A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences

Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille Nord Europe
SequeL Project
40 avenue Halley
59650 Villeneuve dAscq, France

Gilles Stoltz GILLES.STOLTZ@ENS.FR
Ecole Normale Supérieure, CNRS, INRIA
45 rue d'Ulm
75005 Paris, France
&
HEC Paris, CNRS
1 rue de la Libération
78351 Jouy-en-Josas, France

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

We consider a Kullback-Leibler-based algorithm for the stochastic multi-armed bandit problem in the case of distributions with finite supports (not necessarily known beforehand), whose asymptotic regret matches the lower bound of Burnetas and Katehakis (1996). Our contribution is to provide a finite-time analysis of this algorithm; we get bounds whose main terms are smaller than the ones of previously known algorithms with finite-time analyses (like UCB-type algorithms).

Keywords: Multi-armed bandit problem, finite-time analysis, Kullback-Leibler divergence, Sanov's lemma

1. Introduction

The *stochastic* multi-armed bandit problem, introduced by Robbins (1952), formalizes the problem of decision-making under uncertainty, and illustrates the fundamental tradeoff that appears between *exploration*, i.e., making decisions in order to improve the knowledge of the environment, and *exploitation*, i.e., maximizing the payoff.

Setting. In this paper, we consider a multi-armed bandit problem with *finitely* many arms indexed by $\mathcal{A}$, for which each arm $a \in \mathcal{A}$ is associated with an unknown and fixed probability distribution ν_a over $[0,1]$. The game is *sequential* and goes as follows: at each round $t \geqslant 1$, the player first picks an arm $A_t \in \mathcal{A}$ and then receives a stochastic payoff Y_t drawn at random according to ν_{A_t}. He only gets to see the payoff Y_t.

For each arm $a \in \mathcal{A}$, we denote by μ_a the expectation of its associated distribution ν_a and we let $a^\star$ be any optimal arm, i.e., $a^\star \in \operatorname*{argmax}_{a \in \mathcal{A}} \mu_a$.
We write $\mu^\star$ as a short-hand notation for the largest expectation $\mu_{a^\star}$ and denote the gap of the expected payoff μ_a of an arm $a \in \mathcal{A}$ to $\mu^\star$ as $\Delta_a = \mu^\star - \mu_a$. In addition, the number of times each arm $a \in \mathcal{A}$ is pulled between the rounds 1 and T is referred to as $N_T(a)$,

$$N_T(a) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}} \,.$$

The quality of a strategy will be evaluated through the standard notion of *expected regret*, which we recall now. The expected regret (or simply regret) at round $T \geqslant 1$ is defined as

$$R_T \stackrel{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{A_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \; \mathbb{E}\big[N_T(a)\big] \,, \tag{1}$$

where we used the tower rule for the first equality. Note that the expectation is with respect to the random draws of the Y_t according to the ν_{A_t} and also to the possible auxiliary randomizations that the decision-making strategy is resorting to.

The regret measures the cumulative loss resulting from pulling sub-optimal arms, and thus quantifies the amount of exploration required by an algorithm in order to find a best arm, since, as (1) indicates, the regret scales with the expected number of pulls of sub-optimal arms. Since the formulation of the problem by Robbins (1952) the regret has been a popular criterion for assessing the quality of a strategy.

Known lower bounds. Lai and Robbins (1985) showed that for some (one-dimensional) parametric classes of distributions, any consistent strategy (i.e., any strategy not pulling sub-optimal arms more than in a polynomial number of rounds) will despite all asymptotically pull in expectation any sub-optimal arm a at least

$$\mathbb{E}\big[N_T(a)\big] \geqslant \left(\frac{1}{\mathcal{K}(\nu_a, \nu^\star)} + o(1)\right) \log(T)$$

times, where $\mathcal{K}(\nu_a, \nu^\star)$ is the Kullback-Leibler (KL) divergence between ν_a and $\nu^\star$; it measures how close distributions ν_a and $\nu^\star$ are from a theoretical information perspective.

Later, Burnetas and Katehakis (1996) extended this result to some classes of multi-dimensional parametric distributions and proved the following generic lower bound: for a given family $\mathcal{P}$ of possible distributions over the arms,

$$\mathbb{E}\big[N_T(a)\big] \geqslant \left(\frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o(1)\right) \log(T) \,, \qquad \text{where} \quad \mathcal{K}_{\inf}(\nu_a, \mu^\star) \stackrel{\text{def}}{=} \inf_{\nu \in \mathcal{P} \colon E(\nu) > \mu^*} \mathcal{K}(\nu_a, \nu) \,,$$

with the notation $E(\nu)$ for the expectation of a distribution ν. The intuition behind this improvement is to be related to the goal that we want to achieve in bandit problems; it is not detecting whether a distribution is optimal or not (for this goal, the relevant quantity would be $\mathcal{K}(\nu_a, \nu^\star)$), but rather achieving the optimal rate of reward $\mu^\star$ (i.e., one needs to measure how close ν_a is to any distribution $\nu \in \mathcal{P}$ whose expectation is at least $\mu^\star$).

Known upper bounds. Lai and Robbins (1985) provided an algorithm based on the KL divergence, which has been extended by Burnetas and Katehakis (1996) to an algorithm based on $\mathcal{K}_{\inf}$; it is asymptotically optimal since the number of pulls of any sub-optimal arm a satisfies

$$\mathbb{E}\big[N_T(a)\big] \leqslant \left(\frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + o(1)\right)\log(T)\,.$$

This result holds for finite-dimensional parametric distributions under some assumptions, e.g., the distributions having a finite and known support or belonging to a set of Gaussian distributions with known variance. Recently Honda and Takemura (2010a) extended this asymptotic result to the case of distributions $\mathcal{P}$ with support in $[0, 1]$ and such that $\mu^* < 1$; the key ingredient in this case is that $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$ is equal to

$$\mathcal{K}_{\min}(\nu_a, \mu^\star) \overset{\text{def}}{=} \inf_{\nu \in \mathcal{P} : E(\nu) \geqslant \mu^*} \mathcal{K}(\nu_a, \nu).$$

Motivation. All the results mentioned above provide asymptotic bounds only. However, any algorithm is only used for a finite number of rounds and it is thus essential to provide a finite-time analysis of its performance. Auer et al. (2002) initiated this work by providing an algorithm (UCB1) based on a Chernoff-Hoeffding bound; it pulls any sub-optimal arm, till any time T, at most $(8/\Delta_a^2)\log T + 1 + \pi^2/3$ times, in expectation. Although this yields a logarithmic regret, the multiplicative constant depends on the gap $\Delta_a^2 = (\mu^\star - \mu_a)^2$ but not on $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$, which can be seen to be larger than $\Delta_a^2/2$ by Pinsker's inequality; that is, this non-asymptotic bound does not have the right dependence in the distributions. (How much is gained of course depends on the specific families of distributions at hand.) Audibert et al. (2009) provided an algorithm (UCB-V) that takes into account the empirical variance of the arms and exhibited a strategy such that $\mathbb{E}\big[N_T(a)\big] \leqslant 10(\sigma_a^2/\Delta_a^2 + 2/\Delta_a)\log T$ for any time T (where σ_a^2 is the variance of arm a); it improves over UCB1 in case of arms with small variance. Other variants include the MOSS algorithm by Audibert and Bubeck (2010) and Improved UCB by Auer and Ortner (2010).

However, all these algorithms only rely on one moment (for UCB1) or two moments (for UCB-V) of the empirical distributions of the obtained rewards; they do not fully exploit the empirical distributions. As a consequence, the resulting bounds are expressed in terms of the means μ_a and variances σ_a^2 of the sub-optimal arms and not in terms of the quantity $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$ appearing in the lower bounds. The numerical experiments reported in Filippi (2010) confirm that these algorithms are less efficient than those based on $\mathcal{K}_{\inf}$.

Our contribution. In this paper we analyze a $\mathcal{K}_{\inf}$-based algorithm inspired by the ones studied in Lai and Robbins (1985); Burnetas and Katehakis (1996); Filippi (2010); it indeed takes into account the full empirical distribution of the observed rewards. The analysis is performed (with explicit bounds) in the case of Bernoulli distributions over the arms. Less explicit but finite-time bounds are obtained in the case of finitely supported distributions (whose supports do not need to be known in advance). Finally, we pave the way for handling the case of general finite-dimensional parametric distributions. These results improve on the ones by Burnetas and Katehakis (1996); Honda and Takemura (2010a) since finite-time bounds (implying their asymptotic results) are obtained; and on Auer et al. (2002); Audibert et al. (2009) as the dependency of the main term scales with $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$. The proposed $\mathcal{K}_{\inf}$-based algorithm is also more natural and more appealing than the one presented in Honda and Takemura (2010a).

Recent related works. Since our initial submission of the present paper, we got aware of two papers that tackle problems similar to ours. First, a revised version of Honda and Takemura (2010b, personal communication) obtains finite-time regret bounds (with prohibitively large constants) for a *randomized* (less natural) strategy in the case of distributions with finite supports (also not known in advance). Second, another paper at this conference (Garivier and Cappé, 2011) also deals with the $\mathcal{K}$–strategy which we study in Theorem 3; they however do not obtain second-order terms in closed forms as we do and later extend their strategy to exponential families of distributions (while we extend our strategy to the case of distributions with finite supports). On the other hand, they show how the $\mathcal{K}$–strategy can be extended in a straightforward manner to guarantee bounds with respect to the family of all bounded distributions on a known interval; these bounds are suboptimal but improve on the ones of UCB-type algorithms.

2. Definitions and tools

Let $\mathcal{X}$ be a Polish space; in the next sections, we will consider $\mathcal{X} = \{0,1\}$ or $\mathcal{X} = [0,1]$. We denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$ and equip $\mathcal{P}(\mathcal{X})$ with the distance d induced by the norm $\|\cdot\|$ defined by $\|\nu\| = \sup_{f\in\mathcal{L}} \left|\int_{\mathcal{X}} f \,\mathrm{d}\nu\right|$, where $\mathcal{L}$ is the set of Lipschitz functions over $\mathcal{X}$, taking values in $[-1,1]$ and with Lipschitz constant smaller than 1.

Kullback-Leibler divergence: For two elements ν, $\kappa \in \mathcal{P}(\mathcal{X})$, we write $\nu \ll \kappa$ when ν is absolutely continuous with respect to κ and denote in this case by $\mathrm{d}\nu/\mathrm{d}\kappa$ the density of ν with respect to κ. We recall that the Kullback-Leibler divergence between ν and κ is defined as

$$\mathcal{K}(\nu,\kappa) = \int_{[0,1]} \frac{\mathrm{d}\nu}{\mathrm{d}\kappa} \log \frac{\mathrm{d}\nu}{\mathrm{d}\kappa}\,\mathrm{d}\kappa \quad \text{if } \nu \ll \kappa; \qquad \text{and} \quad \mathcal{K}(\nu,\kappa) = +\infty \quad \text{otherwise.} \tag{2}$$

Empirical distribution: We consider a sequence $X_1, X_2, \ldots$ of random variables taking values in $\mathcal{X}$, independent and identically distributed according to a distribution ν. For all integers $t \geqslant 1$, we denote the empirical distribution corresponding to the first t elements of the sequence by

$$\widehat{\nu}_t = \frac{1}{t}\sum_{s=1}^{t} \delta_{X_t}\,.$$

Non-asymptotic Sanov's Lemma: The following lemma follows from a straightforward adaptation of Dinwoodie (1992, Theorem 2.1 and comments on page 372). Details of the proof are provided in the extended version (Maillard et al., 2011) of the present paper.

Lemma 1 *Let $\mathcal{C}$ be an open convex subset of $\mathcal{P}(\mathcal{X})$ such that* $\Lambda(\mathcal{C}) = \inf\limits_{\kappa\in\mathcal{C}} \mathcal{K}(\kappa,\nu) < \infty\,.$
Then, for all $t \geqslant 1$, one has $\mathbb{P}_\nu\bigl\{\widehat{\nu}_t \in \overline{\mathcal{C}}\bigr\} \leqslant e^{-t\Lambda(\overline{\mathcal{C}})}$ *where $\overline{\mathcal{C}}$ is the closure of $\mathcal{C}$.*

This lemma should be thought of as a deviation inequality. The empirical distribution converges (in distribution) to ν. Now, if (and only if) ν is not in the closure of $\mathcal{C}$, then $\Lambda(\mathcal{C}) > 0$ and the lemma indicates how unlikely it is that $\widehat{\nu}_t$ is in this set $\overline{\mathcal{C}}$ not containing the limit ν. The probability of interest decreases at a geometric rate, which depends on $\Lambda(\mathcal{C})$.

3. Finite-time analysis for Bernoulli distributions

In this section, we start with the case of Bernoulli distributions. Although this case is a special case of the general results of Section 4, we provide here a complete and self-contained analysis of this case, where, in addition, we are able to provide closed forms for all the terms in the regret bound. Note however that the resulting bound is slightly worse than what could be derived from the general case (for which more sophisticated tools are used). This result is mainly provided as a warm-up.

3.1. Reminder of some useful results for Bernoulli distributions

We denote by $\mathcal{B}$ the subset of $\mathcal{P}\bigl([0,1]\bigr)$ formed by the Bernoulli distributions; it corresponds to $\mathcal{B} = \mathcal{P}\bigl(\{0,1\}\bigr)$. A generic element of $\mathcal{B}$ will be denoted by $\beta(p)$, where $p \in [0,1]$ is the probability mass put on 1. We consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables, with common distribution $\beta(p)$; for the sake of clarity we will index, in this subsection only, all probabilities and expectations with p.

For all integers $t \geqslant 1$, we denote by $\displaystyle \widehat{p}_t = \frac{1}{t}\sum_{s=1}^{t} X_t$ the empirical average of the first t elements of the sequence.

The lemma below follows from an adaptation of Garivier and Leonardi (2011, Proposition 2).

Lemma 2 *For all $p \in [0,1]$, all $\varepsilon > 1$, and all $t \geqslant 1$,*

$$\mathbb{P}_p\!\left(\bigcup_{s=1}^{t}\left\{s\,\mathcal{K}\Bigl(\beta\bigl(\widehat{p}_s\bigr),\,\beta(p)\Bigr) \geqslant \varepsilon\right\}\right) \leqslant 2e\,\lceil \varepsilon \log t\rceil\, e^{-\varepsilon}\,.$$

In particular, for all random variables N_t taking values in $\{1,\ldots,t\}$,

$$\mathbb{P}_p\left\{N_t\,\mathcal{K}\Bigl(\beta\bigl(\widehat{p}_{N_t}\bigr),\,\beta(p)\Bigr) \geqslant \varepsilon\right\} \leqslant 2e\,\lceil \varepsilon \log t\rceil\, e^{-\varepsilon}\,.$$

Another immediate fact about Bernoulli distributions is that for all $p \in (0,1)$, the mappings $\mathcal{K}_{\,\cdot\,,p} : q \in (0,1) \mapsto \mathcal{K}\bigl(\beta(p),\beta(q)\bigr)$ and $\mathcal{K}_{p,\,\cdot} : q \in [0,1] \mapsto \mathcal{K}\bigl(\beta(q),\beta(p)\bigr)$ are continuous and take finite values. In particular, we have, for instance, that for all $\varepsilon > 0$ and $p \in (0,1)$, the set

$$\Bigl\{q \in [0,1] : \quad \mathcal{K}\bigl(\beta(p),\beta(q)\bigr) \leqslant \varepsilon\Bigr\}$$

is a closed interval containing p. This property still holds when $p \in \{0,1\}$, as in this case, the interval is reduced to $\{p\}$.

3.2. Strategy and analysis

We consider the so-called $\mathcal{K}$*–strategy* of Figure 1, which was already considered in the literature, see Burnetas and Katehakis (1996); Filippi (2010). The numerical computation of the quantities $B^{+}_{a,t}$ is straightforward (by convexity of $\mathcal{K}$ in its second argument, by using iterative methods) and is detailed therein.

Before proceeding, we denote by $\sigma_a^2 = \mu_a(1-\mu_a)$ the variance of each arm $a \in \mathcal{A}$ (and take the short-hand notation $\sigma^{\star,2}$ for the variance of an optimal arm).

```
Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
Initialization: Pull each arm of $\mathcal{A}$ once
For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,
  – compute for each arm $a \in \mathcal{A}$ the quantity
        $B^+_{a,t} = \max\Bigl\{ q \in [0,1] : \;\; N_t(a)\, \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\, \beta(q)\Bigr) \leqslant f(t) \Bigr\}$,
    where   $\widehat{\mu}_{a,N_t(a)} = \dfrac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} Y_s$;
  – in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;
  – pull any arm $A_{t+1} \in \mathop{\mathrm{argmax}}_{a \in \mathcal{A}} B^+_{a,t}$.
```

Figure 1: The $\mathcal{K}$–strategy.

Theorem 3 *When $\mu^\star \in (0,1)$, for all non-decreasing functions $f : \mathbb{N} \to \mathbb{R}_+$ such that $f(1) \geqslant 1$, the expected regret R_T of the strategy of Figure 1 is upper bounded by the infimum, as the $(c_a)_{a\in\mathcal{A}}$ describe $(0,+\infty)$, of the quantities*

$$\sum_{a\in\mathcal{A}} \Delta_a \left(\frac{(1+c_a)\, f(T)}{\mathcal{K}\bigl(\beta(\mu_a),\, \beta(\mu^\star)\bigr)} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \bigl\lceil f(t) \log t \bigr\rceil\, e^{-f(t)} + \frac{(1+c_a)^2}{8\, c_a^2 \Delta_a^2\, \min\bigl\{\sigma_a^4,\, \sigma^{\star,4}\bigr\}} \mathbb{I}_{\{\mu_a \in (0,1)\}} + 3 \right).$$

For $\mu^\star = 0$, its regret is null. For $\mu^\star = 1$, it satisfies $R_T \leqslant 2\bigl(|\mathcal{A}| - 1\bigr)$.

A possible choice for the function f is $f(t) = \log\bigl((et)\log^3(et)\bigr)$, which is non decreasing, satisfies $f(1) \geqslant 1$, and is such that the second term in the sum above is bounded (by a basic result about so-called Bertrand's series). Now, as the constants c_a in the bound are parameters of the analysis (and not of the strategy), they can be optimized. For instance, with the choice of $f(t)$ mentioned above, taking each c_a proportional to $(\log T)^{-1/3}$ (up to a multiplicative constant that depends on the distributions ν_a) entails the regret bound

$$\sum_{a\in\mathcal{A}} \Delta_a \frac{\log T}{\mathcal{K}\bigl(\beta(\mu_a),\, \beta(\mu^\star)\bigr)} + \varepsilon_T\,,$$

where it is easy to give an explicit and closed-form expression of ε_T; in this conference version, we only indicate that ε_T is of order of $(\log T)^{2/3}$ but we do not know whether the order of magnitude of this second-order term is optimal.

Proof We first deal with the case where $\mu^\star \not\in \{0,1\}$ and introduce an additional notation. In view of the remark at the end of Section 3.1, for all arms a and rounds t, we let $B^-_{a,t}$ be the element in $[0,1]$ such that

$$\Bigl\{ q \in [0,1] : \;\; N_t(a)\, \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\, \beta(q)\Bigr) \leqslant f(t) \Bigr\} = \bigl[B^-_{a,t},\; B^+_{a,t}\bigr]\,. \tag{3}$$

As (1) indicates, it suffices to bound $N_T(a)$ for all suboptimal arms a, i.e., for all arms such that $\mu_a < \mu^\star$. We will assume in addition that $\mu_a > 0$ (and we also have $\mu_a \leqslant \mu^\star < 1$); the case where $\mu_a = 0$ will be handled separately.

Step 1: A decomposition of the events of interest. For $t \geqslant |\mathcal{A}|$, when $A_{t+1} = a$, we have in particular, by definition of the strategy, that $B^+_{a,t} \geqslant B^+_{a^\star,t}$. On the event

$$\bigl\{A_{t+1} = a\bigr\} \cap \Bigl\{\mu^\star \in \bigl[B^-_{a^\star,t},\ B^+_{a^\star,t}\bigr]\Bigr\} \cap \Bigl\{\mu_a \in \bigl[B^-_{a,t},\ B^+_{a,t}\bigr]\Bigr\}\,,$$

we therefore have, on the one hand, $\mu^\star \leqslant B^+_{a^\star,t} \leqslant B^+_{a,t}$ and on the other hand, $B^-_{a,t} \leqslant \mu_a \leqslant \mu^\star$, that is, the considered event is included in $\Bigl\{\mu^\star \in \bigl[B^-_{a,t},\ B^+_{a,t}\bigr]\Bigr\}$. We thus proved that

$$\bigl\{A_{t+1} = a\bigr\} \subseteq \Bigl\{\mu^\star \not\in \bigl[B^-_{a^\star,t},\ B^+_{a^\star,t}\bigr]\Bigr\} \cup \Bigl\{\mu_a \not\in \bigl[B^-_{a,t},\ B^+_{a,t}\bigr]\Bigr\} \cup \Bigl\{\mu^\star \in \bigl[B^-_{a,t},\ B^+_{a,t}\bigr]\Bigr\}\,.$$

Going back to the definition (3), we get in particular the inclusion

$$\begin{aligned}\bigl\{A_{t+1} = a\bigr\} \subseteq\ & \Bigl\{N_t(a^\star)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a^\star,N_t(a^\star)}\bigr),\ \beta(\mu^\star)\Bigr) > f(t)\Bigr\} \\ & \cup \Bigl\{N_t(a)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\ \beta(\mu_a)\Bigr) > f(t)\Bigr\} \\ & \cup \biggl(\Bigl\{N_t(a)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\ \beta(\mu^\star)\Bigr) \leqslant f(t)\Bigr\} \cap \bigl\{A_{t+1} = a\bigr\}\biggr)\,.\end{aligned}$$

Step 2: Bounding the probabilities of two elements of the decomposition. We consider the filtration $(\mathcal{F}_t)$, where for all $t \geqslant 1$, the σ–algebra $\mathcal{F}_t$ is generated by $A_1, Y_1, \ldots,$ A_t, Y_t. In particular, A_{t+1} and thus all $N_{t+1}(a)$ are $\mathcal{F}_t$–measurable. We denote by $\tau_{a,1}$ the deterministic round at which a was pulled for the first time and by $\tau_{a,2}$, $\tau_{a,3}$, … the rounds $t \geqslant |\mathcal{A}| + 1$ at which a was then played; since for all $k \geqslant 2$,

$$\tau_{a,k} = \min\bigl\{t \geqslant |\mathcal{A}| + 1 : \quad N_t(a) = k\bigr\}\,,$$

we see that $\bigl\{\tau_{a,k} = t\bigr\}$ is $\mathcal{F}_{t-1}$–measurable. Therefore, for each $k \geqslant 1$, the random variable $\tau_{a,k}$ is a (predictable) stopping time. Hence, by a well-known fact in probability theory (see, e.g., Chow and Teicher 1988, Section 5.3), the random variables $\widetilde{X}_{a,k} = Y_{\tau_{a,k}}$, where $k = 1, 2, \ldots$ are independent and identically distributed according to ν_a. Since on $\bigl\{N_t(a) = k\bigr\}$, we have the rewriting

$$\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k} \qquad \text{where} \qquad \widetilde{\mu}_{a,k} = \frac{1}{k}\sum_{j=1}^{k} \widetilde{X}_{a,j}\,,$$

and since for $t \geqslant |\mathcal{A}| + 1$, one has $N_t(a) \geqslant 1$ with probability 1, we can apply the second statement in Lemma 2 and get, for all $t \geqslant |\mathcal{A}| + 1$,

$$\mathbb{P}\Bigl\{N_t(a)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\ \beta(\mu_a)\Bigr) > f(t)\Bigr\} \leqslant 2e\,\bigl\lceil f(t)\log t\bigr\rceil\, e^{-f(t)}\,.$$

A similar argument shows that for all $t \geqslant |\mathcal{A}|+1$,

$$\mathbb{P}\Bigl\{ N_t(a^\star)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a^\star,N_t(a^\star)}\bigr),\,\beta(\mu^\star)\Bigr) > f(t) \Bigr\} \leqslant 2e \left\lceil f(t)\log t \right\rceil e^{-f(t)}\,.$$

Step 3: Rewriting the remaining terms. We therefore proved that

$$\begin{aligned}\mathbb{E}\bigl[N_T(a)\bigr] \leqslant 1 + 4e \sum_{t=|\mathcal{A}|}^{T-1} \left\lceil f(t)\log t \right\rceil e^{-f(t)} \\ + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\biggl(\Bigl\{ N_t(a)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\,\beta(\mu^\star)\Bigr) \leqslant f(t) \Bigr\} \cap \bigl\{A_{t+1}=a\bigr\} \biggr)\end{aligned}$$

and deal now with the last sum. Since f is non decreasing, it is bounded by

$$\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl(K_t \cap \bigl\{A_{t+1}=a\bigr\}\Bigr) \qquad \text{where} \qquad K_t = \Bigl\{ N_t(a)\ \mathcal{K}\Bigl(\beta\bigl(\widehat{\mu}_{a,N_t(a)}\bigr),\,\beta(\mu^\star)\Bigr) \leqslant f(T) \Bigr\}.$$

Now, $\displaystyle\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl(K_t \cap \bigl\{A_{t+1}=a\bigr\}\Bigr) = \mathbb{E}\left[\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{I}_{\left\{A_{t+1}=a\right\}} \mathbb{I}_{K_t} \right] = \mathbb{E}\left[\sum_{k \geqslant 2} \mathbb{I}_{\left\{\tau_{a,k} \leqslant T\right\}} \mathbb{I}_{K_{\tau_{a,k}-1}} \right].$
We note that, since $N_{\tau_{a,k}-1}(a) = k-1$, we have that

$$K_{\tau_{a,k}-1} = \Bigl\{ (k-1)\ \mathcal{K}\Bigl(\beta\bigl(\widetilde{\mu}_{a,k-1}\bigr),\,\beta(\mu^\star)\Bigr) \leqslant f(T) \Bigr\}.$$

All in all, since $\tau_{a,k} \leqslant T$ implies $k \leqslant T - |\mathcal{A}| + 1$ (as each arm is played at least once during the first $|\mathcal{A}|$ rounds), we have

$$\mathbb{E}\left[\sum_{k \geqslant 2} \mathbb{I}_{\left\{\tau_{a,k} \leqslant T\right\}} \mathbb{I}_{K_{\tau_{a,k}-1}} \right] \leqslant \mathbb{E}\left[\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{I}_{K_{\tau_{a,k}-1}} \right] = \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Bigl\{ (k-1)\ \mathcal{K}\Bigl(\beta\bigl(\widetilde{\mu}_{a,k-1}\bigr),\,\beta(\mu^\star)\Bigr) \leqslant f(T) \Bigr\}. \tag{4}$$

Step 4: Bounding the probabilities of the latter sum via Sanov's lemma. For each $\gamma > 0$, we define the convex open set $\mathcal{C}_\gamma = \Bigl\{ \beta(q) \in \mathcal{B} :\ \ \mathcal{K}\bigl(\beta(q),\,\beta(\mu^\star)\bigr) < \gamma \Bigr\}$, which is a non-empty set (since $\mu^\star < 1$); by continuity of the mapping $\mathcal{K}_{\,\cdot\,,\mu^\star}$ defined after the statement of Lemma 2 when $\mu^\star \in (0,1)$, its closure is $\overline{\mathcal{C}}_\gamma = \Bigl\{ \beta(q) \in \mathcal{B} :\ \ \mathcal{K}\bigl(\beta(q),\,\beta(\mu^\star)\bigr) \leqslant \gamma \Bigr\}$.

In addition, since $\mu_a \in (0,1)$, we have that $\mathcal{K}\bigl(\beta(q),\,\beta(\mu_a)\bigr) < \infty$ for all $q \in [0,1]$. In particular, for all $\gamma > 0$, the condition $\Lambda\bigl(\mathcal{C}_\gamma\bigr) < \infty$ of Lemma 1 is satisfied. Denoting this value by

$$\theta_a(\gamma) = \inf\Bigl\{ \mathcal{K}\bigl(\beta(q),\,\beta(\mu_a)\bigr) :\ \ \beta(q) \in \mathcal{B}\ \text{ such that }\ \mathcal{K}\bigl(\beta(q),\,\beta(\mu^\star)\bigr) \leqslant \gamma \Bigr\},$$

we get by the indicated lemma that for all $k \geqslant 1$,

$$\mathbb{P}\Bigl\{ \mathcal{K}\Bigl(\beta\bigl(\widetilde{\mu}_{a,k}\bigr),\,\beta(\mu^\star)\Bigr) \leqslant \gamma \Bigr\} = \mathbb{P}\Bigl\{ \beta\bigl(\widetilde{\mu}_{a,k}\bigr) \in \overline{\mathcal{C}}_\gamma \Bigr\} \leqslant e^{-k\,\theta_a(\gamma)}\,.$$

Now, since (an open neighborhood of) $\beta(\mu_a)$ is not included in $\overline{\mathcal{C}}_\gamma$ as soon as $0 < \gamma < \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)$, we have that $\theta_a(\gamma) > 0$ for such values of γ. To apply the obtained inequality to the last sum in (4), we fix a constant $c_a > 0$ and denote by k_0 the following upper integer part, $k_0 = \left\lceil \dfrac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} \right\rceil$, so that $f(T)/k \leqslant \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a) < \mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)$ for $k \geqslant k_0$, hence,

$$\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{(k-1)\, \mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k-1}\big),\, \beta(\mu^\star)\Big) \leqslant f(T)\Big\} \;\leqslant\; \sum_{k=1}^{T} \mathbb{P}\Big\{\mathcal{K}\Big(\beta\big(\widetilde{\mu}_{a,k}\big),\, \beta(\mu^\star)\Big) \leqslant \frac{f(T)}{k}\Big\}$$

$$\leqslant\; k_0 - 1 + \sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(f(T)/k\big)\Big)\,.$$

Since θ_a is a non-increasing function,

$$\sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(f(T)/k\big)\Big) \;\leqslant\; \sum_{k=k_0}^{T} \exp\Big(-k\,\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big)$$

$$\leqslant\; \Gamma_a(c_a)\, \exp\Big(-k_0\,\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big) \leqslant \Gamma_a(c_a),$$

where $\Gamma_a(c_a) = \Big[1 - \exp\Big(-\theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)\Big)\Big]^{-1}$.
Putting all pieces together, we thus proved so far that

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \frac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \big\lceil f(t) \log t \big\rceil\, e^{-f(t)} + \Gamma_a(c_a)$$

and it only remains to deal with $\Gamma_a(c_a)$.

Step 5: Getting an upper bound in closed form for $\Gamma_a(c_a)$. We will make repeated uses of Pinsker's inequality: for $p, q \in [0,1]$, one has $\mathcal{K}\big(\beta(p), \beta(q)\big) \geqslant 2\,(p-q)^2$. In what follows, we use the short-hand notation $\Theta_a = \theta_a\big(\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)/(1+c_a)\big)$ and therefore need to upper bound $1/\big(1-e^{-\Theta_a}\big)$. Since for all $u \geqslant 0$, one has $e^{-u} \leqslant 1-u+u^2/2$, we get $\Gamma_a(c_a) \leqslant \dfrac{1}{\Theta_a\big(1-\Theta_a/2\big)} \leqslant \dfrac{2}{\Theta_a}$ for $\Theta_a \leqslant 1$, and $\Gamma_a(c_a) \leqslant \dfrac{1}{1-e^{-1}} \leqslant 2$ for $\Theta_a \geqslant 1$. It thus only remains to lower bound Θ_a in the case when it is smaller than 1.

By the continuity properties of the Kullback-Leibler divergence, the infimum in the definition of θ_a is always achieved; we therefore let $\widetilde{\mu}$ be an element in $[0,1]$ such that

$$\Theta_a = \mathcal{K}\big(\beta(\widetilde{\mu}),\, \beta(\mu_a)\big) \qquad \text{and} \qquad \mathcal{K}\big(\beta(\widetilde{\mu}),\, \beta(\mu^\star)\big) = \frac{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)}{1+c}\,;$$

it is easy to see that we have the ordering $\mu_a < \widetilde{\mu} < \mu^\star$. By Pinsker's inequality, $\Theta_a \geqslant 2\big(\widetilde{\mu}-\mu_a\big)^2$ and we now lower bound the latter quantity. We use the short-hand notation $f(p) = \mathcal{K}\big(\beta(p), \beta(\mu^\star)\big)$ and note that the thus defined mapping f is convex and differentiable on $(0,1)$; its derivative equals $f'(p) = \log\big((1-\mu^\star)/(\mu^\star)\big) + \log\big(p/(1-p)\big)$ for all $p \in (0,1)$ and is therefore non positive for $p \leqslant \mu^\star$. By the indicated convexity of f, using a sub-gradient inequality, we get $f\bigl(\widetilde{\mu}\bigr) - f(\mu_a) \geqslant f'(\mu_a)\,\bigl(\widetilde{\mu} - \mu_a\bigr)$, which entails, since $f'(\mu_a) < 0$,

$$\widetilde{\mu} - \mu_a \geqslant \frac{f\bigl(\widetilde{\mu}\bigr) - f(\mu_a)}{f'(\mu_a)} = \frac{c_a}{1+c_a}\;\frac{f(\mu_a)}{-f'(\mu_a)}\,, \tag{5}$$

where the equality follows from the fact that by definition of μ, we have $f\bigl(\widetilde{\mu}\bigr) = f(\mu_a)/(1+c_a)$. Now, since f' is differentiable as well on $(0,1)$ and takes the value 0 at $\mu^\star$, a Taylor's equality entails that there exists a $\xi \in (\mu_a, \mu^\star)$ such that

$$-f'(\mu_a) = f'(\mu^\star) - f'(\mu_a) = f''(\xi)\,\bigl(\mu^\star - \mu_a\bigr) \;\text{ where }\; f''(\xi) = 1/\xi + 1/(1-\xi) = 1/\bigl(\xi(1-\xi)\bigr).$$

Therefore, by convexity of $\tau \mapsto \tau(1-\tau)$, we get that

$$\frac{1}{-f'(\mu_a)} \geqslant \frac{\min\bigl\{\mu_a(1-\mu_a),\, \mu^\star(1-\mu^\star)\bigr\}}{\mu^\star - \mu_a}\,.$$

Substituting this into (5) and using again Pinsker's inequality to lower bound $f(\mu_a)$, we have proved

$$\widetilde{\mu} - \mu_a \geqslant 2\,\frac{c_a}{1+c_a}\,\bigl(\mu^\star - \mu_a\bigr)\,\min\bigl\{\mu_a(1-\mu_a),\, \mu^\star(1-\mu^\star)\bigr\}\,.$$

Putting all pieces together, we thus proved that

$$\Gamma_a(c_a) \leqslant 2\,\max\left\{\frac{(1+c_a)^2}{8\,c_a^2\bigl(\mu^\star-\mu_a\bigr)^2\Bigl(\min\bigl\{\mu_a(1-\mu_a),\,\mu^\star(1-\mu^\star)\bigr\}\Bigr)^2},\; 1\right\}\,;$$

bounding the maximum of the two quantities by their sum concludes the main part of the proof.

Step 6: For $\mu^\star \in \{0,1\}$ and/or $\mu_a = 0$. When $\mu^\star = 1$, then $\widehat{\mu}_{a^\star,N_t(a\star)} = 1$ for all $t \geqslant |\mathcal{A}|+1$, so that $B^+_{a^\star,t} = 1$ for all $t \geqslant |\mathcal{A}|+1$. Thus, the arm a is played after round $t \geqslant |\mathcal{A}|+1$ only if $B^+_{a,t} = 1$ and $\widehat{\mu}_{a,N_t(a)} = 1$ (in view of the tie-breaking rule of the considered strategy). But this means that a is played as long as it gets payoffs equal to 1 and is stopped being played when it receives the payoff 0 for the first time. Hence, in this case, we have that the sum of payoffs equals at least $T - 2\bigl(|\mathcal{A}|-1\bigr)$ and the regret $R_T = \mathbb{E}[T\mu^\star - (Y_1 + \ldots + Y_t)]$ is therefore bounded by $2\bigl(|\mathcal{A}|-1\bigr)$.

When $\mu^\star = 0$, a Dirac mass over 0 is associated with all arms and the regret of all strategies is equal to 0.

We consider now the case $\mu^\star \in (0,1)$ and $\mu_a = 0$, for which the first three steps go through; only in the upper bound of step 4 we used the fact that $\mu_a > 0$. But in this case, we have a deterministic bound on (4). Indeed, since $\mathcal{K}\bigl(\beta(0), \beta(\mu^\star)\bigr) = -\log \mu^\star$, we have $k\,\mathcal{K}\bigl(\beta(0), \beta(\mu^\star)\bigr) \leqslant f(T)$ if and only if

$$k \leqslant \frac{f(T)}{-\log \mu^\star} = \frac{f(T)}{\mathcal{K}\bigl(\beta(\mu_a), \beta(\mu^\star)\bigr)}\,,$$

which improves on the general bound exhibited in step 4. ■

Remark 4 Note that Step 5 in the proof is specifically designed to provide an upper bound on $\Gamma_a(c_a)$ in the case of Bernoulli distributions. In the general case, getting such an explicit bound seems more involved.

4. A finite-time analysis in the case of distributions with finite support

Before stating and proving our main result, Theorem 10, we introduce the quantity $\mathcal{K}_{\inf}$ and list some of its properties.

4.1. Some useful properties of $\mathcal{K}_{\inf}$ and its level sets

We now introduce the key quantity in order to generalize the previous algorithm to handle the case of distributions with finite support. To that end, we introduce $\mathcal{P}_F\big([0,1]\big)$, the subset of $\mathcal{P}\big([0,1]\big)$ that consists of distributions with finite support.

Definition 5 *For all distributions $\nu \in \mathcal{P}_F\big([0,1]\big)$ and $\mu \in [0,1)$, we define*

$$\mathcal{K}_{\inf}(\nu,\mu) = \inf\Big\{\mathcal{K}(\nu,\nu') : \quad \nu' \in \mathcal{P}_F\big([0,1]\big) \ \ \text{s.t.} \ \ E(\nu') > \mu\Big\},$$

where $E(\nu') = \int_{[0,1]} x \,\mathrm{d}\nu'(x)$ denotes the expectation of the distribution ν'.

We now remind some useful properties of $\mathcal{K}_{\inf}$. Honda and Takemura (2010b, Lemma 6) can be reformulated in our context as follows.

Lemma 6 *For all $\nu \in \mathcal{P}_F\big([0,1]\big)$, the mapping $\mathcal{K}_{\inf}(\nu,\,\cdot\,)$ is continuous and non decreasing in its argument $\mu \in [0,1)$. Moreover, the mapping $\mathcal{K}_{\inf}(\,\cdot\,,\mu)$ is lower semi-continuous on $\mathcal{P}_F\big([0,1]\big)$ for all $\mu \in [0,1)$.*

The next two lemmas bound the variation of $\mathcal{K}_{\inf}$, respectively in its first and second arguments. (For clarity, we denote the expectations with respect to ν by $\mathbb{E}_\nu$.) Their proofs can be found in the extended version of the present conference paper (Maillard et al., 2011). We denote by $\|\cdot\|_1$ the ℓ^1–norm on $\mathcal{P}\big([0,1]\big)$ and recall that the ℓ^1–norm of $\nu-\nu'$ corresponds to twice the distance in variation between ν and ν'.

Lemma 7 *For all $\mu \in (0,1)$ and for all ν, $\nu' \in \mathcal{P}_F\big([0,1]\big)$, the following holds true.*

- *In the case when $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] > 1$, then $\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \leqslant M_{\nu,\mu}\,\|\nu-\nu'\|_1$, for some constant $M_{\nu,\mu} > 0$.*
- *In the case when $\mathbb{E}_\nu\big[(1-\mu)/(1-X)\big] \leqslant 1$, the fact that $\mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu',\mu) \geqslant \alpha\,\mathcal{K}_{\inf}(\nu,\mu)$ for some $\alpha \in (0,1)$ entails that*

$$\|\nu-\nu'\|_1 \geqslant \frac{1-\mu}{(2/\alpha)\big((2/\alpha)-1\big)}\,.$$

Lemma 8 *We have that for any $\nu \in \mathcal{P}_F\big([0,1]\big)$, provided that $\mu \geqslant \mu-\varepsilon > E(\nu)$, the following inequalities hold true:*

$$\varepsilon/(1-\mu) \geqslant \mathcal{K}_{\inf}(\nu,\mu) - \mathcal{K}_{\inf}(\nu,\mu-\varepsilon) \geqslant 2\varepsilon^2$$

Moreover, the first inequality is also valid when $E(\nu) \geqslant \mu > \mu-\varepsilon$ or $\mu > E(\nu) \geqslant \mu-\varepsilon$.

Level sets of $\mathcal{K}_{\inf}$: For each $\gamma > 0$ and $\mu \in (0,1)$, we consider the set

$$\begin{aligned}\mathcal{C}_{\mu,\gamma} &= \Bigl\{\nu' \in \mathcal{P}_F\bigl([0,1]\bigr) : \quad \mathcal{K}_{\inf}(\nu',\mu) < \gamma\Bigr\} \\ &= \Bigl\{\nu' \in \mathcal{P}_F\bigl([0,1]\bigr) : \quad \exists\, \nu'_\mu \in \mathcal{P}_F\bigl([0,1]\bigr) \ \text{ s.t. } \ E\bigl(\nu'_\mu\bigr) > \mu \ \text{ and } \ \mathcal{K}\bigl(\nu', \nu'_\mu\bigr) < \gamma\Bigr\}.\end{aligned}$$

We detail a property in the following lemma, whose proof can be found in the extended version of the present conference paper (Maillard et al., 2011).

Lemma 9 *For all $\gamma > 0$ and $\mu \in (0,1)$, the set $\mathcal{C}_{\mu,\gamma}$ is a non-empty open convex set. Moreover,*

$$\overline{\mathcal{C}}_{\mu,\gamma} \supseteq \Bigl\{\nu' \in \mathcal{P}_F\bigl([0,1]\bigr) : \quad \mathcal{K}_{\inf}(\nu',\mu) \leqslant \gamma\Bigr\}.$$

4.2. The $\mathcal{K}_{\inf}$–strategy and a general performance guarantee

For each arm $a \in \mathcal{A}$ and round t with $N_t(a) > 0$, we denote by $\widehat{\nu}_{a,N_t(a)}$ the empirical distribution of the payoffs obtained till round t when picking arm a, that is,

$$\widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} \delta_{Y_s}\,,$$

where for all $x \in [0,1]$, we denote by δ_x the Dirac mass on x. We define the corresponding empirical averages as

$$\widehat{\mu}_{a^\star,N_t(a^\star)} = E\bigl(\widehat{\nu}_{a^\star,N_t(a^\star)}\bigr) = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} Y_s\,.$$

We then consider the $\mathcal{K}_{\inf}$*–strategy* defined in Figure 2. Note that the use of maxima in the definitions of the $B^+_{a,t}$ is justified by Lemma 6.

As explained in Honda and Takemura (2010b), the computation of the quantities $\mathcal{K}_{\inf}$ can be done efficiently in this case, i.e., when we consider only distributions with finite supports. This is because in the computation of $\mathcal{K}_{\inf}$, it is sufficient to consider only distributions with the same support as the empirical distributions (up to one point). Note that the knowledge of the support of the distributions associated with the arms is not required.

Theorem 10 *Assume that $\nu^\star$ is finitely supported, with expectation $\mu^\star \in (0,1)$ and with support denoted by $\mathcal{S}^\star$. Let $a \in \mathcal{A}$ be a suboptimal arm such that $\mu_a > 0$ and ν_a is finitely supported. Then, for all $c_a > 0$ and all*

$$0 < \varepsilon < \min\left\{\Delta_a,\, \frac{c_a/2}{1+c_a}(1-\mu^\star)\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)\right\},$$

the expected number of times the $\mathcal{K}_{\inf}$–strategy, run with $f(t) = \log t$, pulls arm a satisfies

$$\mathbb{E}\bigl[N_T(a)\bigr] \leqslant 1 + \frac{(1+c_a)\log T}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} + \frac{1}{1-e^{-\Theta_a(c_a,\varepsilon)}} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{1-\mu^* + \varepsilon}\right)\sum_{k=1}^{T}(k+1)^{|\mathcal{S}^\star|}\, e^{-k\varepsilon^2} + \frac{1}{(\Delta_a - \varepsilon)^2}\,,$$

```
Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
Initialization: Pull each arm of $\mathcal{A}$ once
For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,
    – compute for each arm $a \in \mathcal{A}$ the quantity
          $B^+_{a,t} = \max\Bigl\{ q \in [0,1] : \quad N_t(a)\, \mathcal{K}_{\inf}\bigl(\widehat{\nu}_{a,N_t(a)},\, q\bigr) \leqslant f(t) \Bigr\}$,
      where   $\widehat{\nu}_{a,N_t(a)} = \dfrac{1}{N_t(a)} \sum_{s \leqslant t :\, A_s = a} \delta_{Y_s}$ ;
    – in case of a tie, pick an arm with largest value of $\widehat{\mu}_{a,N_t(a)}$;
    – pull any arm $A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} B^+_{a,t}$ .
```

Figure 2: The strategy $\mathcal{K}_{\text{inf}}$.

where

$$\Theta_a(c_a, \varepsilon) = \theta_a\biggl(\frac{\log T}{k_0} + \frac{\varepsilon}{1-\mu^\star} \biggr) \qquad \textit{with} \qquad k_0 = \biggl\lceil \frac{(1+c_a)\,\log T}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^\star)} \biggr\rceil \, .$$

and for all $\gamma > 0$*,*

$$\theta_a(\gamma) = \inf\Bigl\{ \mathcal{K}(\nu', \nu_a) : \quad \nu' \ \text{ s.t. } \ \mathcal{K}_{\text{inf}}(\nu', \mu^\star) < \gamma \Bigr\} \, .$$

As a corollary, we get (by taking some common value for all c_a) that for all $c > 0$,

$$\overline{R}_T \leqslant \sum_{a \in \mathcal{A}} \Delta_a \frac{(1+c)\,\log T}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^\star)} + h(c)\, ,$$

where $h(c) < \infty$ is a function of c (and of the distributions associated with the arms), which is however independent of T. As a consequence, we recover the asymptotic results of Burnetas and Katehakis (1996); Honda and Takemura (2010a), i.e., the guarantee that

$$\limsup_{T \to \infty} \frac{\overline{R}_T}{\log T} \leqslant \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^\star)} \, .$$

Of course, a sharper optimization can be performed by carefully choosing the constants c_a, that are parameters of the analysis; similarly to the comments after the statement of Theorem 3, we would then get a dominant term with a constant factor 1 instead of $1 + c$ as above, plus an additional second-order term. Details are left to a journal version of this paper.

Proof By arguments similar to the ones used in the first step of the proof of Theorem 3, we have

$$\bigl\{A_{t+1} = a\bigr\} \subseteq \Bigl\{ \mu^\star - \varepsilon < \widehat{\mu}_{a,N_t(a)} \Bigr\} \cup \Bigl\{ \mu^\star - \varepsilon > B^+_{a^\star,t} \Bigr\} \cup \Bigl\{ \mu^\star - \varepsilon \in \bigl[\widehat{\mu}_{a,N_t(a)},\ B^+_{a,t} \bigr] \Bigr\} \, ;$$

indeed, on the event $\big\{A_{t+1} = a\big\} \cap \Big\{\mu^\star - \varepsilon \geqslant \widehat{\mu}_{a,N_t(a)}\Big\} \cap \Big\{\mu^\star - \varepsilon \leqslant B^+_{a^\star,t}\Big\}$, we have, $\widehat{\mu}_{a,N_t(a)} \leqslant \mu^\star - \varepsilon \leqslant B^+_{a^\star,t} \leqslant B^+_{a,t}$ (where the last inequality is by definition of the strategy). Before proceeding, we note that

$$\Big\{\mu^\star - \varepsilon \in \big[\widehat{\mu}_{a,N_t(a)},\ B^+_{a,t}\big]\Big\} \subseteq \Big\{N_t(a)\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\, \mu^\star - \varepsilon\big) \leqslant f(t)\Big\}\,,$$

since $\mathcal{K}_{\inf}$ is a non-decreasing function in its second argument and $\mathcal{K}_{\inf}\big(\nu, E(\nu)\big) = 0$ for all distributions ν. Therefore,

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{\mu^\star - \varepsilon < \widehat{\mu}_{a,N_t(a)} \ \text{ and }\ A_{t+1} = a\Big\} + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{\mu^\star - \varepsilon > B^+_{a^\star,t}\Big\}$$
$$+ \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{N_t(a)\ \mathcal{K}_{\inf}\big(\widehat{\nu}_{a,N_t(a)},\, \mu^\star - \varepsilon\big) \leqslant f(t) \ \text{ and }\ A_{t+1} = a\Big\}\,;$$

now, the two sums with the events "and $A_{t+1} = a$" can be rewritten by using the stopping times $\tau_{a,k}$ introduced in the proof of Theorem 3; more precisely, by mimicking the transformations performed in its step 3, we get the simpler bound

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{\mu^\star - \varepsilon < \widetilde{\mu}_{a,k-1}\Big\} + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{\mu^\star - \varepsilon > B^+_{a^\star,t}\Big\}$$
$$+ \sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{(k-1)\ \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\, \mu^\star - \varepsilon\big) \leqslant f(t)\Big\}\,, \tag{6}$$

where the $\widetilde{\nu}_{a,s}$ and $\widetilde{\mu}_{a,s}$ are respectively the empirical distributions and empirical expectations computed on the first s elements of the sequence of the random variables $\widetilde{X}_{a,j} = Y_{\tau_{a,j}}$, which are i.i.d. according to ν_a.

Step 1: The first sum in (6) is bounded by resorting to Hoeffding's inequality, whose application is legitimate since $\mu^\star - \mu_a - \varepsilon > 0$;

$$\begin{aligned}\sum_{k=2}^{T-|\mathcal{A}|+1} \mathbb{P}\Big\{\mu^\star - \varepsilon < \widetilde{\mu}_{a,k-1}\Big\} &= \sum_{k=1}^{T-|\mathcal{A}|} \mathbb{P}\Big\{\mu^\star - \mu_a - \varepsilon < \widetilde{\mu}_{a,k} - \mu_a\Big\} \\ &\leqslant \sum_{k=1}^{T-|\mathcal{A}|} e^{-2k(\mu^\star-\mu_a-\varepsilon)^2} \leqslant \frac{1}{1 - e^{-2(\mu^\star-\mu_a-\varepsilon)^2}} \leqslant \frac{1}{(\mu^\star - \mu_a - \varepsilon)^2}\end{aligned}$$

where we used for the last inequality the general upper bounds provided at the beginning of step 5 in the proof of Theorem 3.

Step 2: The second sum in (6) is bounded by first using the definition of $B^+_{a^\star,t}$, then, decomposing the event depending on the values taken by $N_t(a^\star)$; and finally using the fact that on $\big\{N_t(a^\star) = k\big\}$, we have the rewriting $\widehat{\nu}_{a,N_t(a)} = \widetilde{\nu}_{a,k}$ and $\widehat{\mu}_{a,N_t(a)} = \widetilde{\mu}_{a,k}$;

more precisely,

$$\begin{aligned}\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl\{\mu^\star - \varepsilon > B^+_{a^\star,t}\Bigr\} &\leqslant \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl\{N_t(a^\star)\ \mathcal{K}_{\inf}\bigl(\widehat{\nu}_{a^\star,N_t(a^\star)},\ \mu^\star-\varepsilon\bigr) > f(t)\Bigr\} \\ &= \sum_{t=|\mathcal{A}|}^{T-1} \sum_{k=1}^{t} \mathbb{P}\Bigl\{N_t(a^\star) = k \ \text{ and } \ k\ \mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) > f(t)\Bigr\} \\ &\leqslant \sum_{k=1}^{T} \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl\{k\ \mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) > f(t)\Bigr\}\,.\end{aligned}$$

Since $f = \log$ is increasing, we can rewrite the bound, using a Fubini-Tonelli argument, as

$$\begin{aligned}\sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl\{\mu^\star - \varepsilon > B^+_{a^\star,t}\Bigr\} &\leqslant \sum_{k=1}^{T} \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Bigl\{f^{-1}\Bigl(k\,\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr)\Bigr) > t\Bigr\} \\ &\leqslant \sum_{k=1}^{T} \mathbb{E}\Bigl[f^{-1}\Bigl(k\,\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr)\Bigr)\ \mathbb{I}_{\bigl\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon)>0\bigr\}}\Bigr]\,.\end{aligned}$$

Now, Honda and Takemura (2010a, Lemma 13) indicates that, since $\mu^\star - \varepsilon \in [0,1)$,

$$\sup_{\nu\in\mathcal{P}_F([0,1])} \mathcal{K}_{\inf}\bigl(\nu, \mu^\star-\varepsilon\bigr) \leqslant \log\bigl(1/(1-\mu^\star+\varepsilon)\bigr) \overset{\text{def}}{=} K_{\max}\,;$$

we define $Q = K_{\max}/\varepsilon^2$ and introduce the following sets $(V_q)_{1\leqslant q\leqslant Q}$:

$$V_q = \Bigl\{\nu \in \mathcal{P}_F\bigl([0,1]\bigr) : \quad (q-1)\varepsilon^2 < \mathcal{K}_{\inf}\bigl(\nu, \mu^* - \varepsilon\bigr) \leqslant q\varepsilon^2\Bigr\}.$$

A peeling argument (and by using that $f^{-1} = \exp$ is increasing as well) entails, for all $k \geqslant 1$,

$$\begin{aligned}&\mathbb{E}\Bigl[f^{-1}\Bigl(k\,\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr)\Bigr)\ \mathbb{I}_{\bigl\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon)>0\bigr\}}\Bigr] \qquad (7)\\ &= \sum_{q=1}^{Q} \mathbb{E}\Bigl[f^{-1}\Bigl(k\,\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr)\Bigr)\ \mathbb{I}_{\bigl\{\widetilde{\nu}_{a^\star,k}\in V_q\bigr\}}\Bigr] \\ &\leqslant \sum_{q=1}^{Q} \mathbb{P}\bigl\{\widetilde{\nu}_{a^\star,k} \in V_q\bigr\}\, f^{-1}(kq\varepsilon^2) \leqslant \sum_{q=1}^{Q} \mathbb{P}\Bigl\{\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) > (q-1)\varepsilon^2\Bigr\}\, f^{-1}(kq\varepsilon^2) \qquad (8)\end{aligned}$$

where we used the definition of V_q to obtain each of the two inequalities. Now, by Lemma 8, when $E\bigl(\widetilde{\nu}_{a^\star,k}\bigr) < \mu^\star - \varepsilon$, which is satisfied whenever $\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) > 0$, we have

$$\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) \leqslant \mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star\bigr) - 2\varepsilon^2 \leqslant \mathcal{K}\bigl(\widetilde{\nu}_{a^\star,k},\ \nu^\star\bigr) - 2\varepsilon^2\,,$$

where the last inequality is by mere definition of $\mathcal{K}_{\inf}$. Therefore,

$$\mathbb{P}\Bigl\{\mathcal{K}_{\inf}\bigl(\widetilde{\nu}_{a^\star,k},\ \mu^\star-\varepsilon\bigr) > (q-1)\varepsilon^2\Bigr\} \leqslant \mathbb{P}\Bigl\{\mathcal{K}\bigl(\widetilde{\nu}_{a^\star,k},\ \nu^\star\bigr) > (q+1)\varepsilon^2\Bigr\}\,.$$

We note that for all $k \geqslant 1$, $\mathbb{P}\Big\{\mathcal{K}\big(\widetilde{\nu}_{a^\star,k},\,\nu^\star\big) > (q+1)\varepsilon^2\Big\} \leqslant (k+1)^{|\mathcal{S}^\star|}\,e^{-k(q+1)\varepsilon^2}$, where we recall that $\mathcal{S}^\star$ denotes the finite support of $\nu^\star$ and where we applied the method of types; see, e.g., the extended version of the present paper (Maillard et al., 2011) for more details about this standard inequality. Now, (8) then yields, via the choice $f = \log$ and thus $f^{-1} = \exp$, that

$$\mathbb{E}\Big[f^{-1}\Big(k\,\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon\big)\Big)\,\mathbb{I}_{\big\{\mathcal{K}_{\inf}(\widetilde{\nu}_{a^\star,k},\,\mu^\star-\varepsilon)>0\big\}}\Big] \leqslant \underbrace{\sum_{q=1}^{Q}(k+1)^{|\mathcal{S}^\star|}\,e^{-k(q+1)\varepsilon^2}e^{kq\varepsilon^2}}_{=Q\,(k+1)^{|\mathcal{S}^\star|}\,e^{-k\varepsilon^2}}\,.$$

Substituting the value of Q, we therefore have proved that

$$\sum_{t=|\mathcal{A}|}^{T-1}\mathbb{P}\Big\{\mu^\star-\varepsilon > B^+_{a^\star,t}\Big\} \leqslant \frac{1}{\varepsilon^2}\log\!\left(\frac{1}{1-\mu^*+\varepsilon}\right)\sum_{k=1}^{T}(k+1)^{|\mathcal{S}^\star|}\,e^{-k\varepsilon^2}.$$

Step 3: The third sum in (6) is first upper bounded by Lemma 8, which states that

$$\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\,\mu^\star\big) - \varepsilon/(1-\mu^\star) \leqslant \mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k-1},\,\mu^\star-\varepsilon\big)$$

for all $k \geqslant 1$, and by using $f(t) \leqslant f(T)$; this gives

$$\sum_{k=1}^{T-|\mathcal{A}|}\mathbb{P}\Big\{k\;\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k},\,\mu^\star-\varepsilon\big) \leqslant f(t)\Big\}$$
$$\leqslant \sum_{k=1}^{T-|\mathcal{A}|}\mathbb{P}\left\{k\;\mathcal{K}_{\inf}\big(\widetilde{\nu}_{a,k},\,\mu^\star\big) \leqslant f(T) + \frac{k\,\varepsilon}{1-\mu^\star}\right\} = \sum_{k=1}^{T-|\mathcal{A}|}\mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma_k}\Big\}\,,$$

where $\gamma_k = f(T)/k + \varepsilon/(1-\mu^\star)$ and where the set $\overline{\mathcal{C}}_{\mu^\star,\gamma_k}$ was defined in Section 4.1. For all $\gamma > 0$, we then introduce

$$\theta_a(\gamma) = \inf\Big\{\mathcal{K}(\nu',\nu_a) : \quad \nu' \in \mathcal{C}_{\mu^\star,\gamma}\Big\} = \inf\Big\{\mathcal{K}(\nu',\nu_a) : \quad \nu' \in \overline{\mathcal{C}}_{\mu^\star,\gamma}\Big\}\,,$$

(where the second equality follows from the lower semi-continuity of $\mathcal{K}$) and aim at bounding $\mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma}\Big\}$.

As shown in Section 4.1, the set $\mathcal{C}_{\mu^\star,\gamma}$ is a non-empty open convex set. If we prove that $\theta_a(\gamma)$ is finite for all $\gamma > 0$, then all the conditions will be required to apply Lemma 1 and get the upper bound

$$\sum_{k=1}^{T-|\mathcal{A}|}\mathbb{P}\Big\{\widetilde{\nu}_{a,k} \in \overline{\mathcal{C}}_{\mu^\star,\gamma_k}\Big\} \leqslant \sum_{k=1}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_k)}\,.$$

To that end, we use the fact that ν_a is finitely supported. Now, either the probability of interest is null and we are done; or, it is not null, which implies that there exists a possible value of $\widetilde{\nu}_{a,k}$ that is in $\overline{\mathcal{C}}_{\mu^\star,\gamma}$; since this value is a distribution with a support included in the one of ν_a, it is absolutely continuous with respect to ν_a and hence, the Kullback-Leibler divergence between this value and ν_a is finite; in particular, $\theta_a(\gamma)$ is finite.

Finally, we bound the $\theta_a(\gamma_k)$ for values of k larger than $k_0 = \left\lceil \dfrac{(1+c_a)\,f(T)}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)} \right\rceil$; we have that for all $k \geqslant k_0$, in view of the bound put on ε,

$$\gamma_k \leqslant \gamma_{k_0} = \frac{f(T)}{k_0} + \frac{\varepsilon}{1-\mu^\star} < \frac{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}{1+c_a} + \frac{c_a/2}{1+c_a}\,\mathcal{K}_{\inf}(\nu_a,\mu^\star) = \frac{1+c_a/2}{1+c_a}\,\mathcal{K}_{\inf}(\nu_a,\mu^\star)\,. \quad (9)$$

Since θ_a is non increasing, we have

$$\sum_{k=1}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_k)} \leqslant k_0 - 1 + \sum_{k=k_0}^{T-|\mathcal{A}|} e^{-k\,\theta_a(\gamma_{k_0})} \leqslant k_0 - 1 + \frac{1}{1-e^{-\Theta_a(c_a,\varepsilon)}}\,,$$

provided that the quantity $\Theta_a(c_a,\varepsilon) = \theta_a\big(\gamma_{k_0}\big)$ is positive, which we prove now.

Indeed for all $\nu' \in \mathcal{C}_{\mu^\star,\gamma_{k_0}}$, we have by definition and by (9) that

$$\mathcal{K}_{\inf}(\nu',\mu^\star) - \mathcal{K}_{\inf}(\nu_a,\mu^\star) < \gamma_{k_0} - \mathcal{K}_{\inf}(\nu_a,\mu^\star) < -\big((c_a/2)/(1+c_a)\big)\mathcal{K}_{\inf}(\nu_a,\mu^\star)\,.$$

Now, in the case where $\mathbb{E}_{\nu_a}\big[(1-\mu^\star)/(1-X)\big] > 1$, we have, first by application of Pinsker's inequality and then by Lemma 7, that

$$\mathcal{K}\big(\nu',\nu_a\big) \;\geqslant\; \frac{\|\nu'-\nu_a\|_1^2}{2} \;\geqslant\; \frac{1}{2\,M^2_{\nu_a,\mu^\star}}\big(\mathcal{K}_{\inf}(\nu_a,\mu^\star) - \mathcal{K}_{\inf}(\nu',\mu^\star)\big)^2 > \frac{c_a^2\,\big(\mathcal{K}_{\inf}(\nu_a,\mu^\star)\big)^2}{8\,(1+c_a)^2\,M^2_{\nu_a,\mu^\star}}\,;$$

since, again by Pinsker's inequality, $\mathcal{K}_{\inf}(\nu_a,\mu^\star) \geqslant (\mu_a-\mu^\star)^2/2 > 0$, we have exhibited a lower bound independent of ν' in this case. In the case where $\mathbb{E}_{\nu_a}\big[(1-\mu^\star)/(1-X)\big] \leqslant 1$, we apply the second part of Lemma 7, with $\alpha_a = (c_a/2)/(1+c_a)$, and get

$$\mathcal{K}\big(\nu',\nu_a\big) \;\geqslant\; \frac{\|\nu'-\nu_a\|_1^2}{2} \;\geqslant\; \frac{1}{2}\left(\frac{1-\mu^\star}{(2/\alpha_a)\,\big((2/\alpha_a)-1\big)}\right)^2 > 0\,.$$

Thus, in both cases we found a positive lower bound independent of ν', so that the infimum over $\nu' \in \mathcal{C}_{\mu^\star,\gamma_{k_0}}$ of the quantities $\mathcal{K}_{\inf}(\nu',\mu^\star)$, which precisely equals $\theta_a\big(\gamma_{k_0}\big)$, is also positive. This concludes the proof. ■

Conclusion. We provided a finite-time analysis of the (asymptotically optimal) $\mathcal{K}_{\inf}$–strategy in the case of finitely supported distributions. The extension to the case of general distributions (e.g., by histogram-based approximations of such general distributions) is left for future work.

Acknowledgments

The authors wish to thank *Peter Auer* and *Daniil Ryabko* for insightful discussions. They acknowledge support from the French National Research Agency (ANR-08-COSI-004 project EXPLO-RA), the European Community's Seventh Framework Programme (project COMPLACS, grant agreement no. 231495), and the PASCAL2 Network of Excellence (EC grant no. 506778).

References

J-Y. Audibert, R. Munos, and C. Szepesvari. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410:1876–1902, 2009.

J.Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

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

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

Y. Chow and H. Teicher. *Probability Theory.* Springer, 1988.

I.H. Dinwoodie. Mesures dominantes et théorème de Sanov. *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques*, 28(3):365–373, 1992.

S. Filippi. *Stratégies optimistes en apprentissage par renforcement.* PhD thesis, Télécom ParisTech, 2010.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of COLT*, 2011.

A. Garivier and F. Leonardi. Context tree selection: A unifying view. *Stochastic Processes and their Applications*, 2011. In press.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of COLT*, pages 67–79, 2010a.

J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. arXiv:0905.2776, 2010b.

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

O.-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. 2011. URL `http://hal.archives-ouvertes.fr/inria-00574987/`.

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