one hand, for $a^\star$,

$$\sum_{n\geqslant 1} \mathbb{P}\left\{\overline{\mu}_{a^\star,n} - \mu^\star \leqslant -\sqrt{\frac{4\ln n}{n}}\right\} \leqslant \sum_{n\geqslant 1} n^{-2} \leqslant 2\,. \tag{31}$$

On the other hand, for a, we rewrite $\mu^\star = \mu_a + \Delta_a$ and get

$$\sum_{n\geqslant 1} \mathbb{P}\left\{\overline{\mu}_{a,n} - \mu^\star > -\sqrt{\frac{4\ln n}{n}}\right\} = \sum_{n\geqslant 1} \mathbb{P}\left\{\overline{\mu}_{a,n} - \mu_a > \Delta_a - \sqrt{\frac{4\ln n}{n}}\right\}.$$

To upper bound the latter sum, we denote by n_0 the smallest integer $k \geqslant 3$, if it exists, such that:

$$\Delta_a - \sqrt{\frac{4\ln k}{k}} \geqslant \frac{\Delta_a}{2}\,, \qquad \text{that is,} \qquad \sqrt{\frac{4\ln k}{k}} \leqslant \frac{\Delta_a}{2}\,. \tag{32}$$

As $x \mapsto \sqrt{(\ln x)/x}$ is decreasing on $[3, +\infty)$, we have

$$\forall n \geqslant n_0, \qquad \Delta_a - \sqrt{\frac{4\ln n}{n}} \geqslant \frac{\Delta_a}{2}\,,$$

and thus

$$\sum_{n\geqslant 1} \mathbb{P}\left\{\overline{\mu}_{a,n} - \mu_a > \Delta_a - \sqrt{\frac{4\ln n}{n}}\right\} \leqslant n_0 - 1 + \sum_{n\geqslant n_0} \mathbb{P}\left\{\overline{\mu}_{a,n} - \mu_a > \frac{\Delta_a}{2}\right\}.$$

Note that the above inequality also holds with $n_0 = 2$ when no $k \geqslant 3$ satisfies (32). We use (30) and a comparison to an integral to get

$$\sum_{n\geqslant n_0} \mathbb{P}\left\{\overline{\mu}_{a,n} - \mu_a > \frac{\Delta_a}{2}\right\} \leqslant \sum_{n\geqslant n_0} \exp\left(-\frac{n\Delta_a^2}{8}\right) \leqslant \int_{n_0-1}^{+\infty} \exp\left(-\frac{x\Delta_a^2}{8}\right) \mathrm{d}x \leqslant \frac{8}{\Delta_a^2}\,.$$

Substituting the above bounds and (31) into (29), we showed so far that

$$\mathbb{E}_\nu\big[N_a(T)\big] \leqslant n_0 + 2 + \frac{8}{\Delta_a^2}\,.$$

The proof is concluded by upper bounding n_0, based on (32). If $\Delta_a \leqslant 4\sqrt{(\ln 3)/3}$, then the n_0 defined in (32) exists. In this case, we denote by $x_0 \in [3, +\infty)$ the real number such that

$$\sqrt{\frac{4\ln x_0}{x_0}} \leqslant \frac{\Delta_a}{2} \qquad \text{that is,} \qquad x_0 = \frac{16\ln x_0}{\Delta_a^2}\,.$$

We have $n_0 = \lceil x_0 \rceil \leqslant x_0 + 1$. Since

$$x_0 = \frac{16\ln x_0}{\Delta_a^2} = \frac{32\ln(4/\Delta)}{\Delta_a^2} + \frac{16}{\Delta_a^2}\ln\big(\ln x_0\big)\,,$$

we suspect that x_0 should not be too much larger than $32\ln(4/\Delta)/\Delta_a^2$. Indeed, using the inequality $\ln(u) \leqslant u$, we see that

$$x_0 = \frac{16\ln x_0}{\Delta_a^2} = \frac{160\ln x_0^{1/10}}{\Delta_a^2} \leqslant \frac{160\, x_0^{1/10}}{\Delta_a^2}, \qquad \text{thus} \qquad x_0 \leqslant \left(\frac{160}{\Delta_a^2}\right)^{10/9}.$$

Therefore,

$$x_0 = \frac{16\ln x_0}{\Delta_a^2} \leqslant \frac{16}{\Delta_a^2}\ln\left(\frac{160}{\Delta_a^2}\right)^{10/9} \leqslant \frac{16\times(10/9)\times 2}{\Delta_a^2}\ln\frac{13}{\Delta^2} \leqslant \frac{36}{\Delta_a^2}\ln\frac{13}{\Delta^2}.$$

When the n_0 defined in (32) does not exist and we take $n_0 = 2$, we may still bound n_0 by 1 plus the bound above on x_0 (as the latter is larger than 1). The theorem follows, after substitution of all the bounds, together with the inequality $8 \leqslant 36\ln(17) - 36\ln(13)$. □

Acknowledgments. The authors thank Sébastien Gerchinovitz and Vianney Perchet for stimulating discussions and comments. They are grateful to the anonymous associate editor and reviewers for their thoughtful feedback and remarks.

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.

References

[1] Ali, S. M., S. D. Silvey. 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B. Methodological* **28** 131–142.

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

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

[4] Auer, P., R. Ortner. 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* **61**(1) 55–65.

[5] Bubeck, S. 2010. Bandits games and clustering foundations. Ph.D. thesis, Université Lille 1, France.

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

[7] Bubeck, S., V. Perchet, P. Rigollet. 2013. Bounded regret in stochastic multi-armed bandits. *Proceedings of the 26th Annual Conference on Learning Theory* (COLT), *JMLR W&CP*, vol. 30. 122–134.

[8] Bubeck, S., V. Perchet, P. Rigollet. 2013. Erratum to [7]. URL http://research.microsoft.com/en-us/um/people/sebubeck/pub.html. "The proof of Theorem 8 is not correct. We do not know if the theorem holds true.".

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

[10] Calabro, Chris. 2009. The exponential complexity of satisfiability problems. Ph.D. thesis, University of California, San Diego.

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

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

[13] Combes, R., A. Proutière. 2014. Unimodal bandits without smoothness. ArXiv:1406.7447.

[14] Cowan, W., M.N. Katehakis. 2015. Asymptotically optimal sequential experimentation under generalized ranking. ArXiv:1510.02041.

[15] Faure, M., P. Gaillard, B. Gaujal, V. Perchet. 2015. Online learning and game theory. a quick overview with recent results and applications. *ESAIM: Proceedings and Surveys* **51** 246–271.

[16] Garivier, A., E. Kaufmann, T. Lattimore. 2016. On explore-then-commit strategies. D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, eds., *Advances in Neural Information Processing Systems 29* (NIPS 2016). Curran Associates, Inc., 784–792.

[17] Honda, J., A. Takemura. 2015. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research* **16**(Dec) 3721–3756.

[18] Jiang, C. 2015. Online advertisements and multi-armed bandits. Ph.D. thesis, University of Illinois at Urbana-Champaign, USA.

[19] Kaufmann, E., O. Capp, A. Garivier. 2016. On the complexity of best arm identication in multi-armed bandit models. *Journal of Machine Learning Research* **7** 1–42.

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

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

[22] Lehmann, E.L., G. Casella. 1998. *Theory of Point Estimation*. Springer.

[23] Mannor, S., J.N. Tsitsiklis. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* **5** 623–648.

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

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