



**HAL**  
open science

# Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos,  
Gilles Stoltz

► **To cite this version:**

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz. Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation. 2012. hal-00738209v1

**HAL Id: hal-00738209**

**<https://hal.science/hal-00738209v1>**

Preprint submitted on 3 Oct 2012 (v1), last revised 21 Mar 2013 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## KULLBACK-LEIBLER UPPER CONFIDENCE BOUNDS FOR OPTIMAL SEQUENTIAL ALLOCATION

BY OLIVIER CAPPÉ<sup>1</sup>, AURÉLIEN GARIVIER<sup>2</sup>, ODALRIC-AMBRYM  
MAILLARD<sup>3</sup>, RÉMI MUNOS<sup>4</sup> AND GILLES STOLTZ<sup>5</sup>

<sup>1</sup>*LTCI, Telecom ParisTech, CNRS*

<sup>2</sup>*IMT, Université Paul Sabatier*

<sup>3</sup>*University of Leoben*

<sup>4</sup>*INRIA Lille, SequeL Project*

<sup>5</sup>*Ecole Normale Supérieure, CNRS, INRIA & HEC Paris, CNRS*

We consider optimal sequential allocation in the context of the so-called stochastic multi-armed bandit model. We describe a generic index policy, in the sense of Gittins (1979), based on upper confidence bounds of the arm payoffs computed using the Kullback-Leibler divergence. We consider two classes of distributions for which instances of this general idea are analyzed: The `kl-UCB` algorithm is designed for one-parameter exponential families and the empirical `KL-UCB` algorithm for bounded and finitely supported distributions. Our main contribution is a unified finite-time analysis of the regret of these algorithms that asymptotically matches the lower bounds of Lai and Robbins (1985) and Burnetas and Katehakis (1996), respectively. We also investigate the behavior of these algorithms when used with general bounded rewards, showing in particular that they provide significant improvements over the state-of-the-art.

**1. Introduction.** This paper is about optimal sequential allocation in unknown random environments. More precisely, we consider the setting known under the conventional, if not very explicit, name of (stochastic) *multi-armed bandit*, in reference to the 19th century gambling game. In the multi-armed bandit model, the emphasis is put on focussing as quickly as possible on the best available option(s) rather than on estimating precisely the efficiency of each option. These options are referred to as arms and each of them is associated with a distribution; arms are indexed by  $a$  and associated distributions are denoted by  $\nu_a$ .

The archetypal example occurs in clinical trials where the options (or arms) correspond to available treatments whose efficiencies are unknown a priori and patients arrive sequentially; the action consists of prescribing a particular treatment to the patient and the observation corresponds (for

---

*Keywords and phrases:* Multi-armed bandit problems, Upper confidence bound, Kullback-Leibler divergence, Sequential testing

instance) to the success or failure of the treatment. The goal is clearly here to achieve as many successes as possible. A strategy for doing so is said to be *anytime* if it does not require to know in advance the number of patients that will participate to the experiment. Although the term multi-armed bandit was probably coined in the late 1960's (Gittins, 1979), the origin of the problem can be traced back to fundamental questions about optimal stopping policies in the context of clinical trials (see Thompson, 1933, 1935) raised since the 1930's (see also Wald, 1945; Robbins, 1952).

In his celebrated work, Gittins (1979) considered the *Bayesian-optimal* solution to the discounted infinite-horizon multi-armed bandit problem. Gittins first showed that the Bayesian optimal policy could be determined by dynamic programming in an extended Markov decision process. The second key element is the fact that the optimal policy search can be factored into a set of simpler computations to determine *indices* that fully characterize each arm given the current history of the game (Gittins, 1979; Whittle, 1980; Weber, 1992). The optimal policy is then an *index policy* in the sense that at each time round, the (or an) arm with highest index is selected. Hence, index policies only differ in the way the indices are computed.

From a practical perspective however, the use of Gittins indices is limited to specific arm distributions and is computationally challenging (Gittins, Glazebrook and Weber, 2011). In the 1980's, pionnering works by Lai and Robbins (1985), Chang and Lai (1987), Burnetas and Katehakis (1996, 1997, 2003) suggested that Gittins indices can be approximated by quantities that can be interpreted as upper bounds of confidence intervals. This insight was used in the landmark paper by Auer, Cesa-Bianchi and Fischer (2002) who popularized the acronym UCB (for upper confidence bounds) to refer to a particular variant of indices obtained using Hoeffding's inequality.

There are however significant differences between the algorithms and results of Gittins (1979) and Auer, Cesa-Bianchi and Fischer (2002). First, UCB is an anytime algorithm that does not rely on the use of a discount factor or even on the knowledge of the horizon of the problem. More significantly, the Bayesian perspective is absent and UCB is analyzed in terms of its frequentist (distribution-dependent or distribution-free) performance, by exhibiting so-called *finite-time*, non-asymptotic, regret bounds.

UCB is a very robust algorithm that is suitable for and has strong performance guarantees, including distribution-free ones, in all problems with bounded stochastic rewards. However, a closer examination of the arguments in the proof reveals that the form of the upper confidence bounds used in UCB is a direct consequence of the use of Hoeffding's inequality and significantly differs from the approximate form of Gittins indices suggested by Lai

and Robbins (1985) or Burnetas and Katehakis (1996). Furthermore, the frequentist asymptotic lower bounds for the regret obtained by these authors also suggest that the behavior of UCB can be far from optimal. Indeed, under suitable conditions on the model  $\mathcal{D}$  (the class of possible distributions associated with each arm), any policy that is “admissible” (i.e., not grossly under-performing, see Lai and Robbins, 1985 for details) must satisfy the following asymptotic inequality on its “regret” (a quantity to be defined in Section 2):

$$(1) \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[R_T]}{\log(T)} \geq \sum_{a: \mu_a < \mu^*} \frac{\mu^* - \mu_a}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)},$$

where  $R_T$  is the regret at round  $T$  and  $\mu_a$  denotes the expectation of the distribution  $\nu_a$  of arm  $a$ , while  $\mu^*$  is the maximal expectation among all arms. The quantity

$$(2) \quad \mathcal{K}_{\text{inf}}(\nu, \mu) = \inf \left\{ \text{KL}(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } \mathbb{E}(\nu') > \mu \right\},$$

which measures the difficulty of the problem, is the minimal Kullback-Leibler divergence between the arm distribution  $\nu$  and distributions in the model  $\mathcal{D}$  that have expectations larger than  $\mu$ . By comparison, the bound obtained in Auer, Cesa-Bianchi and Fischer (2002) for UCB is of the form

$$\mathbb{E}[R_T] \leq C \left( \sum_{a: \mu_a < \mu^*} \frac{1}{\mu^* - \mu_a} \right) \log(T) + o(\log(T)),$$

for some numerical constant  $C$ , e.g.,  $C = 8$  (we provide a refinement of the result of Auer, Cesa-Bianchi and Fischer, 2002 as Corollary 2 below). These two results coincide as to the logarithmic rate of the regret but the (distribution-dependent) constants differ, sometimes significantly. Based on this observation, Honda and Takemura (2010) proposed an algorithm, called DMED, that is not an index policy but was shown to improve over UCB in some situations.

Building on similar ideas, we show in this paper that for a large class of problems there does exist a generic index policy—already suggested by Lai and Robbins (1985) and Burnetas and Katehakis (2003)—that guarantees a bound on the expected regret of the form

$$\mathbb{E}[R_T] \leq \sum_{a: \mu_a < \mu^*} \left( \frac{\mu^* - \mu_a}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} \right) \log(T) + o(\log(T)),$$

and which is thus asymptotically optimal. Interestingly, the index used in this algorithm can be interpreted as the upper bound of a confidence region for the expectation constructed using an empirical likelihood principle (Owen, 2001).

We describe the implementation of this algorithm and analyze its performance in two practically important cases where the lower bound of (1) was shown to hold (Lai and Robbins, 1985; Burnetas and Katehakis, 1996)—namely, for one-parameter canonical exponential families of distributions (Section 4), in which case the algorithm is referred to as **k1-UCB**; and for finitely supported distributions (Section 5), where the algorithm is called empirical **KL-UCB**. Determining the empirical **KL-UCB** index requires solving a convex program (maximizing a linear function on the probability simplex under Kullback-Leibler constraints) for which we provide in Appendix C.1 a simple algorithm inspired by Filippi, Cappé and Garivier (2010).

The analysis presented here greatly improves over the preliminary results presented, on the one hand in Garivier and Cappé (2011), and on the other hand in Maillard, Munos and Stoltz (2011); more precisely, the improvements lie in the greater generality of the analysis and by the more precise evaluation of the remainder terms in the regret bounds. We believe that the result obtained in this paper for **k1-UCB** (Theorem 1) is not improvable. For empirical **KL-UCB** the bounding of the remainder term could be improved upon obtaining a sharper version of the contraction lemma for  $\mathcal{K}_{\text{inf}}$  (Lemma 6). Our numerical simulations indeed suggest that the factor  $(n+2)$  appearing in its bound may be superfluous. The proofs rely on results of independent statistical interest: non-asymptotic bounds on the level of sequential confidence intervals for the expectation of independent, identically distributed variables in canonical exponential families (Lemma 5) using, in the bounded case, the empirical likelihood method (Proposition 1).

For general bounded distributions, we further make three important observations. First, the particular instance of the **k1-UCB** algorithm based on the Kullback-Leibler divergence between normal distributions is the **UCB** algorithm, which allows us to provide an improved optimal finite-time analysis of its performance (Corollary 2). Next, the **k1-UCB** algorithm, when used with the Kullback-Leibler divergence between Bernoulli distributions, obtains a strictly better performance than **UCB**, for any bounded distribution (Corollary 1). Finally, although a complete analysis of the empirical **KL-UCB** algorithm is subject to further investigations, we show here that the empirical **KL-UCB** index is indeed a valid empirical-likelihood upper-confidence bound for the expectation, that is, that it has a guaranteed, non-asymptotic bound on the coverage probability, for any bounded distribution (Propo-

sition 1). We provide some empirical evidence that empirical KL-UCB also performs well for general bounded distributions and illustrate the tradeoffs around the use of the two algorithms, in particular for short horizons.

*Outline.* The paper is organized as follows. Section 2 introduces the necessary notations and defines the notion of regret. Section 3 presents the generic form of the KL-UCB algorithm and provides the main steps for its analysis, leaving two facts to be proven under each specific instantiation of the algorithm to a model. The  $\text{k1-UCB}$  algorithm in the case of one-dimensional exponential families is considered in Section 4, and the empirical KL-UCB algorithm for bounded and finitely supported distributions is presented in Section 5. Finally, the behavior of these algorithms in the case of general bounded distributions is investigated in Section 6; and numerical experiments comparing  $\text{k1-UCB}$  and empirical KL-UCB to their competitors are reported in Section 7. Proofs are postponed to the appendix (Sections A–C).

**2. Setup and notation.** We consider a bandit problem with finitely many arms indexed by  $a \in \{1, \dots, K\}$ , with  $K \geq 2$ , each associated with an (unknown) probability distribution  $\nu_a$  over  $\mathbb{R}$ . We assume however that a model  $\mathcal{D}$  is known: a family of probability distributions such that  $\nu_a \in \mathcal{D}$  for all arms  $a$ .

The game is sequential and goes as follows: At each round  $t \geq 1$ , the player picks an arm  $A_t$  (based on the information gained in the past) and receives a stochastic payoff  $Y_t$  drawn independently at random according to the distribution  $\nu_{A_t}$ . He only gets to see the payoff  $Y_t$ .

2.1. *Assessment of the quality of a strategy via its expected regret.* For each arm  $a \in \{1, \dots, K\}$ , we denote by  $\mu_a$  the expectation of its associated distribution  $\nu_a$  and we let  $a^*$  be any optimal arm, i.e.,

$$a^* \in \operatorname{argmax}_{a \in \{1, \dots, K\}} \mu_a.$$

We write  $\mu^*$  as a short-hand notation for the largest expectation  $\mu_{a^*}$  and denote the gap of the expected payoff  $\mu_a$  of an arm  $a$  to  $\mu^*$  as  $\Delta_a = \mu^* - \mu_a$ . In addition, the number of times each arm  $a$  is pulled between the rounds 1 and  $T$  is referred to as  $N_a(T)$ ,

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

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

$$(3) \quad R_T \stackrel{\text{def}}{=} \mathbb{E} \left[ T\mu^* - \sum_{t=1}^T Y_t \right] = \mathbb{E} \left[ T\mu^* - \sum_{t=1}^T \mu_{A_t} \right] = \sum_{a=1}^K \Delta_a \mathbb{E} [N_a(T)],$$

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  $\nu_{A_t}$  and also to the possible auxiliary randomizations that the decision-making strategy is resorting to.

The regret measures the cumulative loss resulting from pulling suboptimal arms, and thus quantifies the amount of exploration required by an algorithm in order to find a best arm, since, as (3) indicates, the regret scales with the expected number of pulls of suboptimal arms.

*2.2. Empirical distributions.* We will denote them in two related ways, depending on whether random averages indexed by the global time  $t$  or averages of a given number  $n$  of pulls of a given arms are considered. The first series of averages will be referred to by using a functional notation for the indexation in the global time:  $\hat{\nu}_a(t)$ , while the second series will be indexed with the local times  $n$  in subscripts:  $\hat{\nu}_{a,n}$ . These two related indexations, functional for global times and random averages versus subscript indexes for local times, will be consistent throughout the paper for all quantities at hand, not only empirical averages.

More formally, for all arms  $a$  and all rounds  $t$  such that  $N_a(t) \geq 1$ ,

$$\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^t \delta_{Y_s} \mathbb{I}_{\{A_s=a\}},$$

where  $\delta_x$  denotes the Dirac distribution on  $x \in \mathbb{R}$ .

For averages based on local times we need to introduce stopping times. To that end, we consider the filtration  $(\mathcal{F}_t)$ , where for all  $t \geq 1$ , the  $\sigma$ -algebra  $\mathcal{F}_t$  is generated by  $A_1, Y_1, \dots, A_t, Y_t$ . In particular,  $A_{t+1}$  and all  $N_a(t+1)$  are  $\mathcal{F}_t$ -measurable. For all  $n \geq 1$ , we denote by  $\tau_{a,n}$  the round at which  $a$  was pulled for the  $n$ -th time; since

$$\tau_{a,n} = \min\{t \geq 1 : N_a(t) = n\},$$

we see that  $\{\tau_{a,n} = t\}$  is  $\mathcal{F}_{t-1}$ -measurable. That is, each random variable  $\tau_{a,s}$  is a (predictable) stopping time. Hence, as shown for instance in (Chow

and Teicher, 1988, Section 5.3), the random variables  $X_{a,n} = Y_{\tau_{a,n}}$ , where  $n = 1, 2, \dots$ , are independent and identically distributed according to  $\nu_a$ . For all arms  $a$ , we then denote by

$$\hat{\nu}_{a,n} = \frac{1}{n} \sum_{k=1}^n \delta_{X_{a,k}}$$

the empirical distributions corresponding to local times  $n \geq 1$ .

All in all, we of course have the rewriting

$$\hat{\nu}_a(t) = \hat{\nu}_{a,N_a(t)}.$$

**3. The KL-UCB algorithm.** We fix an interval or discrete subset  $\mathcal{S} \subseteq \mathbb{R}$  and denote by  $\mathfrak{M}_1(\mathcal{S})$  the set of all probability distributions over  $\mathcal{S}$ . For two distributions  $\nu, \nu' \in \mathfrak{M}_1(\mathcal{S})$ , we denote by  $\text{KL}(\nu, \nu')$  their Kullback-Leibler divergence and by  $\mathbb{E}(\nu)$  and  $\mathbb{E}(\nu')$  their expectations. (This expectation operator is denoted by  $\mathbb{E}$  while expectations with respect to underlying randomizations are referred to as  $\mathbb{E}$ .)

---

**Algorithm 1:** The KL-UCB algorithm (generic form).

---

**Parameters:** An operator  $\Pi_{\mathcal{D}} : \mathfrak{M}_1(\mathcal{S}) \rightarrow \mathcal{D}$ ; a non-decreasing function  $f : \mathbb{N} \rightarrow \mathbb{R}$

**Initialization:** Pull each arm of  $\{1, \dots, K\}$  once

**for**  $t = K$  **to**  $T - 1$ , **do**

compute for each arm  $a$  the quantity

$$(4) \quad U_a(t) = \sup \left\{ \mathbb{E}(\nu) : \nu \in \mathcal{D} \quad \text{and} \quad \text{KL}(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \nu) \leq \frac{f(t)}{N_a(t)} \right\}$$

pick an arm  $A_{t+1} \in \underset{a \in \{1, \dots, K\}}{\text{argmax}} U_a(t)$

---

The generic form of the algorithm of interest in this paper is described as Algorithm 1. It relies on two parameters: an operator  $\Pi_{\mathcal{D}}$  (in spirit, a projection operator) that associates with each empirical distribution  $\hat{\nu}_a(t)$  an element of the model  $\mathcal{D}$ ; and a non-decreasing function  $f$ , which is typically such that  $f(t) \approx \log(t)$ .

At each round  $t \geq K$ , an upper confidence bound  $U_a(t)$  is associated with the expectation  $\mu_a$  of the distribution  $\nu_a$  of each arm; an arm  $A_{t+1}$  with highest upper confidence bound is then played. Note that the algorithm does not need to know the time horizon  $T$  in advance. Furthermore, the UCB algorithm of Auer, Cesa-Bianchi and Fischer (2002) may be recovered



by replacing  $\text{KL}(\Pi_{\mathcal{D}}(\widehat{\nu}_a(t)), \nu)$  with a quantity proportional to  $(\mathbb{E}(\widehat{\nu}_a(t)) - \mathbb{E}(\nu))^2$ ; the implications of this observation will be made more explicit in Section 6.

*3.1. General analysis of performance.* In Sections 4 and 5, we prove non-asymptotic regret bounds for Algorithm 1 in two different settings. These bounds match the asymptotic lower bound (1) in the sense that, according to (3), bounding the expected regret is equivalent to bounding the number of suboptimal draws. We show that, for any suboptimal arm  $a$ , we have

$$\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} (1 + o(1)) ,$$

where the quantity  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$  was defined in the introduction. This result appears as a consequence of non-asymptotic bounds, which are derived using a common analysis framework detailed in the rest of this section.

Note that the term  $\log(T)/\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$  has an heuristic interpretation in terms of large deviations, which gives some insight on the regret analysis to be presented below. Let  $\nu' \in \mathcal{D}$  be such that  $\mathbb{E}(\nu') \geq \mu^*$ , let  $X'_1, \dots, X'_n$  be independent variables with distribution  $\nu'$ , and let  $\widehat{\nu}'_n = (\delta_{X'_1} + \dots + \delta_{X'_n})/n$ . By Sanov's theorem, for a small neighborhood  $\mathcal{V}_a$  of  $\nu_a$ , the probability that  $\widehat{\nu}'_n$  belongs to  $\mathcal{V}_a$  is such that

$$-\frac{1}{n} \log \mathbb{P}\{\widehat{\nu}'_n \in \mathcal{V}_a\} \xrightarrow{n \rightarrow \infty} \inf_{\nu \in \mathcal{V}_a} \text{KL}(\nu, \nu') \approx \text{KL}(\nu_a, \nu') \geq \mathcal{K}_{\text{inf}}(\nu_a, \mu^*) .$$

In the limit, ignoring the sub-exponential terms, this means that for  $n = \log(T)/\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$ , the probability  $\mathbb{P}\{\widehat{\nu}'_n \in \mathcal{V}_a\}$  is smaller than  $1/T$ . Hence,  $\log(T)/\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$  appears as the minimal number  $n$  of draws ensuring that the probability under any distribution with expectation at least  $\mu^*$  of the event “the empirical distribution of  $n$  independent draws belongs to a neighborhood of  $\nu_a$ ” is smaller than  $1/T$ . This event, of course, has an overwhelming probability under  $\nu_a$ . The significance of  $1/T$  as a cutoff value can be understood as follows: if the suboptimal arm  $a$  is chosen along the  $T$  draws, then the regret is at most equal to  $(\mu^* - \mu_a)T$ ; thus, keeping the probability of this event under  $1/T$  bounds the contribution of this event to the average regret by a constant. Incidentally, this explains why knowing  $\mu^*$  in advance does not significantly reduce the number of necessary suboptimal draws. The analysis that follows shows that the bandit problem, despite its sequential aspect and the absence of prior knowledge on the expectation of the arms, is indeed comparable to a sequence of tests of level  $1 - 1/T$  with null hypothesis  $H_0 : \mathbb{E}(\nu') > \mu^*$  and alternative hypothesis  $H_1 : \nu' = \nu_a$ , for

which Stein's lemma (see, e.g., van der Vaart 2000, Theorem 16.12) states that the best error exponent is  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$ .

Let us now turn to the main lines of the regret proof. By definition of the algorithm, at rounds  $t \geq K$ , one has  $A_{t+1} = a$  only if  $U_a(t) \geq U_{a^*}(t)$ . Therefore, one has the decomposition

$$(5) \quad \begin{aligned} \{A_{t+1} = a\} &\subseteq \{\mu^\dagger \geq U_{a^*}(t)\} \cup \{\mu^\dagger < U_{a^*}(t) \text{ and } A_{t+1} = a\} \\ &\subseteq \{\mu^\dagger \geq U_{a^*}(t)\} \cup \{\mu^\dagger < U_a(t) \text{ and } A_{t+1} = a\}, \end{aligned}$$

where  $\mu^\dagger$  is a parameter which is taken either equal to  $\mu^*$ , or slightly smaller when required by technical arguments. The event  $\{\mu^\dagger < U_a(t)\}$  can be rewritten as

$$\begin{aligned} \{\mu^\dagger < U_a(t)\} &= \left\{ \exists \nu' \in \mathcal{D} : \mathbb{E}(\nu') > \mu^\dagger \text{ and } \text{KL}(\Pi_{\mathcal{D}}(\widehat{\nu}_a(t)), \nu') \leq \frac{f(t)}{N_a(t)} \right\} \\ &= \left\{ \widehat{\nu}_a(t) \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)} \right\} = \left\{ \widehat{\nu}_{a, N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)} \right\}, \end{aligned}$$

where for  $\mu \in \mathbb{R}$  and  $\gamma > 0$ , the set  $\mathcal{C}_{\mu, \gamma}$  is defined as

$$(6) \quad \mathcal{C}_{\mu, \gamma} = \left\{ \nu \in \mathfrak{M}_1(\mathcal{S}) : \exists \nu' \in \mathcal{D} \text{ with } \mathbb{E}(\nu') > \mu \text{ and } \text{KL}(\Pi_{\mathcal{D}}(\nu), \nu') \leq \gamma \right\}.$$

By definition of  $\mathcal{K}_{\text{inf}}$ ,

$$(7) \quad \mathcal{C}_{\mu, \gamma} \subseteq \left\{ \nu \in \mathfrak{M}_1(\mathcal{S}) : \mathcal{K}_{\text{inf}}(\Pi_{\mathcal{D}}(\nu), \mu) \leq \gamma \right\}.$$

Using (5), and recalling that for rounds  $t \in \{1, \dots, K\}$ , each arm is played once, one obtains

$$\begin{aligned} \mathbb{E}[N_T(a)] &\leq 1 + \sum_{t=K}^{T-1} \mathbb{P}\{\mu^\dagger \geq U_{a^*}(t)\} \\ &\quad + \sum_{t=K}^{T-1} \mathbb{P}\left\{ \widehat{\nu}_{a, N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)} \text{ and } A_{t+1} = a \right\}. \end{aligned}$$

The two sums in this decomposition are handled separately. The first sum is negligible with respect to the second sum: case-specific arguments, given in Sections 4 and 5, prove the following statement.

**FACT TO BE PROVEN 1.** *For proper choices of  $\Pi_{\mathcal{D}}$ ,  $f$ , and  $\mu^\dagger$ , the sum  $\sum \mathbb{P}\{\mu^\dagger \geq U_{a^*}(t)\}$  is negligible with respect to  $\log T$ .*

The second sum is thus the leading term in the bound. It is first rewritten using the stopping times  $\tau_{a,2}, \tau_{a,3}, \dots$  introduced in Section 2. Indeed,  $A_{t+1} = a$  happens for  $t \geq K$  if and only if  $\tau_{a,n} = t + 1$  for some  $n \in \{2, \dots, T - K + 1\}$ ; and of course, two stopping times  $\tau_{a,n}$  and  $\tau_{a,n'}$  cannot be equal when  $n \neq n'$ . We also note that  $N_a(\tau_{a,n} - 1) = n - 1$  for  $n \geq 2$ . Therefore,

$$\begin{aligned}
(8) \quad & \sum_{t=K}^{T-1} \mathbb{P} \left\{ \widehat{\nu}_{a, N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(t)/N_a(t)} \text{ and } A_{t+1} = a \right\} \\
& \leq \sum_{t=K}^{T-1} \mathbb{P} \left\{ \widehat{\nu}_{a, N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(T)/N_a(t)} \text{ and } A_{t+1} = a \right\} \\
& = \sum_{t=K}^{T-1} \sum_{n=2}^{T-K+1} \mathbb{P} \left\{ \widehat{\nu}_{a, N_a(t)} \in \mathcal{C}_{\mu^\dagger, f(T)/N_a(t)} \text{ and } \tau_{a,n} = t + 1 \right\} \\
& = \sum_{n=2}^{T-K+1} \sum_{t=K}^{T-1} \mathbb{P} \left\{ \widehat{\nu}_{a, n-1} \in \mathcal{C}_{\mu^\dagger, f(T)/(n-1)} \text{ and } \tau_{a,n} = t + 1 \right\} \\
& \leq \sum_{n=1}^{T-K} \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n} \right\},
\end{aligned}$$

where we used, successively, the following facts: the sets  $\mathcal{C}_{\mu^\dagger, \gamma}$  grow with  $\gamma$ ; the event  $\{A_{t+1} = a\}$  can be written as a disjoint union of the events  $\{\tau_{a,n} = t + 1\}$ , for  $2 \leq n \leq T - K + 1$ ; the events  $\{\tau_{a,n} = t + 1\}$  are disjoint as  $t$  varies between  $K$  and  $T - 1$ , with a possibly empty union (as  $\tau_{a,n}$  may be larger than  $T$ ).

By upper bounding the first

$$(9) \quad n_0 = \left\lceil \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} \right\rceil$$

terms of the sum in (8) by 1, we obtain

$$\sum_{n=1}^{T-K} \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n} \right\} \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + 1 + \sum_{n \geq n_0+1} \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n} \right\}.$$

It remains to upper bound the remaining sum: this is the object of the following statement, which will also be proved using case-specific arguments.

**FACT TO BE PROVEN 2.** *For proper choices of  $\Pi_{\mathcal{D}}$ ,  $f$ , and  $\mu^\dagger$ , the sum  $\sum \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n} \right\}$  is negligible with respect to  $\log T$ .*

Putting everything together, one obtains

$$(10) \quad \mathbb{E}[N_a(T)] \leq \frac{f(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^*)} + \underbrace{\sum_{n \geq n_0+1} \mathbb{P}\{\widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\dagger, f(T)/n}\}}_{o(\log T)} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\{\mu^\dagger \geq U_{a^*}(t)\}}_{o(\log T)} + 2.$$

Theorems 1 and 2 are instances of this general bound providing non-asymptotic controls for  $\mathbb{E}[N_T(a)]$  in the two settings considered in this paper.

#### 4. Rewards in a canonical one-dimensional exponential family.

We consider in this section the case when  $\mathcal{D}$  is a canonical exponential family of probability distributions  $\nu_\theta$ , indexed by  $\theta \in \Theta$ ; that is, the distributions  $\nu_\theta$  are absolutely continuous with respect to a dominating measure  $\rho$  on  $\mathbb{R}$ , with probability density

$$\frac{d\nu_\theta}{d\rho}(x) = \exp(x\theta - b(\theta)), \quad x \in \mathbb{R};$$

we assume in addition that  $b : \Theta \rightarrow \mathbb{R}$  is twice differentiable. We also assume that  $\Theta \subseteq \mathbb{R}$  is the natural parameter space, that is, the set

$$\Theta = \left\{ \theta \in \mathbb{R} : \int_{\mathbb{R}} \exp(x\theta) d\rho(x) < \infty \right\},$$

and that the exponential family  $\mathcal{D}$  is regular, i.e., that  $\Theta$  is an open interval (an assumption that turns out to be true in all the examples listed below). For a thorough introduction to canonical exponential families, as well as proofs of the following properties, the reader is referred to Lehmann and Casella (1998).

The derivative  $\dot{b}$  of  $b$  is an increasing continuous function such that  $\mathbb{E}(\nu_\theta) = \dot{b}(\theta)$  for all  $\theta \in \Theta$ ; in particular,  $b$  is strictly convex. Thus,  $\dot{b}$  is one-to-one with a continuous inverse  $\dot{b}^{-1}$  and the distributions  $\nu_\theta$  of  $\mathcal{D}$  can also be parameterized by their expectations  $\mathbb{E}(\nu_\theta)$ . Defining the open interval of all expectations,  $I = \dot{b}(\Theta) = (\mu_-, \mu_+)$ , there exists a unique distribution of  $\mathcal{D}$  with expectation  $\mu \in I$ , namely,  $\nu_{\dot{b}^{-1}(\mu)}$ .

The Kullback-Leibler divergence between two distributions  $\nu_\theta, \nu_{\theta'} \in \mathcal{D}$  is given by

$$\text{KL}(\nu_\theta, \nu_{\theta'}) = (\theta - \theta') \dot{b}(\theta) - b(\theta) + b(\theta'),$$

which, writing  $\mu = \mathbb{E}(\nu_\beta)$  and  $\mu' = \mathbb{E}(\nu_{\theta'})$ , can be reformulated as  
(11)

$$d(\mu, \mu') \stackrel{\text{def}}{=} \text{KL}(\nu_\theta, \nu_{\theta'}) = (\dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu')) \mu - b(\dot{b}^{-1}(\mu)) + b(\dot{b}^{-1}(\mu')) .$$

This defines a divergence  $d : I \times I \rightarrow \mathbb{R}_+$  that inherits from the Kullback-Leibler divergence the property that  $d(\mu, \mu') = 0$  if and only if  $\mu = \mu'$ . In addition,  $d$  is (strictly) convex and differentiable over  $I \times I$ .

As the examples below of specific canonical exponential families illustrate, the closed-form expression for this re-parameterized Kullback-Leibler divergence is usually simple.

EXAMPLE 1 (Binomial distributions for  $n$ -samples).  $\theta = \log(\mu/(n-\mu))$ ,  $\Theta = \mathbb{R}$ ,  $b(\theta) = n \log(1 + \exp(\theta))$ ,  $I = (0, n)$ ,

$$d(\mu, \mu') = \mu \log \frac{\mu}{\mu'} + (n - \mu) \log \frac{n - \mu}{n - \mu'} .$$

*The case  $n = 1$  corresponds to Bernoulli distributions.*

EXAMPLE 2 (Poisson distributions).  $\theta = \log(\mu)$ ,  $\Theta = \mathbb{R}$ ,  $b(\theta) = \exp(\theta)$ ,  $I = (0, +\infty)$ ,

$$d(\mu, \mu') = \mu' - \mu + \mu \log \frac{\mu}{\mu'} .$$

EXAMPLE 3 (Negative binomial distributions with known shape parameter  $r$ ).  $\theta = \log(\mu/(r + \mu))$ ,  $\Theta = (-\infty, 0)$ ,  $b(\theta) = -r \log(1 - \exp(\theta))$ ,  $I = (0, +\infty)$ ,

$$d(\mu, \mu') = r \log \frac{r + \mu'}{r + \mu} + \mu \log \frac{\mu(r + \mu')}{\mu'(r + \mu)} .$$

*The case  $r = 1$  corresponds to geometric distributions.*

EXAMPLE 4 (Gaussian distributions with known variance  $\sigma^2$ ).  $\theta = \mu/\sigma^2$ ,  $\Theta = \mathbb{R}$ ,  $b(\theta) = \sigma^2 \theta^2/2$ ,  $I = \mathbb{R}$ ,

$$d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2} .$$

EXAMPLE 5 (Gamma distributions with known shape parameter  $\alpha$ ).  $\theta = -\alpha/\mu$ ,  $\Theta = (-\infty, 0)$ ,  $b(\theta) = -\alpha \log(-\theta)$ ,  $I = (0, +\infty)$ ,

$$d(\mu, \mu') = \alpha \left( \frac{\mu}{\mu'} - 1 - \log \frac{\mu}{\mu'} \right) .$$

*The case  $\alpha = 1$  corresponds to exponential distributions.*

For all  $\mu \in I$  the convex functions  $d(\cdot, \mu)$  and  $d(\mu, \cdot)$  can be extended by continuity to  $\bar{I} = [\mu_-, \mu_+]$  as follows:

$$d(\mu_-, \mu) = \lim_{\mu' \rightarrow \mu_-} d(\mu', \mu), \quad d(\mu_+, \mu) = \lim_{\mu' \rightarrow \mu_+} d(\mu', \mu),$$

with similar statements for the second function. Note that these limits may equal  $+\infty$ ; the extended function  $d : \bar{I} \times I \cup \bar{I} \times \bar{I} \rightarrow [0, +\infty]$  is still a convex function. By convention, we also define  $d(\mu_-, \mu_-) = d(\mu_+, \mu_+) = 0$ .

Note that our exponential family models are minimal in the sense of Wainwright and Jordan (2008, Section 3.2) and thus that  $I$  coincides with the interior of the set of realizable expectations for all distributions that are absolutely continuous with respect to  $\rho$  (see Wainwright and Jordan 2008, Theorem 3.3 and Appendix B). In particular, this implies that distributions in  $\mathcal{D}$  have supports in  $\bar{I}$  and that, consequently, the empirical means  $\hat{\nu}_a(t)$  are in  $\bar{I}$  for all  $a$  and  $t$ . (Note however that they may not be in  $I$  itself: think in particular of the case of Bernoulli distributions when  $t$  is small.)

4.1. *The k1-UCB algorithm.* As the distributions in  $\mathcal{D}$  can be parameterized by their expectation,  $\Pi_{\mathcal{D}}$  associates with each  $\nu \in \mathfrak{M}_1(\bar{I})$  such that  $\mathbb{E}(\nu) \in I$  the distribution  $\nu_{b^{-1}(\mathbb{E}(\nu))} \in \mathcal{D}$ , which has the same expectation.

As shown above, for all  $\nu' \in \mathcal{D}$  it then holds that  $\text{KL}(\Pi_{\mathcal{D}}(\nu), \nu') = d(\mathbb{E}(\nu), \mathbb{E}(\nu'))$ ; and this equality can be extended to the case where  $\mathbb{E}(\nu) \in \bar{I}$ . In this setting, sufficient statistics for  $\hat{\nu}_a(t)$  and  $\hat{\nu}_{a,n}$  are given by, respectively,

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^t Y_s \mathbb{I}_{\{A_s=a\}} \quad \text{and} \quad \hat{\mu}_{a,n} = \frac{1}{n} \sum_{k=1}^n X_{a,k},$$

where the former is defined as soon as  $N_a(t) \geq 1$ .

The upper-confidence bound  $U_a(t)$  may be defined in this model not only in terms of  $\mathcal{D}$  but also of its “boundaries,” namely, in terms of  $\bar{I}$  and not only  $I$ , as

$$(12) \quad U_a(t) = \sup \left\{ \mu \in \bar{I} : d(\hat{\mu}_a(t), \mu) \leq \frac{f(t)}{N_a(t)} \right\}.$$

This supremum is achieved: in the case when  $\hat{\mu}_a(t) \in I$ , this follows from the fact that  $d$  is continuous on  $I \times \bar{I}$ ; when  $\hat{\mu}_a(t) = \mu_+$ , this is because  $U_a(t) = \mu_+$ ; in the case when  $\hat{\mu}_a(t) = \mu_-$ , either  $\mu_-$  is the only  $\mu \in \bar{I}$  for which  $d(\mu_-, \mu)$  is finite, or  $d(\mu_-, \cdot)$  is convex thus continuous on the open interval where it is finite.

Thus, in the setting of this section, Algorithm 1 rewrites as Algorithm 2 below, which will be referred to as k1-UCB.

---

**Algorithm 2:** The k1-UCB algorithm.

---

**Parameters:** A non-decreasing function  $f : \mathbb{N} \rightarrow \mathbb{R}$ 
**Initialization:** Pull each arm of  $\{1, \dots, K\}$  once

**for**  $t = K$  **to**  $T - 1$ , **do**

    compute for each arm  $a$  the quantity

$$U_a(t) = \sup \left\{ \mu \in \bar{I} : d(\hat{\mu}_a(t), \mu) \leq \frac{f(t)}{N_a(t)} \right\}$$

    pick an arm  $A_{t+1} \in \operatorname{argmax}_{a \in \{1, \dots, K\}} U_a(t)$ 


---

In practice, the computation of  $U_a(t)$  boils down to finding the zero of an increasing and convex scalar function. This can be done either by dichotomic search or by Newton iterations. In all the examples given above, well-known inequalities (e.g., Hoeffding's inequality) may be used to obtain an initial upper bound on  $U_a(t)$ .

4.2. *Regret analysis.* In this parametric context we have  $\mathcal{K}_{\inf}(\nu, \mu) = d(\mathbb{E}(\nu), \mu)$  when  $\mathbb{E}(\nu) \in I$  and  $\mu \in I$ . The following theorem thus proves, in light of the lower bound of Lai and Robbins (1985), the asymptotic optimality of the k1-UCB algorithm. Moreover, it provides an explicit, non-asymptotic bound on the regret.

**THEOREM 1.** *Assume that all arms belong to a canonical, regular, exponential family  $\mathcal{D} = \{\nu_\theta : \theta \in \Theta\}$  of probability distributions indexed by its natural parameter space  $\Theta \subseteq \mathbb{R}$ . Then, using Algorithm 2 with the divergence  $d$  given in (11) and with the choice  $f(t) = \log(t) + 3 \log \log(t)$  for  $t \geq 3$  and  $f(1) = f(2) = f(3)$ , the number of draws of any suboptimal arm  $a$  is upper bounded for any horizon  $T \geq 3$  as*

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{\log(T)}{d(\mu_a, \mu^*)} + 2 \sqrt{\frac{2\pi\sigma_{a,*}^2 (d'(\mu_a, \mu^*))^2}{(d(\mu_a, \mu^*))^3} \sqrt{\log(T) + 3 \log \log(T)}} \\ &\quad + \left(4e + \frac{3}{d(\mu_a, \mu^*)}\right) \log \log(T) + 8\sigma_{a,*}^2 \left(\frac{d'(\mu_a, \mu^*)}{d(\mu_a, \mu^*)}\right)^2 + 6, \end{aligned}$$

where  $\sigma_{a,*}^2 = \max \{ \operatorname{Var}(\nu_\theta) : \mu_a \leq \mathbb{E}(\nu_\theta) \leq \mu^* \}$  and where  $d'(\cdot, \mu^*)$  denotes the derivative of  $d(\cdot, \mu^*)$ .

**5. Bounded and finitely supported rewards.** In this section,  $\mathcal{D}$  is the set  $\mathcal{F}$  of finitely supported probability distributions over  $\mathcal{S} = [0, 1]$ . In this case, the empirical measures  $\widehat{\nu}_a(t)$  belong to  $\mathcal{F}$  and hence the operator  $\Pi_{\mathcal{D}}$  is taken to be the identity. We denote by  $\text{Supp}(\nu)$  the finite support of an element  $\nu \in \mathcal{F}$ .

The maximization program (4) defining  $U_a(t)$  admits in this case the simpler formulation

$$\begin{aligned} U_a(t) &\stackrel{\text{def}}{=} \sup \left\{ \mathbb{E}(\nu) : \nu \in \mathcal{F} \text{ and } \text{KL}(\widehat{\nu}_a(t), \nu) \leq \frac{f(t)}{N_a(t)} \right\} \\ &= \sup \left\{ \mathbb{E}(\nu) : \nu \in \mathfrak{M}_1(\text{Supp}(\widehat{\nu}_a(t)) \cup \{1\}) \text{ and } \text{KL}(\widehat{\nu}_a(t), \nu) \leq \frac{f(t)}{N_a(t)} \right\}, \end{aligned}$$

and also admits an explicit computational solution; these two points are detailed in Appendix C.1. Thus Algorithm 1 takes the following simpler form, which will be referred to as the empirical KL-UCB algorithm.

---

**Algorithm 3:** The empirical KL-UCB algorithm.

---

**Parameters:** A non-decreasing function  $f : \mathbb{N} \rightarrow (0, +\infty)$

**Initialization:** Pull each arm of  $\{1, \dots, K\}$  once

**for**  $t = K$  **to**  $T - 1$ , **do**

compute for each arm  $a$  the quantity

$$U_a(t) = \sup \left\{ \mathbb{E}(\nu) : \nu \in \mathfrak{M}_1(\text{Supp}(\widehat{\nu}_a(t)) \cup \{1\}) \text{ and } \text{KL}(\widehat{\nu}_a(t), \nu) \leq \frac{f(t)}{N_a(t)} \right\}$$

pick an arm  $A_{t+1} \in \underset{a \in \{1, \dots, K\}}{\text{argmax}} U_a(t)$

---

Like the DMED algorithm, for which asymptotic bounds are proved in Honda and Takemura (2010), Algorithm 1 relies on the empirical likelihood method (see Owen 2001) for the construction of the confidence bounds. However, DMED is not an index policy, but it maintains a list of active arms—an approach that, generally speaking, seems to be less satisfactory and slightly less efficient in practice. Besides, the analyses of the two algorithms, even though they both rely on some technical properties of the function  $\mathcal{K}_{\text{inf}}$ , differ significantly.

**THEOREM 2.** *Assume that  $\mu_a > 0$  for all arms  $a$  and that  $\mu^* < 1$ . There exists a constant  $M(\nu_a, \mu^*) > 0$  only depending on  $\nu_a$  and  $\mu^*$  such that, with*



the choice  $f(t) = \log(t) + \log(\log(t))$  for  $t \geq 2$ , the expected number of times that any suboptimal arm  $a$  is pulled by Algorithm 3 is smaller, for all  $T \geq 3$ , than

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{\log(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{36}{(\mu^*)^4} (\log(T))^{4/5} \log(\log(T)) \\ &\quad + \left( \frac{72}{(\mu^*)^4} + \frac{2\mu^*}{(1-\mu^*) \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2} \right) (\log(T))^{4/5} \\ &\quad + \frac{(1-\mu^*)^2 M(\nu_a, \mu^*)}{2(\mu^*)^2} (\log(T))^{2/5} \\ &\quad + \frac{\log(\log(T))}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{2\mu^*}{(1-\mu^*) \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2} + 4. \end{aligned}$$

Theorem 2 implies a non-asymptotic bound of the form

$$\mathbb{E}[N_a(T)] \leq \frac{\log(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + O\left((\log(T))^{4/5} \log(\log(T))\right).$$

The exact value of the constant  $M(\nu_a, \mu^*)$  is provided in the proof of Theorem 2 (see Section B.3) and relies on the variational form of  $\mathcal{K}_{\text{inf}}$  introduced in Section B.1 (see Lemma 4).

**6. Algorithms for general bounded rewards.** In this section, we consider the case where the arms are only known to have bounded distributions. As in Section 5, we assume without loss of generality that the rewards are bounded in  $[0, 1]$ . This is the setting considered by Auer, Cesa-Bianchi and Fischer (2002), where the UCB algorithm was described and analyzed. We first prove that **k1-UCB** (Algorithm 2) with Kullback-Leibler divergence for Bernoulli distributions is always preferable to UCB, in the sense that a smaller finite-time regret bound is guaranteed. UCB is indeed nothing but **k1-UCB** with quadratic divergence and we obtain a refined analysis of UCB as a consequence of Theorem 1. We then discuss the use of the empirical **KL-UCB** approach, in which one directly applies Algorithm 3. We provide preliminary results to support the observation that empirical **KL-UCB** achieves improved performance on sufficiently long horizons (see simulation results in Section 7), at the price however of a significantly higher computational complexity.

6.1. *The **k1-UCB** algorithm for bounded distributions.* A careful reading of the proof of Theorem 1 (see Section A) shows that **k1-UCB** enjoys regret

guarantees in models with arbitrary bounded distributions  $\nu$  over  $[0, 1]$  as long as it is used with a divergence  $d$  over  $[0, 1]^2$  satisfying the following double property: There exists a family of strictly convex and continuously differentiable functions  $\phi_\mu : \mathbb{R} \rightarrow [0, +\infty)$ , indexed by  $\mu \in [0, 1]$ , such that first,  $d(\cdot, \mu)$  is the convex conjugate of  $\phi_\mu$  for all  $\mu \in [0, 1]$ ; and, second, the domination condition  $\mathcal{L}_\nu(\lambda) \leq \phi_{\mathbb{E}(\nu)}(\lambda)$  for all  $\lambda \in \mathbb{R}$  and all  $\nu \in \mathfrak{M}_1([0, 1])$  holds, where  $\mathcal{L}_\nu$  denotes the moment-generating function of  $\nu$ ,

$$\mathcal{L}_\nu : \lambda \in \mathbb{R} \mapsto \mathcal{L}_\nu(\lambda) = \int_{[0,1]} e^{\lambda x} d\nu(x).$$

The following elementary lemma dates back to Hoeffding (1963); it upper bounds the moment-generating function of any probability distribution over  $[0, 1]$  with expectation  $\mu$  by the moment-generating function of the Bernoulli distribution with parameter  $\mu$ , which is further bounded by the moment-generating function of the normal distribution with mean  $\mu$  and variance  $1/4$ . All these moment-generating functions are defined on the whole real line  $\mathbb{R}$ . In light of the above, it thus shows that the Kullback-Leibler divergence  $d_{\text{BER}}$  between Bernoulli distributions and the Kullback-Leibler divergence  $d_{\text{QUAD}}$  between normal distributions with variance  $1/4$  are adequate candidates for use in the kl-UCB algorithm in the case of bounded distributions.

LEMMA 1. *Let  $\nu \in \mathfrak{M}_1([0, 1])$  and let  $\mu = \mathbb{E}(\nu)$ . Then, for all  $\lambda \in \mathbb{R}$ ,*

$$\mathcal{L}_\nu(\lambda) = \int_{[0,1]} e^{\lambda x} d\nu(x) \leq 1 - \mu + \mu \exp(\lambda) \leq \exp(\lambda\mu + 2\lambda^2).$$

The proof of this lemma is straightforward; the first inequality is by convexity, as  $e^{\lambda x} \leq xe^\lambda + (1-x)$  for all  $x \in [0, 1]$ , the second inequality follows by standard analysis.

We therefore have the following corollaries to Theorem 1. (They are obtained by bounding in particular the variance term  $\sigma_{a,\star}^2$  by  $1/4$ .)

COROLLARY 1. *Consider a bandit problem with rewards bounded in  $[0, 1]$ . Choosing the parameters  $f(t) = \log(t) + 3 \log \log(t)$  for  $t \geq 3$  and  $f(1) = f(2) = f(3)$ , and*

$$d_{\text{BER}}(\mu, \mu') = \mu \log \frac{\mu}{\mu'} + (1 - \mu) \log \frac{1 - \mu}{1 - \mu'}$$

*in Algorithm 3, the number of draws of any suboptimal arm  $a$  is upper*

bounded for any horizon  $T \geq 3$  as

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{\log(T)}{d_{\text{BER}}(\mu_a, \mu^*)} + \frac{\sqrt{2\pi} \log\left(\frac{\mu^*(1-\mu_a)}{\mu_a(1-\mu^*)}\right)}{(d_{\text{BER}}(\mu_a, \mu^*))^{3/2}} \sqrt{\log(T) + 3 \log(\log(T))} \\ &\quad + \left(4e + \frac{3}{d_{\text{BER}}(\mu_a, \mu^*)}\right) \log(\log(T)) + \frac{2\left(\log\left(\frac{\mu^*(1-\mu_a)}{\mu_a(1-\mu^*)}\right)\right)^2}{(d_{\text{BER}}(\mu_a, \mu^*))^2} + 6. \end{aligned}$$

We denote by  $\phi_{\mathbb{E}(\nu)} = 1 - \mathbb{E}(\nu) + \mathbb{E}(\nu) \exp(\cdot)$  the upper bound on  $\mathcal{L}_\nu$  exhibited in Lemma 1. Standard results on Kullback-Leibler divergences are that for all  $\mu, \mu' \in [0, 1]$  and all  $\nu, \nu' \in \mathfrak{M}_1([0, 1])$ ,

$$d_{\text{BER}}(\mu, \mu') = \sup_{\lambda \in \mathbb{R}} \{\lambda \mu - \phi_{\mu'}(\lambda)\} \quad \text{and} \quad \text{KL}(\nu, \nu') \geq \sup_{\lambda \in \mathbb{R}} \{\lambda \mathbb{E}(\nu) - \mathcal{L}_{\nu'}(\lambda)\}$$

(see Massart 2007, pages 21 and 28, see also Dembo and Zeitouni 1998). Because of Lemma 1, it thus holds that for all distributions  $\nu, \nu' \in \mathfrak{M}_1([0, 1])$ ,

$$d_{\text{BER}}(\mathbb{E}(\nu), \mathbb{E}(\nu')) \leq \text{KL}(\nu, \nu'),$$

and it follows that in the model  $\mathcal{D} = \mathfrak{M}_1([0, 1])$  one has

$$\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) \geq d_{\text{BER}}(\mu_a, \mu^*).$$

As expected, the **kl-UCB** algorithm may not be optimal for all sub-families of bounded distributions. Yet, this algorithm has stronger guarantees than the **UCB** algorithm. It is readily checked that the latter exactly corresponds to the choice of

$$d_{\text{QUAD}}(\mu, \mu') = 2(\mu - \mu')^2$$

in Algorithm 2. Indeed, the analysis derived in this paper gives an improved analysis of the performance of the **UCB** algorithm.

**COROLLARY 2.** *Consider the **kl-UCB** algorithm with  $d_{\text{QUAD}}$ , or equivalently, the **UCB** algorithm tuned as follows: at step  $t + 1 > K$ , an arm maximizing the upper-confidence bounds*

$$\hat{\mu}_a(t) + \sqrt{(\log(t) + 3 \log \log(t)) / (2N_a(t))}$$

*is chosen. Then the number of draws of a suboptimal arm  $a$  is upper bounded as*

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{\log(T)}{2(\mu^* - \mu_a)^2} + \frac{2\sqrt{\pi}}{(\mu^* - \mu_a)^2} \sqrt{\log(T) + 3 \log(\log(T))} \\ &\quad + \left(4e + \frac{3}{2(\mu^* - \mu_a)^2}\right) \log(\log(T)) + \frac{8}{(\mu^* - \mu_a)^2} + 6. \end{aligned}$$

As claimed, it can be checked that the leading term in the bound of Corollary 1 is smaller than the one of Corollary 2 by applying Pinsker's inequality  $d_{\text{BER}} \leq d_{\text{QUAD}}$ . The bound obtained in Corollary 2 above also improves on the one of Auer, Cesa-Bianchi and Fischer (2002, Theorem 1) and it is "optimal" in the sense that the constant 1/2 in the logarithmic term cannot be improved. Note that a constant in front on the leading term of the regret bound is proven to be arbitrarily close to (but strictly greater than) 1/2 for the UCB2 algorithm of Auer, Cesa-Bianchi and Fischer (2002), when the parameter  $\alpha$  goes to 0 as the horizon grows, but then other terms are unbounded. In comparison, Corollary 2 provides a bound for UCB with a leading optimal constant 1/2 and all the remaining terms of the bound are finite and made explicit. Note, in addition, that the choice of the parameter  $\alpha$ , which drives the length of the phases during which a single arm is played, is important but difficult in practice, where UCB2 does not really prove more efficient than UCB.

6.2. *The empirical KL-UCB algorithm for bounded distributions.* The justification of the use of empirical KL-UCB for general bounded distributions  $\mathfrak{M}_1([0, 1])$  relies on the following result.

*A result of independent interest, connected to the empirical-likelihood method.* The empirical-likelihood (or EL in short) method provides a way to construct confidence bounds for the true expectation of i.i.d. observations; for a thorough introduction to this theory, see Owen (2001). We only recall briefly its principle. Given a sample  $X_1, \dots, X_n$  of an unknown distribution  $\nu_0$ , and denoting  $\hat{\nu}_n = n^{-1} \sum_{k=1}^n \delta_{X_k}$  the empirical distribution of this sample, an EL upper-confidence bound for the expectation  $E(\nu_0)$  of  $\nu_0$  is given by

$$(13) \quad U_{\text{EL}}(\hat{\nu}_n, \varepsilon) = \sup \left\{ E(\nu') : \nu' \in \mathfrak{M}_1(\text{Supp}(\hat{\nu}_n)) \text{ and } \text{KL}(\hat{\nu}_n, \nu') \leq \varepsilon \right\},$$

where  $\varepsilon > 0$  is a parameter controlling the confidence level.

An apparent impediment to the application of this method in bandit problems is the impossibility of obtaining non-asymptotic guarantees for the covering probability of EL upper-confidence bounds. In fact, it appears in (13) that  $U_{\text{EL}}(\hat{\nu}_n, \varepsilon)$  necessarily belongs to the convex envelop of the observations. If, for example, all the observations are equal to 0, then  $U_{\text{EL}}(\hat{\nu}_n, \varepsilon)$  is also equal to 0, no matter what the value of  $\varepsilon$  is; therefore, it is not possible to obtain an upper-confidence bounds for all confidence levels.

In the case of (upper-)bounded variables, this problem can be circumvented by adding to the support of  $\hat{\nu}_n$  the maximal possible value. In our

case, instead of considering  $U_{\text{EL}}(\widehat{\nu}_n, \varepsilon)$ , one should use

$$(14) \quad U(\widehat{\nu}_n, \varepsilon) = \sup \left\{ \mathbb{E}(\nu') : \nu' \in \mathfrak{M}_1(\text{Supp}(\widehat{\nu}_n) \cup \{1\}) \text{ and } \text{KL}(\widehat{\nu}_n, \nu') \leq \varepsilon \right\}.$$

This idea was introduced in Honda and Takemura (2010), independently of the EL literature. The following guarantee can be obtained; its proof is provided in Section C.2.

**PROPOSITION 1.** *Let  $\nu_0 \in \mathfrak{M}_1([0, 1])$  with  $\mathbb{E}(\nu_0) \in (0, 1)$  and let  $X_1, \dots, X_n$  be independent random variables with common distribution  $\nu_0 \in \mathfrak{M}_1([0, 1])$ , not necessarily with finite support. Then, for all  $\varepsilon > 0$ ,*

$$\mathbb{P}\{U(\widehat{\nu}_n, \varepsilon) \leq \mathbb{E}(\nu_0)\} \leq \mathbb{P}\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mathbb{E}(\nu_0)) \geq \varepsilon\} \leq e(n+2) \exp(-n\varepsilon),$$

where  $\mathcal{K}_{\text{inf}}$  is defined in terms of the model  $\mathcal{D} = \mathcal{F}$ .

For  $\{0, 1\}$ -valued observations, it is readily seen that  $U(\widehat{\nu}_n, \varepsilon)$  boils down to the upper-confidence bound given by (12). This example suggests that the above proposition is not (always) optimal: the presence of the factor  $n$  in front of the exponential  $\exp(-n\varepsilon)$  term is indeed questionable.

*Conjectured regret guarantees of empirical KL-UCB.* The analysis of empirical KL-UCB in the case where the arms are associated with general bounded distributions is a work in progress. In view of Proposition 1 and of the discussion above, it is only the proof of Fact 2 that needs to be extended.

As a preliminary results, we can prove an asymptotic regret bound, which is indeed optimal, but for a variant of Algorithm 3; it consists of playing in regimes  $r$  of increasing lengths instances of the empirical KL-UCB algorithm in which the upper confidence bounds are given by

$$\sup \left\{ \mathbb{E}(\nu) : \nu \in \mathfrak{M}_1(\text{Supp}(\widehat{\nu}_a(t)) \cup \{1 + \delta_r\}) \text{ and } \text{KL}(\widehat{\nu}_a(t), \nu) \leq \frac{f(t)}{N_a(t)} \right\},$$

where  $\delta_r \rightarrow 0$  as the index of the regime  $r$  increases.

The open questions would be to get an optimal bound for Algorithm 3 itself, preferably a non-asymptotic one like those of Theorems 1 and 2. Also, a computational issue arises: as the support of each empirical distribution may contain as many points as the number of times the corresponding arm was pulled, the computational complexity of the empirical KL-UCB algorithm grows, approximately linearly, with the number of rounds. Hence the empirical KL-UCB algorithm as it stands is only suitable for small to medium horizons (typically less than ten thousands rounds). To reduce the numerical

complexity of this algorithm without renouncing to performance, a possible direction could be to cluster the rewards on adaptive grids that are to be refined over time.

**7. Numerical experiments.** The results of the previous sections show that the **k1-UCB** and the empirical **KL-UCB** algorithms are efficient not only in the special frameworks for which they were developed, but also for general bounded distributions. In the rest of this section, we support this claim by numerical experiments that compare these methods with competitors such as **UCB** and **UCB-Tuned** (Auer, Cesa-Bianchi and Fischer, 2002), **MOSS** (Audibert and Bubeck, 2010), **UCB-V** (Audibert, Munos and Szepesvári, 2009) or **DMED** (Honda and Takemura, 2011). In these simulations, similar confidence levels are chosen for all the upper confidence bounds, corresponding to  $f(t) = \log(t)$ —a choice which we recommend in practice. Indeed, using  $f(t) = \log(t) + 3 \log \log(t)$  or  $f(t) = (1 + \varepsilon) \log(t)$  (with a small  $\varepsilon > 0$ ) yields similar conclusions regarding the ranking of the performance of the algorithms, but leads to slightly higher average regrets. More precisely, the upper-confidence bounds we used were  $U_a(t) = \hat{\mu}_a(t) + \sqrt{\log(t)/(2N_a(t))}$  for **UCB**,

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\hat{v}_a(t)\log(t)}{N_a(t)}} + 3\frac{\log(t)}{N_a(t)}$$

with

$$(15) \quad \hat{v}_a(t) = \left( \frac{1}{N_a(t)} \sum_{s=1}^t Y_s^2 \mathbb{I}_{\{A_s=a\}} \right) - \hat{\mu}_a(t)^2$$

for **UCB-V**, and, following Auer, Cesa-Bianchi and Fischer (2002),

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\min\left\{1/4, \hat{v}_a(t) + \sqrt{2\log(t)/N_a(t)}\right\} \log(t)}{N_a(t)}}$$

for **UCB-Tuned**. Both **UCB-V** and **UCB-Tuned** are expected to improve over **UCB** by estimating the variance of the rewards; but **UCB-Tuned** was introduced as an heuristic improvement over **UCB** (and does not come with a performance bound) while **UCB-V** was analysed by Audibert, Munos and Szepesvári (2009).

Different choices of the divergence function  $d$  lead to different variants of the **k1-UCB** algorithm, which are sometimes compared with one another in the sequel. In order to clarify this point, we reserve the term **k1-UCB** for the variant using the *binary* Kullback-Leibler divergence (i.e., between

Bernoulli distributions), while other choices are explicitly specified by their denomination (e.g., `kl-poisson-UCB` or `kl-exp-UCB` for families of Poisson or exponential distributions). The simulations presented in this section have been performed using the `py/mABandits` package (Cappé, Garivier and Kaufmann, 2012), which is publicly available from the `mloss.org` website and can be used to replicate these experiments.

7.1. *Bernoulli rewards.* We first consider the case of Bernoulli rewards, which has a special historical importance and which covers several important practical applications of bandit algorithms (see Robbins (1952); Gittins (1979); and references therein). With  $\{0,1\}$ -valued rewards and with the binary Kullback-Leibler divergence as a divergence function, it is readily checked that the `kl-UCB` algorithm coincides exactly with empirical KL-UCB.

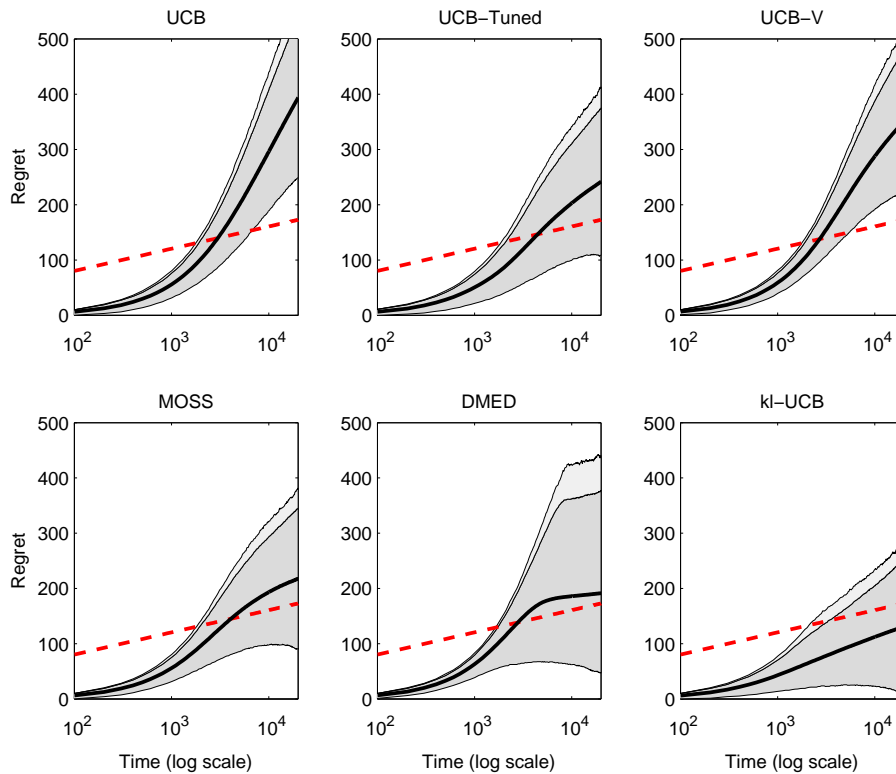


FIG 1. Regret of the various algorithms as a function of time (on a log-scale) in the Bernoulli ten-arm scenario. On each figure, the dashed line shows the asymptotic lower bound; the solid bold curve corresponds to the mean regret; while the dark and light shaded regions show respectively the central 99% region and the upper 99.95% quantile.

In Figure 1 we consider a difficult scenario, inspired by a situation (frequent in applications like marketing or Internet advertising) where the mean reward of each arm is very low. In our scenario, there are ten arms: the optimal arm has expected reward 0.1, and the nine suboptimal arms consist of three different groups of three (stochastically) identical arms, each with respective expected rewards 0.05, 0.02 and 0.01. We resorted to  $N = 50,000$  simulations to obtain the regret plots of Figure 1. These plots show, for each algorithm, the average cumulated regret together with quantiles of the cumulated regret distribution as a function of time (on a logarithmic scale).

Here, there is a huge gap in performance between UCB and  $\text{k1-UCB}$ . This is explained by the fact that the variances of all reward distributions are much smaller than  $1/4$ , the pessimistic upper bound used in Hoeffding's inequality (that is, in the design of UCB). The gain in performance of UCB-Tuned is not very significant.  $\text{k1-UCB}$  and DMED reach a performance that is on par with the lower bound (1) of Burnetas and Katehakis (1996) (shown in strong dashed line); the performance of  $\text{k1-UCB}$  is somewhat better than the one of DMED. Notice that for the best methods, and in particular for  $\text{k1-UCB}$ , the mean regret is below the lower bound, even for larger horizons, which reveals and illustrates the asymptotic nature of this bound.

*7.2. Truncated Poisson rewards.* In this second scenario, we consider 6 arms with truncated Poisson distributions. More precisely, each arm  $1 \leq a \leq 6$  is associated with  $\nu_a$ , a Poisson distribution with expectation  $(2 + a)/4$ , truncated at 10. The experiment consisted of  $N = 10,000$  Monte-Carlo replications on an horizon of  $T = 20,000$  steps. Note that the truncation does not alter much the distributions here, as the probability of draws larger than 10 is small for all arms. In fact, the role of this truncation is only to provide an explicit upper bound on the possible rewards, which is required for most algorithms.

Figure 2 shows that, in this case again, the UCB algorithm is significantly worse than some of its competitors. The UCB-V algorithm, which appears to have a larger regret on the first 5,000 steps, progressively improves thanks to its use of variance estimates for the arms. But the horizon  $T = 20,000$  is (by far) not sufficient for UCB-V to provide an advantage over  $\text{k1-UCB}$ , which is thus seen to offer an interesting alternative even in non-binary cases.

These three methods, however, are outperformed by the  $\text{k1-poisson-UCB}$  algorithm: using the properties of the Poisson distributions (but not taking truncation into account, however), this algorithm achieves a regret that is about ten times smaller. In-between stands the empirical KL-UCB algorithm; it relies on non-parametric empirical-likelihood-based upper bounds and is



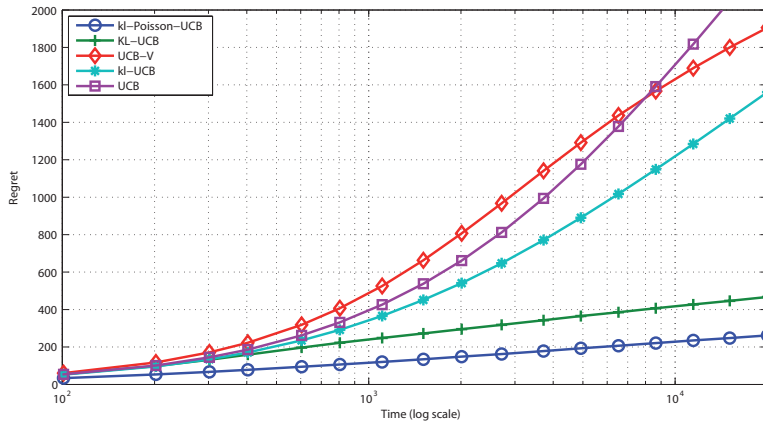


FIG 2. Regret of the various algorithms as a function of time in the truncated Poisson scenario.

therefore is distribution-free as explained in Section 6.2, yet, it proves remarkably efficient.

**7.3. Truncated exponential rewards.** In the third and last example, there are 5 arms associated with continuous distributions: the rewards are exponential variables, with respective parameters  $1/5$ ,  $1/4$ ,  $1/3$ ,  $1/2$  and  $1$ , truncated at  $x_{\max} = 10$  (i.e., they are bounded in  $[0, 10]$ ).

In this scenario, UCB and MOSS are clearly suboptimal. This time, the  $\text{kl-UCB}$  does not provide a significant improvement over UCB as the expectations of the arms are not particularly close to 0 or to  $x_{\max} = 10$ ; hence the confidence intervals computed by  $\text{kl-UCB}$  are close to those used by UCB. UCB-V, by estimating the variances of the distributions of the rewards, which are much smaller than the variances of  $\{0, 10\}$ -valued distributions with the same expectations, would be expected to perform significantly better. But here again, UCB-V is not competitive, at least for an horizon  $T = 20,000$ . This can be explained by the fact that the upper confidence bound of any suboptimal arm  $a$ , as stated in (15), contains a residual term  $3 \log(t)/N_a(t)$ ; this term is negligible in common applications of Bernstein's inequality, but it does not vanish here because  $N_a(t)$  is precisely of order  $\log(t)$  (see also Garivier and Cappé 2011 for further discussion of this issue).

The  $\text{kl-exp-UCB}$  algorithm uses the divergence  $d(x, y) = x/y - 1 - \log(x/y)$  prescribed for genuine exponential distributions, but it ignores the fact that the rewards are truncated. However, contrary to the previous scenario, the truncation has an important effect here, as values larger than 10 are rela-

tively probable for each arm. Because `k1-exp-UCB` is not aware of the truncation, it uses upper bounds that are slightly too large. Yet, the performance is still excellent, stable, and the algorithm is particularly simple.

But the best-performing algorithm in this case is the non-parametric algorithm, empirical `KL-UCB`. This method appears to reach here the best compromise between efficiency and versatility, at the price of a larger computational complexity.

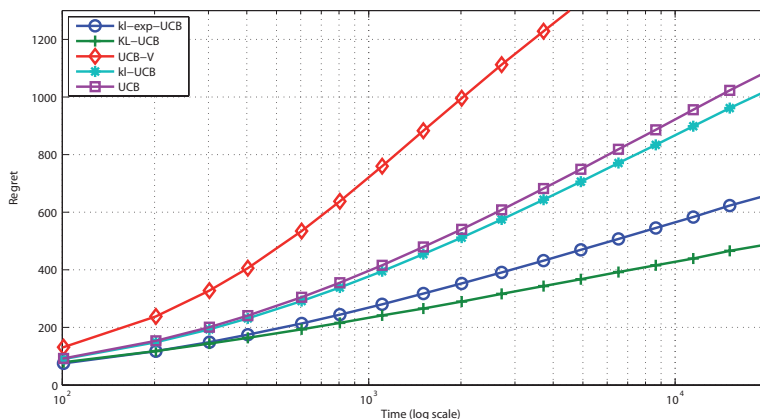


FIG 3. *Regret of the various algorithms as a function of time in the truncated exponential scenario.*

**8. Conclusion.** The `k1-UCB` algorithm is a quasi-optimal method for multi-armed bandits whenever the distributions associated with the arms are known to belong to a simple parametric family. For each one-dimensional exponential family, a specific divergence function has to be used in order to achieve the lower bound (1) of Lai and Robbins (1985).

However, the binary Kullback-Leibler divergence plays a special role: it is a conservative, universal choice for bounded distributions. The resulting algorithm is versatile, fast and simple, and proves to be a significant improvement, both in theory and in practice, over the widely used `UCB` algorithm.

The more elaborate `KL-UCB` algorithm relies on non-parametric inference, by using the so-called empirical likelihood method. It is optimal if the distributions of the arms are only known to be bounded (with a known upper bound) and finitely supported. For general bounded arms, the empirical-likelihood-based upper confidence bounds, which are the core of the algorithm, still have a adequate level; but obtaining explicit finite-time regret

bounds for the algorithm itself and/or reducing its computational complexity is still the object of further investigations (see the discussion in Section 6.2). The simulation results show that empirical KL-UCB is efficient in general cases when the distributions are far from being members of simple parametric families.

In a nutshell, empirical KL-UCB is to be preferred when the distributions of the arms are not known to belong (or be close) to a simple parametric family and when the `kl-UCB` algorithm is known not to get satisfactory performance—that is, for instance, when the variance of a  $[0, 1]$ -valued arm with expectation  $\mu$  is much smaller than  $\mu(1 - \mu)$ .

**Acknowledgements.** Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz acknowledge support from the French National Research Agency (ANR) under grant EXPLO/RA (“Exploration–exploitation for efficient resource allocation”), by the PASCAL2 Network of Excellence under EC grant no. 506778 and from the EC’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 270327.

## APPENDIX A: PROOF OF THEOREM 1

To prove Theorem 1, one needs to check the two “facts to be proven” of Section 3.1. We will do so using the choice  $\mu^\dagger = \mu^*$  for the analysis parameter.

**A.1. Proof of Fact 1.** Our goal is to upper bound the quantity

$$\sum_{t=K}^{T-1} \mathbb{P}\{\mu^* \geq U_{a^*}(t)\}.$$

The control consists of two steps: first, a reduction of this question to the application of a deviation result; second, an instantiation of the resulting bound to our choice of  $f$ . The deviation inequality itself is deferred to section A.1.1 below.

For all  $t \in \{K, \dots, T-1\}$ , we have on the event  $\{U_{a^*}(t) \leq \mu^*\}$  that  $\widehat{\mu}_{a^*}(t) \leq U_{a^*}(t) \leq \mu^* < \mu_+$ . Therefore, for all  $0 < \delta < \mu_+ - U_{a^*}(t)$ ,

$$d(\widehat{\mu}_{a^*}(t), U_{a^*}(t) + \delta) > \frac{f(t)}{N_{a^*}(t)}.$$

Since  $\widehat{\mu}_{a^*}(t) < \mu_+$ , we then have, except when  $\widehat{\mu}_{a^*}(t) = U_{a^*}(t) = \mu_-$ , that  $\widehat{\mu}_{a^*}(t)$  belongs to the open interval  $I = (\mu_-, \mu_+)$ ; thus the discussion after (12) on the continuity of  $d$  shows that, letting  $\delta \rightarrow 0$ ,

$$d(\widehat{\mu}_{a^*}(t), U_{a^*}(t)) \geq \frac{f(t)}{N_{a^*}(t)}.$$

Therefore, since  $d(\widehat{\mu}_{a^*}(t), \cdot)$  is non-decreasing on  $[\widehat{\mu}_{a^*}(t), \mu^*]$ , we get the inclusion

$$\{\mu^* \geq U_{a^*}(t)\} \subseteq \left\{ \mu^* > \widehat{\mu}_{a^*}(t) \quad \text{and} \quad d(\widehat{\mu}_{a^*}(t), \mu^*) \geq \frac{f(t)}{N_{a^*}(t)} \right\};$$

we note that this inclusion is also valid when  $\widehat{\mu}_{a^*}(t) = U_{a^*}(t) = \mu_-$ . Decomposing according to the values of  $N_{a^*}(t)$  yields

$$\{\mu^* \geq U_{a^*}(t)\} \subseteq \bigcup_{n=1}^{t-K+1} \left\{ \mu^* > \widehat{\mu}_{a^*,n} \quad \text{and} \quad d(\widehat{\mu}_{a^*,n}, \mu^*) \geq \frac{f(t)}{n} \right\}.$$

By application of the deviation result (Lemma 2 below), the sum of in-

terest is thus bounded as

$$(16) \quad \sum_{t=K}^{T-1} \mathbb{P}\{\mu^* \geq U_{a^*}(t)\} \leq \sum_{t=K}^{T-1} \min\left\{1, e^{\lceil f(t) \log(t) \rceil} e^{-f(t)}\right\} \\ \leq 1 + \sum_{t=3}^{T-1} e^{\lceil f(t) \log(t) \rceil} e^{-f(t)},$$

where we used the fact that  $K \geq 2$  for the second inequality. We recall that  $f(t) = \log(t) + 3 \log(\log(t))$  for  $t \geq 3$  and these values are indeed such that  $f(t) > 1$ , as needed to apply Lemma 2. As  $\log(u) \leq u - 1$  for all  $u > -1$  and  $\log(t) > 1$  for  $t \geq 3$ , we have

$$f(t) \log(t) \leq \log^2(t) + 3(\log(t) - 1) \log(t) \leq 4 \log^2(t) - 3,$$

thus  $e^{\lceil f(t) \log(t) \rceil} \leq 4e \log^2(t)$ . It follows that

$$(17) \quad \sum_{t=3}^{T-1} e^{\lceil f(t) \log(t) \rceil} e^{-f(t)} \\ \leq 4e \sum_{t=3}^{T-1} \frac{1}{t \log(t)} \leq 4e \left( \frac{1}{3 \log(3)} + \int_3^{T-1} \frac{1}{t \log(t)} dt \right) \\ \leq 4e \left( \frac{1}{3 \log(3)} + \log(\log(T-1)) - \log(\log(3)) \right) \leq 3 + 4e \log(\log(T)),$$

where we resorted to a sum-integral comparison and used the fact that  $\log(\log(\cdot))$  is the primitive function of  $t \mapsto 1/(t \log(t))$ .

A.1.1. *Deviation inequality.* It remains to state and prove Lemma 2, which is actually a consequence of the more general results provided in Appendix C.4.

LEMMA 2. *For all  $\varepsilon > 1$ , provided that  $\mu_- < \mu^* < \mu_+$ ,*

$$\mathbb{P}\left(\bigcup_{k=1}^n \left\{ \mu^* > \hat{\mu}_{a^*,k} \quad \text{and} \quad k d(\hat{\mu}_{a^*,k}, \mu^*) \geq \varepsilon \right\}\right) \leq e^{\lceil \varepsilon \log(n) \rceil} e^{-\varepsilon}.$$

PROOF. We consider, with the notation of Lemma 11, the random variables  $Z_k = X_{a^*,k}$ . We denote by  $\theta^* = \dot{b}^{-1}(\mu^*)$  the parameter in the exponential family corresponding to their common distribution;  $\theta^*$  lies in the open set  $\Theta$ . The random variable  $e^{\lambda X_{a^*,1}}$  is integrable for all  $\lambda \in \mathbb{R}$  such

that  $\theta + \lambda \in \Theta$ ; these  $\lambda$  are in an open interval containing 0 and denoted by  $(\lambda_1, \lambda_2)$ . In addition, by definition of the densities,

$$(18) \quad \mathbb{E}[e^{\lambda X_{a^*,1}}] = \int_{\mathbb{R}} \exp(\lambda x + \theta^* x - b(\theta^*)) \, d\rho(x) = \exp(b(\theta^* + \lambda) - b(\theta^*));$$

that is,  $\phi(\lambda) = \log \mathbb{E}[e^{\lambda X_{a^*,1}}] = b(\theta^* + \lambda) - b(\theta^*)$ . As indicated when introducing the canonical exponential family,  $b$  is strictly convex and (twice) differentiable; therefore, so is  $\phi$ . We only need to show that  $\phi^* = d(\cdot, \mu^*)$  at least on  $[\mu_-, \mu_+]$

Indeed, note that for all  $\mu \in (\mu_-, \mu_+)$ , the function  $\lambda \in (\lambda_1, \lambda_2) \mapsto \lambda\mu - \phi(\lambda)$  is strictly concave and twice differentiable, with derivative equal to  $\mu - \dot{b}(\theta^* + \lambda)$ ; this derivative is null at a unique point  $\lambda_\mu$  given by

$$(19) \quad \theta^* + \lambda_\mu = \dot{b}^{-1}(\mu)$$

and therefore, the concave function of interest is maximized at this point, with value

$$(20) \quad \begin{aligned} \phi^*(\mu) &= \lambda_\mu \mu - \phi(\lambda_\mu) = (\dot{b}^{-1}(\mu) - \theta^*)\mu - (b(\theta^* + \lambda_\mu) - b(\theta^*)) \\ &= (\dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu^*))\mu - b(\dot{b}^{-1}(\mu)) + b(\dot{b}^{-1}(\mu^*)) = d(\mu, \mu^*), \end{aligned}$$

where the final equality follows from (11). For the other values of  $\mu$ , namely,  $\mu = \mu_-$  and  $\mu = \mu_+$ , we argue by continuity, as  $d$  was extended by continuity and since  $\phi^*$  is convex thus continuous on  $\mathbb{R}$ .  $\square$

**A.2. Proof of Fact 2.** Our goal is to upper bound the quantity

$$\sum_{n \geq n_0+1} \mathbb{P}\left\{\widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^*, f(T)/n}\right\}.$$

The control consists of four main steps. Some rewriting of the events of interest is first performed, to get a form that is suitable for an application of a Markov–Chernoff bounding (which is the second step). In the third step the obtained bound is further bounded in an integral form using the intuition of Laplace’s method. This integral bound is finally controlled in an explicit way, using an auxiliary result proved in Section A.2.1.

*Rewriting step.* Note first that in our case

$$(21) \quad \mathcal{K}_{\text{inf}}(\nu_a, \mu^*) = \inf\left\{d(\mu_a, \mu) : \mu > \mu^*\right\} = d(\mu_a, \mu^*),$$

where the second equality is because the mapping  $d(\mu_a, \cdot)$  is strictly convex and continuous over  $I$  and achieves its minimum at  $\mu_a$ , thus is increasing on  $[\mu_a, \mu_+)$ . Similarly, for all  $\gamma > 0$ ,

$$\mathcal{C}_{\mu^*, \gamma} = \left\{ \nu \in \mathfrak{M}_1(\bar{I}) : \exists \mu \in (\mu^*, \mu_+] \text{ with } d(\mathbb{E}(\nu), \mu) \leq \gamma \right\}.$$

Distributions  $\nu \in \mathfrak{M}_1(\bar{I})$  with  $\mathbb{E}(\nu) > \mu^*$  all belong to  $\mathcal{C}_{\mu^*, \gamma}$ ; for distributions  $\nu \in \mathfrak{M}_1(\bar{I})$  with  $\mathbb{E}(\nu) \leq \mu^*$ , it follows from the same arguments as above that  $\nu \in \mathcal{C}_{\mu^*, \gamma}$  if and only if  $\mathbb{E}(\nu) = \mu_+$  or  $d(\mathbb{E}(\nu), \mu^*) < \gamma$ . For  $0 \leq \gamma \leq d(\mu_a, \mu^*)$ , by continuity and strict convexity of  $d(\cdot, \mu^*)$  on  $I$  there exists a unique  $\mu_\gamma^* \in [\mu_a, \mu^*]$  such that

$$d(\mu_\gamma^*, \mu^*) = \gamma;$$

distributions  $\nu$  with  $d(\mathbb{E}(\nu), \mu^*) \leq \gamma$  together with  $\mathbb{E}(\nu) \leq \mu^*$  are then exactly the ones with  $\mathbb{E}(\nu) \geq \mu_\gamma^*$ . All in all, we proved that the sets of interest can be rewritten, for all  $0 \leq \gamma \leq d(\mu_a, \mu^*)$ ,

$$\mathcal{C}_{\mu^*, \gamma} = \left\{ \nu \in \mathfrak{M}_1(\bar{I}) : \mathbb{E}(\nu) \geq \mu_\gamma^* \right\}.$$

*Markov–Chernoff bounding.* Now, for  $n \geq n_0 + 1$ , by definition of  $n_0$ , we have  $f(T)/n < d(\mu_a, \mu^*)$ ; the probabilities of interest hence equal

$$\begin{aligned} \mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{\mu^*, f(T)/n} \right\} &= \mathbb{P}\left\{ \hat{\mu}_{a,n} > \mu_{f(T)/n}^* \right\} \\ &\leq e^{-\lambda \mu_{f(T)/n}^*} \mathbb{E}\left[ e^{\lambda \hat{\mu}_{a,n}} \right] = e^{-\lambda \mu_{f(T)/n}^*} \left( \mathbb{E}\left[ e^{(\lambda/n) X_{a,1}} \right] \right)^n, \end{aligned}$$

where the upper bound holds for all  $\lambda > 0$  such that  $e^{(\lambda/n) X_{a,1}}$  is integrable and comes from a Markov–Chernoff bounding, while the last equality is by independence and identical distribution. Denoting  $\lambda' = \lambda/n$ , we have proved that for all  $\lambda' > 0$  such that  $e^{\lambda' X_{a,1}}$  is integrable,

$$\mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{\mu^*, f(T)/n} \right\} \leq \exp\left( -n(\lambda' \mu_{f(T)/n}^* - \phi_a(\lambda')) \right),$$

where  $\phi_a(\lambda') = \log \mathbb{E}[e^{\lambda' X_{a,1}}]$ . Now, the same argument as in the proof of Lemma 2 shows that  $\lambda' \mapsto \lambda' \mu_{f(T)/n}^* - \phi_a(\lambda')$  is defined on an open interval of  $\mathbb{R}$  containing 0, is maximized at the value  $\lambda' > 0$  such that  $\mu_{f(T)/n}^* = \dot{b}(\theta_a + \lambda')$ , where  $\theta_a$  is the parameter in  $\Theta$  corresponding to  $\nu_a$ , with maximal value equal to  $d(\mu_{f(T)/n}^*, \mu_a)$ . Therefore,

$$\mathbb{P}\left\{ \hat{\nu}_{a,n} \in \mathcal{C}_{\mu^*, f(T)/n} \right\} \leq e^{-n d(\mu_{f(T)/n}^*, \mu_a)} = e^{-f(T) \varphi(n/f(T))},$$

where we introduced the mapping

$$\varphi : x \in \left[ \frac{1}{d(\mu_a, \mu^*)}, +\infty \right] \mapsto x d(\mu_{1/x}^*, \mu_a).$$

*Integral bound (Laplace's method).* Since the two mappings

$$\gamma \in [0, d(\mu_a, \mu^*)] \longmapsto \mu_\gamma^* \in [\mu_a, \mu^*] \longmapsto d(\mu_\gamma^*, \mu_a)$$

are respectively decreasing and increasing, their composition is a decreasing application. Hence, the mapping  $\varphi$  is nonnegative and increasing; as a consequence, the sum of interest can be bounded as

$$\begin{aligned} (22) \quad & \sum_{n \geq n_0+1} \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^*, f(T)/n} \right\} \leq \sum_{n \geq n_0+1} e^{-f(T)\varphi(n/f(T))} \\ & \leq \int_{n_0}^{\infty} e^{-f(T)\varphi(x/f(T))} dx = f(T) \int_{n_0/f(T)}^{\infty} e^{-f(T)\varphi(u)} du \\ & \leq f(T) \int_{1/d(\mu_a, \mu^*)}^{\infty} e^{-f(T)\varphi(u)} du, \end{aligned}$$

where the last equality follows by a change of variable and the last inequality holds because  $n_0/f(T) \geq 1/d(\mu_a, \mu^*)$  by definition of  $n_0$ . An equivalent of the last bound in (22) can be obtained by using the standard Laplace's method; it is of the order of  $\sqrt{f(T)}$ . A non-asymptotic upper bound is now obtained via an explicit lower bound on  $\varphi$ .

*Control of the integral bound.* We first prove that

$$(23) \quad \forall \gamma \in [0, d(\mu_a, \mu^*)], \quad \mu_\gamma^* - \mu_a \geq -\frac{d(\mu_a, \mu^*) - \gamma}{d'(\mu_a, \mu^*)} \geq 0,$$

where we denoted by  $d'(\mu_a, \mu^*)$  the derivative of the mapping  $\mu \mapsto d(\mu, \mu^*)$ ; the latter exists in view of the defining expression (11) of  $d$  and is negative, as  $d$  is decreasing on  $(\mu_-, \mu^*]$ . Indeed,  $d$  is above any tangent line as it is a convex function:

$$\forall \mu \in (\mu_-, \mu_+), \quad d(\mu_a, \mu^*) - d(\mu, \mu^*) \leq d'(\mu_a, \mu^*) (\mu_a - \mu),$$

which in particular entails that

$$\forall \gamma \in [0, d(\mu_a, \mu^*)], \quad d(\mu_a, \mu^*) - \gamma \leq d'(\mu_a, \mu^*) (\mu_a - \mu_\gamma^*),$$

leading to the claimed inequality as we recall that  $d'(\mu_a, \mu^*) < 0$ .

Combining Lemma 3 below with (23), we get that

$$\forall \gamma \in [0, d(\mu_a, \mu^*)], \quad d(\mu_\gamma^*, \mu_a) \geq \frac{1}{2\sigma_{a,\star}^2} \left( \frac{d(\mu_a, \mu^*) - \gamma}{d'(\mu_a, \mu^*)} \right)^2,$$



where we recall the definition  $\sigma_{a,\star}^2 \stackrel{\text{def}}{=} \max\{\text{Var}[\nu_\theta] : \mu_a \leq \mathbb{E}(\nu_\theta) \leq \mu^\star\}$ . Thus,

$$\forall x \in \left[ \frac{1}{d(\mu_a, \mu^\star)}, +\infty \right], \quad \varphi(x) \geq \frac{x}{2\sigma_{a,\star}^2} \left( \frac{d(\mu_a, \mu^\star) - 1/x}{d'(\mu_a, \mu^\star)} \right)^2.$$

Bounding either  $x$  or  $1/x$  in the expression above, the integral on right-hand side of (22) may be further upper bounded as

$$\begin{aligned} & \int_{1/d(\mu_a, \mu^\star)}^{\infty} e^{-f(T)\varphi(u)} \, du \leq \\ & \int_{1/d(\mu_a, \mu^\star)}^{2/d(\mu_a, \mu^\star)} e^{-(d(\mu_a, \mu^\star) - 1/u)^2 D_{a,\star} f(T)} \, du + \underbrace{\int_{2/d(\mu_a, \mu^\star)}^{\infty} e^{-u D'_{a,\star} f(T)} \, du}_{\leq 1/(D'_{a,\star} f(T))}, \end{aligned}$$

where the upper bound on the second integral is by a direct calculation and where the constants equal

$$D_{a,\star} = \frac{1}{2\sigma_{a,\star}^2} \frac{1}{d(\mu_a, \mu^\star)(d'(\mu_a, \mu^\star))^2} \quad \text{and} \quad D'_{a,\star} = \frac{1}{2\sigma_{a,\star}^2} \left( \frac{d(\mu_a, \mu^\star)}{2d'(\mu_a, \mu^\star)} \right)^2.$$

The remaining integral is controlled by performing a change of variable,  $v = d(\mu_a, \mu^\star) - 1/u$ , that is,  $du = (d(\mu_a, \mu^\star) - v)^{-2} dv$ . Thus,

$$\begin{aligned} & \int_{1/d(\mu_a, \mu^\star)}^{2/d(\mu_a, \mu^\star)} e^{-(d(\mu_a, \mu^\star) - 1/u)^2 D_{a,\star} f(T)} \, du \\ &= \int_0^{d(\mu_a, \mu^\star)/2} e^{-v^2 D_{a,\star} f(T)} \frac{1}{(d(\mu_a, \mu^\star) - v)^2} \, dv \\ &\leq \frac{4}{(d(\mu_a, \mu^\star))^2} \int_0^{+\infty} e^{-v^2 D_{a,\star} f(T)} \, dv = \frac{4}{(d(\mu_a, \mu^\star))^2} \frac{\sqrt{\pi}}{2\sqrt{D_{a,\star} f(T)}}. \end{aligned}$$

Putting everything together, we obtain the bound

$$(24) \quad \sum_{n \geq n_0+1} \mathbb{P}\left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\star, f(T)/n} \right\} \leq 2\sqrt{2\pi\sigma_{a,\star}^2} \sqrt{\frac{(d'(\mu_a, \mu^\star))^2}{(d(\mu_a, \mu^\star))^3} \sqrt{f(T)}} + 8\sigma_{a,\star}^2 \left( \frac{d'(\mu_a, \mu^\star)}{d(\mu_a, \mu^\star)} \right)^2.$$

A.2.1. *A variant of Pinsker's inequality.* It only remains to state and prove Lemma 3. Note that in the case of Bernoulli distributions, it corresponds to (a refinement of) Pinsker's inequality.

LEMMA 3. *Let  $\mu_1 < \mu_2$  be two elements of  $I$ . Then*

$$d(\mu_2, \mu_1) \geq \frac{(\mu_2 - \mu_1)^2}{2\sigma^2},$$

where  $\sigma^2 = \max \{ \text{Var}(\nu_\theta) : \text{E}(\nu_\theta) \in [\mu_1, \mu_2] \}$ .

The proof of this inequality shows that, symmetrically, the inequality  $d(\mu_1, \mu_2) \geq (\mu_2 - \mu_1)^2 / (2\sigma^2)$  is also true.

PROOF. Denote

$$\phi(\lambda) = \log \left( \int \exp(\lambda x) d\nu_{\dot{b}^{-1}(\mu_1)}(x) \right),$$

and recall that by (20),  $\phi$  is twice differentiable. For  $\mu \in [\mu_1, \mu_2]$ , Equation (20) states that  $\phi^*(\mu) = d(\mu, \mu_1)$ . According to Crouzeix (1977),  $\phi^*$  is twice differentiable, with second derivative equal to

$$(\phi^*)''(\mu) = \frac{1}{\phi''(\lambda_\mu)},$$

where  $\lambda_\mu = \dot{b}^{-1}(\mu) - \dot{b}^{-1}(\mu_1)$  by Equation (19). From (18) and the general links between  $\ddot{b}$  and variances in exponential families,

$$\phi''(\lambda) = \ddot{b}(\dot{b}^{-1}(\mu_1) + \lambda) = \text{Var}(\nu_{\dot{b}^{-1}(\mu_1) + \lambda}).$$

Hence, by definition of  $\sigma^2$ ,

$$(\phi^*)''(\mu) = \frac{1}{\text{Var}(\nu_{\dot{b}^{-1}(\mu)})} \geq \frac{1}{\sigma^2}.$$

Thus, by Taylor's formula and as  $\phi^*(\mu_1) = (\phi^*)'(\mu_1) = 0$  because  $\phi^* = d(\cdot, \mu_1)$  is minimal at  $\mu_1$ , we obtain

$$\phi^*(\mu_2) = \int_{\mu_1}^{\mu_2} (\mu_2 - \mu) (\phi^*)''(\mu) d\mu \geq \frac{(\mu_2 - \mu_1)^2}{2\sigma^2}.$$

□

**A.3. Proof of Theorem 1.** Combining (10) with the proofs of Fact 1 and of Fact 2—respectively, equations (16)–(17) and (21)—, substituting the expression for  $f$ , and using the fact that  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^\star) = d(\mu_a, \mu^\star)$  as asserted in (24), we get that for  $T \geq 3$ ,

$$\begin{aligned} \mathbb{E}[N_a(T)] \leq & \frac{\log(T)}{d(\mu_a, \mu^\star)} + \frac{2 \log(\log(T))}{d(\mu_a, \mu^\star)} + 1 + 3 + 4e \log(\log(T)) + 2 \\ & + 2\sqrt{2\pi\sigma_{a,\star}^2} \sqrt{\frac{(d'(\mu_a, \mu^\star))^2}{(d(\mu_a, \mu^\star))^3}} \sqrt{\log(T) + 3 \log(\log(T))} \\ & + 8\sigma_{a,\star}^2 \left( \frac{d'(\mu_a, \mu^\star)}{d(\mu_a, \mu^\star)} \right)^2. \end{aligned}$$

Rearranging the terms concludes the proof.

## APPENDIX B: PROOF OF THEOREM 2

To prove Theorem 2, we consider an analysis parameter  $\mu^\dagger = \mu^\star - \varepsilon$  and check the two “facts to be proven” discussed in Section 3.1. We however first require an important result which provides an alternative variational expression for  $\mathcal{K}_{\text{inf}}$ .

**B.1. Variational form of  $\mathcal{K}_{\text{inf}}$ .** A key element is that  $\mathcal{K}_{\text{inf}}$  defined in (2) with  $\mathcal{D} = \mathcal{F}$  may be given the following variational expression; see Borwein and Lewis (1991); Harari-Kermadec (2006) as well as the re-derivation of this result by Honda and Takemura (2011, Theorem 3). The notation  $\mathbb{E}_\nu$  here indicates that the random variable  $X$  has distribution  $\nu$ .

LEMMA 4. *For all  $\nu \in \mathcal{F}$  and all  $\mu \in (0, 1)$ ,*

$$\mathcal{K}_{\text{inf}}(\nu, \mu) = \max_{\lambda \in [0, 1]} \mathbb{E}_\nu[h_{\lambda, \mu}(X)],$$

where  $h_{\lambda, \mu}$  is the mapping

$$h_{\lambda, \mu} : x \in [0, 1] \mapsto \log\left(1 - \lambda \frac{x - \mu}{1 - \mu}\right).$$

The following regularity lemma will be used throughout this section; it corresponds to Honda and Takemura (2011, Lemma 6).

LEMMA 5. *For all  $\nu \in \mathcal{F}$ , all  $\mu \in (0, 1)$ , and all  $0 < \varepsilon < \mu$ ,*

$$\mathcal{K}_{\text{inf}}(\nu, \mu) \leq \mathcal{K}_{\text{inf}}(\nu, \mu - \varepsilon) + \frac{\varepsilon}{1 - \mu};$$

and, under the additional condition that  $E(\nu) < \mu - \varepsilon$ ,

$$\mathcal{K}_{\text{inf}}(\nu, \mu) \geq \mathcal{K}_{\text{inf}}(\nu, \mu - \varepsilon) + \frac{\varepsilon^2}{2}.$$

Lemma 5 in particular implies that for a distribution  $\nu$  with finite support,

$$\mathcal{K}_{\text{inf}}(\nu, \mu) > 0 \quad \text{if and only if} \quad E(\nu) < \mu.$$

Indeed, in view of the original expression of  $\mathcal{K}_{\text{inf}}$  in (2), and by continuity of  $\text{KL}(\nu, \cdot)$  on the set of distributions with same support as  $\nu$ , we have, for all  $\mu \in (0, 1)$ , that  $\mathcal{K}_{\text{inf}}(\nu, \mu) = 0$  as soon as  $E(\nu) \geq \mu$ . Now, the second part of Lemma 5 entails that  $\mathcal{K}_{\text{inf}}(\nu, \mu) > 0$  when  $E(\nu) < \mu$ .

**B.2. Proof of Fact 1.** Our goal is to control the sum

$$\sum_{t=K}^{T-1} \mathbb{P}\{\mu^* - \varepsilon \geq U_{a^*}(t)\}.$$

First notice that the inequality  $\mu^* - \varepsilon \geq U_{a^*}(t)$  means that all  $\nu \in \mathcal{F}$  with  $E(\nu) > \mu^* - \varepsilon$  are such that  $\text{KL}(\widehat{\nu}_{a^*}(t), \nu) > f(t)/N_{a^*}(t)$ ; in particular, it thus implies that

$$\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*}(t), \mu^* - \varepsilon) = \inf\left\{\text{KL}(\widehat{\nu}_{a^*}(t), \nu) : \nu \in \mathcal{F} \text{ and } E(\nu) > \mu^* - \varepsilon\right\} \geq \frac{f(t)}{N_{a^*}(t)}.$$

The sum of interest is thus be bounded as follows,

$$\begin{aligned} \sum_{t=K}^{T-1} \mathbb{P}\{\mu^* - \varepsilon \geq U_{a^*}(t)\} &\leq \sum_{t=K}^{T-1} \mathbb{P}\left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*}(t), \mu^* - \varepsilon) \geq \frac{f(t)}{N_{a^*}(t)}\right\} \\ &\leq \sum_{t=K}^{T-1} \mathbb{P}\left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*}(t), \mu^*) \geq \frac{f(t)}{N_{a^*}(t)} + \frac{\varepsilon^2}{2}\right\}, \end{aligned}$$

where the second inequality is obtained by application of the second part of Lemma 5, which is legitimate as  $\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*}(t), \mu^* - \varepsilon) > 0$  entails  $E(\widehat{\nu}_{a^*}(t)) < \mu^* - \varepsilon$ , as recalled after the statement of Lemma 5.

By decomposing according to the values of  $N_{a^*}(t)$  as in (8), we have the inclusion

$$\left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*}(t), \mu^*) \geq \frac{f(t)}{N_{a^*}(t)} + \frac{\varepsilon^2}{2}\right\} \subseteq \bigcup_{n=1}^{t-K+1} \left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*,n}, \mu^*) \geq \frac{f(t)}{n} + \frac{\varepsilon^2}{2}\right\},$$

and the probability of each event in the union is upper bounded, resorting to Lemma 6 below, by

$$\begin{aligned} \mathbb{P}\left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_{a^*,n}, \mu^*) \geq \frac{f(t)}{n} + \frac{\varepsilon^2}{2}\right\} &\leq e(n+2) \exp\left(-n(\varepsilon^2/2 + f(t)/n)\right) \\ &= e^{-f(t)}(n+2)e^{1-n\varepsilon^2/2}. \end{aligned}$$

The union bound then leads to

$$\begin{aligned} \sum_{t=K}^{T-1} \mathbb{P}\{\mu^* - \varepsilon \geq U_{a^*}(t)\} &\leq \sum_{t=K}^{T-1} \sum_{n=1}^{t-K+1} e^{-f(t)}(n+2)e^{1-n\varepsilon^2/2} \\ &= e\left(\sum_{t=K}^{T-1} e^{-f(t)}\right) \left(\sum_{n=1}^{t-K+1} (n+2)e^{-n\varepsilon^2/2}\right). \end{aligned}$$

It only remains to deal with the term

$$\sum_{n=1}^{t-K+1} (n+2)e^{-n\varepsilon^2/2} \leq 3 + 2 \sum_{n=2}^{\infty} n e^{-n\varepsilon^2/2}.$$

The positive mapping  $n \mapsto n e^{-n\varepsilon^2/2}$  is increasing on  $[0, 2/\varepsilon^2]$  and decreasing on  $[2/\varepsilon^2, +\infty)$ , so that, for  $\varepsilon \leq 1$ , the following series can be bounded by integrals,

$$\begin{aligned} \sum_{n=2}^{\infty} n e^{-n\varepsilon^2/2} &= \sum_{n=2}^{\lceil 2/\varepsilon^2 \rceil - 1} n e^{-n\varepsilon^2/2} + \lceil 2/\varepsilon^2 \rceil e^{-\lceil 2/\varepsilon^2 \rceil \varepsilon^2/2} + \sum_{n=\lceil 2/\varepsilon^2 \rceil + 1}^{\infty} n e^{-n\varepsilon^2/2} \\ &\leq \int_{x=2}^{\lceil 2/\varepsilon^2 \rceil} x e^{-x\varepsilon^2/2} dx + \lceil 2/\varepsilon^2 \rceil e^{-1} + \int_{x=\lceil 2/\varepsilon^2 \rceil}^{\infty} x e^{-x\varepsilon^2/2} dx \\ &\leq e^{-1} \left(1 + \frac{2}{\varepsilon^2}\right) + \int_{s=0}^{\infty} x e^{-x\varepsilon^2/2} dx \\ &= e^{-1} \left(1 + \frac{2}{\varepsilon^2}\right) + \frac{4}{\varepsilon^4} \underbrace{\int_0^{\infty} u e^{-u} du}_{=1} \end{aligned}$$

(where we performed the change of variable  $u = x\varepsilon^2/2$  to obtain the last equality). Putting everything together, we have the following bound that completes the proof of Fact 1:

$$(25) \quad \sum_{t=K}^{T-1} \mathbb{P}\{\mu^* - \varepsilon \geq U_{a^*}(t)\} \leq \left(\sum_{t=K}^{T-1} e^{-f(t)}\right) \left(3e + 2 + \frac{4}{\varepsilon^2} + \frac{8e}{\varepsilon^4}\right).$$

B.2.1. *Deviation inequality.*

LEMMA 6. *Let  $\nu \in \mathcal{F}$  be a distribution with expectation  $\mathbb{E}(\nu) = \mu \in (0, 1)$  and denote by  $Z_1, \dots, Z_n$  a  $n$ -sample of random variables with common distribution  $\nu$ . For all  $\varepsilon > 0$  and  $n \geq 1$ ,*

$$\mathbb{P}\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mu) \geq \varepsilon\} \leq e(n+2)e^{-n\varepsilon}, \quad \text{where } \widehat{\nu}_n = \frac{1}{n} \sum_{k=1}^n \delta_{Z_k}.$$

PROOF. This result is inspired by Honda and Takemura (2012, Theorem 11) and borrows some elements of its proof, in particular the fact that, for all  $\gamma > 0$ , there exists a set  $\Lambda_\gamma \subset [0, 1]$  with cardinality at most  $2 + 1/\gamma$  such that

$$(26) \quad \mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mu) \leq \gamma + \max_{\lambda \in \Lambda_\gamma} \frac{1}{n} \sum_{k=1}^n \log \left( 1 - \lambda \frac{Z_k - \mu}{1 - \mu} \right).$$

Assuming for the moment the existence of such a grid  $\Lambda_\gamma$  (which will be re-proved below), we obtain using the union bound

$$(27) \quad \mathbb{P}\left\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mu) \geq \varepsilon\right\} \leq \sum_{\lambda \in \Lambda_\gamma} \mathbb{P}\left\{\frac{1}{n} \sum_{k=1}^n \log \left( 1 - \lambda \frac{Z_k - \mu}{1 - \mu} \right) \geq \varepsilon - \gamma\right\}.$$

By Markov–Chernoff inequality, for all  $\lambda \in [0, 1]$ ,

$$\begin{aligned} \mathbb{P}\left\{\frac{1}{n} \sum_{k=1}^n \log \left( 1 - \lambda \frac{Z_k - \mu}{1 - \mu} \right) \geq \varepsilon - \gamma\right\} &\leq e^{-n(\varepsilon - \gamma)} \mathbb{E}\left[\prod_{k=1}^n \left( 1 - \lambda \frac{Z_k - \mu}{1 - \mu} \right)\right] \\ &= e^{-n(\varepsilon - \gamma)} \prod_{k=1}^n \mathbb{E}\left[1 - \lambda \frac{Z_k - \mu}{1 - \mu}\right] = e^{-n(\varepsilon - \gamma)}, \end{aligned}$$

using both the independence of the  $Z_k$  and the fact that  $\mathbb{E}(\nu) = \mu$  for the final equalities. The bound (27) and the observation that  $\Lambda_\gamma$  has cardinality at most  $2 + 1/\gamma$  yield

$$\mathbb{P}\{\mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mu) \geq \varepsilon\} \leq \left(2 + \frac{1}{\gamma}\right) e^{-n(\varepsilon - \gamma)}.$$

We conclude by taking  $\gamma = 1/n$ .

It only remains to prove (26). Thanks to Lemma 4, this inequality can be rewritten as

$$\sup_{\lambda \in [0, 1]} \frac{1}{n} \sum_{k=1}^n \log \left( 1 - \lambda \frac{Z_k - \mu}{1 - \mu} \right) \leq \gamma + \max_{\lambda' \in \Lambda_\gamma} \frac{1}{n} \sum_{k=1}^n \log \left( 1 - \lambda' \frac{Z_k - \mu}{1 - \mu} \right),$$

and thus follows from the fact that for all  $\lambda \in [0, 1]$ , there exists  $\lambda' \in \Lambda_\gamma$  such that, for all  $x \in [0, 1]$ ,

$$\log\left(1 - \lambda \frac{x - \mu}{1 - \mu}\right) \leq \gamma + \log\left(1 - \lambda' \frac{x - \mu}{1 - \mu}\right).$$

This fact is a consequence of Lemma 7 below, by choosing the set

$$\begin{aligned} \Lambda_\gamma = \{1/2, 1\} \cup \{1/2 + \gamma, \dots, 1/2 + \lfloor 1/(2\gamma) \rfloor \gamma\} \\ \cup \{1/2 - \gamma, \dots, 1/2 - \lfloor 1/(2\gamma) \rfloor \gamma\}, \end{aligned}$$

which has, at most,  $2 + 1/\gamma$  elements (Lemma 7 applies for  $\lambda \in [0, 1]$  and  $\lambda = 1$  belongs to the grid  $\Lambda_\gamma$ ).  $\square$

LEMMA 7. *For all  $\lambda, \lambda' \in [0, 1]$  such that either  $\lambda \leq \lambda' \leq 1/2$  or  $\lambda \geq \lambda' \geq 1/2$ , for all real numbers  $c \leq 1$ ,*

$$\log(1 - \lambda c) \leq \log(1 - \lambda' c) + 2|\lambda' - \lambda|.$$

PROOF. First note that the quantities  $1 - \lambda c$  and  $1 - \lambda' c$  are indeed positive as they are respectively larger than  $1 - \lambda > 0$  and  $1 - \lambda' > 0$ ; this is where the condition  $\lambda, \lambda' < 1$  plays a role. The claimed inequality is straightforward in the case  $c \in [0, 1]$  and  $\lambda \geq \lambda' \geq 1/2$ , as well as in the case  $c \leq 0$  and  $\lambda \leq \lambda' \leq 1/2$ . In the rest of the proof we consider only the other cases.

The mapping  $\psi_c : \lambda \in [0, 1] \mapsto \log(1 - \lambda c)$  is concave, differentiable, with a non-increasing and continuous derivative  $\psi'_c(\lambda) = -c/(1 - \lambda c)$ , therefore  $\psi'_c(u) \geq \psi'_c(1/2) \geq \psi'_c(v)$  for all  $0 \leq u \leq 1/2$  and  $1/2 \leq v < 1$ . This entails, via Taylor's equality with integral remainder, that

$$\begin{aligned} \log(1 - \lambda c) - \log(1 - \lambda' c) &= \int_{\lambda'}^{\lambda} \psi'_c(x) dx \\ &= \begin{cases} \int_{\lambda'}^{\lambda} \psi'_c(x) dx \leq \psi'_c(1/2) (\lambda - \lambda') & \text{if } \lambda \geq \lambda' \geq 1/2 \text{ and } c \leq 0; \\ \int_{\lambda}^{\lambda'} (-\psi'_c(x)) dx \leq -\psi'_c(1/2) (\lambda' - \lambda) & \text{if } \lambda \leq \lambda' \leq 1/2 \text{ and } c \in [0, 1]. \end{cases} \end{aligned}$$

In the first case, we note that  $-1/c \geq 0$  and thus

$$\psi'_c(1/2) = \frac{-c}{1 - c/2} = \frac{1}{1/2 - 1/c} \leq 2;$$

while in the second case,  $-1/c \leq -1$  and thus

$$\psi'_c(1/2) = \frac{1}{1/2 - 1/c} \geq \frac{1}{1/2 - 1} = -2;$$

in all cases, the bound  $2|\lambda' - \lambda|$  holds.  $\square$

**B.3. Proof of Fact 2.** In this section we upper bound the residual sum  $\sum_{n \geq n_0+1} \mathbb{P}\{\widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^*-\varepsilon, f(T)/n}\}$ , where  $n_0$  has been chosen in (9) as  $n_0 = \lceil f(T)/\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) \rceil$ .

*Main argument.* For all  $\gamma > 0$ , we have the following inclusions

$$\begin{aligned} \mathcal{C}_{\mu^*-\varepsilon, \gamma} &\subseteq \left\{ \nu \in \mathcal{F} : \mathcal{K}_{\text{inf}}(\nu, \mu^* - \varepsilon) \leq \gamma \right\} \\ &\subseteq \left\{ \nu \in \mathcal{F} : \mathcal{K}_{\text{inf}}(\nu, \mu^*) \leq \gamma + \frac{\varepsilon}{1 - \mu^*} \right\}, \end{aligned}$$

where we have used successively (7) and the first statement of Lemma 5 (for  $0 < \varepsilon < \mu^*$ ). Hence,

(28)

$$\sum_{n \geq n_0+1} \mathbb{P}\left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^*-\varepsilon, f(T)/n} \right\} \leq \mathbb{P}\left\{ \mathcal{K}_{\text{inf}}(\widehat{\nu}_{a,n}, \mu^*) \leq \frac{f(T)}{n} + \frac{\varepsilon}{(1 - \mu^*)} \right\}.$$

The right-hand side is the probability that  $\widehat{\nu}_{a,n}$  belongs to the set

$$\left\{ \mathcal{K}_{\text{inf}}(\cdot, \mu^*) \leq \frac{f(T)}{n} + \frac{\varepsilon}{(1 - \mu^*)} \right\} \cap \mathfrak{M}_1(\text{Supp}(\nu_a)).$$

Lemma 8 below asserts that this set is closed and convex and thus that Sanov's inequality (Lemma 10) may be applied to upper bound the right-hand side of (28) as

$$\begin{aligned} \sum_{n \geq n_0+1} \mathbb{P}\left\{ \mathcal{K}_{\text{inf}}(\widehat{\nu}_{a,n}, \mu^*) \leq \frac{f(T)}{n} + \frac{\varepsilon}{(1 - \mu^*)} \right\} \\ \leq \sum_{n \geq n_0+1} \exp\left(-n \kappa_a\left(\frac{f(T)}{n} + \frac{\varepsilon}{(1 - \mu^*)}\right)\right), \end{aligned}$$

where

(29)

$$\kappa_a(\gamma) = \inf\left\{ \text{KL}(\nu, \nu_a) : \nu \in \mathfrak{M}_1(\text{Supp}(\nu_a)) \text{ such that } \mathcal{K}_{\text{inf}}(\nu, \mu^*) \leq \gamma \right\}.$$

Figure 4 displays pictorially the connection between the quantities  $\nu_a$ ,  $\mu^*$ ,  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)$ ,  $\mathcal{C}_{\mu^*, \gamma}$  and  $\kappa_a(\gamma)$ . Defining the index  $n_1 \geq n_0$  such that

$$n_1 = \left\lceil \frac{(1 + \varepsilon_a) f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} \right\rceil,$$



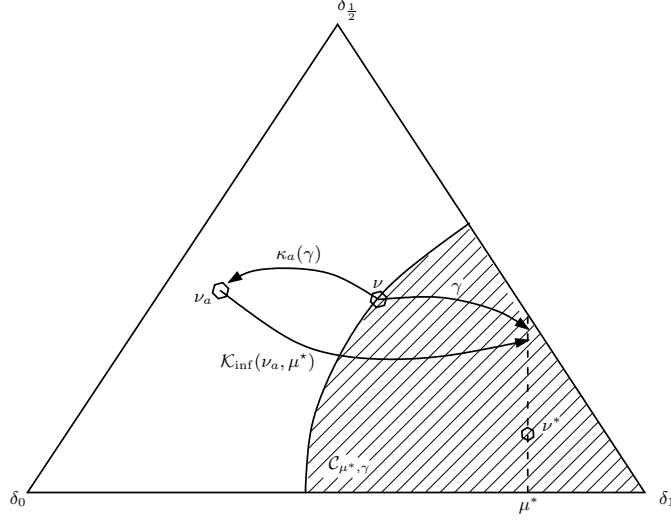


FIG 4. Quantities appearing in the proof represented in the two-dimensional probability simplex  $\mathcal{M}_\infty(\{0, 1/2, 1\})$ ; distributions with the same expectation lie on a common vertical line. Arrows stand for KL-divergence. The distribution  $\nu$  in the figure reaches the minimum in the definition of  $\kappa_a(\gamma)$  given by Equation (29).

where  $\varepsilon_a > 0$  is an analysis parameter, we have

$$\begin{aligned}
 (30) \quad & \sum_{n \geq n_0+1} \exp\left(-n \kappa_a\left(\frac{f(T)}{n} + \frac{\varepsilon}{(1-\mu^*)}\right)\right) \\
 & \leq n_1 - n_0 - 1 + \sum_{n \geq n_1} \exp\left(-n \kappa_a\left(\frac{f(T)}{n} + \frac{\varepsilon}{(1-\mu^*)}\right)\right) \\
 & \leq \frac{\varepsilon_a f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \sum_{n \geq n_1} \exp\left(-n \kappa_a\left(\underbrace{\frac{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)}{1 + \varepsilon_a}}_{\stackrel{\text{def}}{=} \gamma_a(\varepsilon_a, \varepsilon)} + \frac{\varepsilon}{(1-\mu^*)}\right)\right),
 \end{aligned}$$

where we used, for the second inequality, the fact that  $\gamma \mapsto \kappa_a(\gamma)$  is non-increasing and where we introduced for convenience the short-hand notation  $\gamma_a(\varepsilon_a, \varepsilon)$ . We show below that there exists a constant  $M(\nu_a, \mu^*) > 0$  only depending on  $\nu_a$  and  $\mu^*$  such that

$$(31) \quad \forall \varepsilon \leq \frac{1-\mu^*}{2} \varepsilon_a \mathcal{K}_{\text{inf}}(\nu_a, \mu^*), \quad \kappa_a(\gamma_a(\varepsilon_a, \varepsilon)) \geq \frac{\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2}{M(\nu_a, \mu^*)}.$$

Recall that  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) > 0$  as  $\mathbb{E}(\nu_a) < \mu^*$ ; therefore, after substitution

in (30), and under the condition  $\varepsilon \leq (1 - \mu^*) \varepsilon_a \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)/2$ , one obtains

$$\begin{aligned}
(32) \quad & \sum_{n \geq n_0+1} \mathbb{P} \left\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^* - \varepsilon, f(T)/n} \right\} \\
& \leq \frac{\varepsilon_a f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \sum_{n \geq n_1} \exp \left( -s \frac{\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2}{M(\nu_a, \mu^*)} \right) \\
& \leq \frac{\varepsilon_a f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{1}{1 - \exp \left( -\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2 / M(\nu_a, \mu^*) \right)},
\end{aligned}$$

which suffices to prove Fact 2 (in Section B.4 we provide a simpler upper bound of the right-hand-side which coincides with the form given in the statement of Theorem 2).

*Bounding  $\kappa_a$ .* It remains to show that (31) holds. First note that the definition of  $\kappa_a$  in (29) implies that the infimum is achieved: because of the finiteness of the support of  $\nu_a$ , the function  $\text{KL}(\cdot, \nu_a)$  is continuous over the compact set  $\mathfrak{M}_1(\text{Supp}(\nu_a))$ , and its infimum is taken over a closed thus compact set (see Lemma 8). Thus, there exists an element  $\tilde{\nu}$  of  $\mathfrak{M}_1(\text{Supp}(\nu_a))$  such that  $\kappa_a(\gamma_a(\varepsilon_a, \varepsilon)) = \text{KL}(\tilde{\nu}, \nu_a)$  and  $\mathcal{K}_{\text{inf}}(\tilde{\nu}, \mu^*) \leq \gamma_a(\varepsilon_a, \varepsilon)$ . Note that  $\tilde{\nu}$  depends on the two analysis parameters  $\varepsilon$  and  $\varepsilon_a$ . By Pinsker's inequality,

$$\text{KL}(\tilde{\nu}, \nu_a) \geq \frac{\|\tilde{\nu} - \nu_a\|_1^2}{2}, \quad \text{where } \|\tilde{\nu} - \nu_a\|_1 = \sum_{x \in \text{Supp}(\nu_a)} \left| \tilde{\nu}(\{x\}) - \nu_a(\{x\}) \right|.$$

To obtain (31) we show below the existence of a constant  $C(\nu_a, \mu^*) > 0$  only depending on  $\nu_a$  and  $\mu^*$  such that  $\varepsilon_a \mathcal{K}_{\text{inf}}(\nu_a, \mu^*) \leq 2C(\nu_a, \mu^*) \|\tilde{\nu} - \nu_a\|_1$  for all relevant values of  $\varepsilon_a$  and  $\varepsilon$ ; then, (31) holds with  $M(\nu_a, \mu^*)$  equal  $8(C(\nu_a, \mu^*))^2$ .

Because the supremum is achieved in the alternative expression of  $\mathcal{K}_{\text{inf}}$  provided by Lemma 4, there exists  $\lambda_a \in [0, 1]$  be such that

$$\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) = \sum_{x \in \text{Supp}(\nu_a)} \nu_a(\{x\}) \log \left( 1 - \lambda_a \frac{x - \mu^*}{1 - \mu^*} \right).$$

Note that  $\lambda_a$  only depends on  $\nu_a$  and  $\mu^*$ . There are two cases: either  $\nu_a(\{1\}) > 0$  and then necessarily  $\lambda_a < 1$ , or,  $1 \notin \text{Supp}(\nu_a)$ . Using again Lemma 4 to lower bound  $\mathcal{K}_{\text{inf}}(\tilde{\nu}, \mu^*)$ , we have

$$\begin{aligned}
(33) \quad & \mathcal{K}_{\text{inf}}(\nu_a, \mu^*) - \mathcal{K}_{\text{inf}}(\tilde{\nu}, \mu^*) \\
& \leq \sum_{x \in \text{Supp}(\nu_a)} \left( \nu_a(\{x\}) - \tilde{\nu}(\{x\}) \right) \log \left( 1 - \lambda_a \frac{x - \mu^*}{1 - \mu^*} \right).
\end{aligned}$$

Introducing

$$C(\nu_a, \mu^*) = \max_{x \in \text{Supp}(\nu_a)} \left| \log \left( 1 - \lambda_a \frac{x - \mu^*}{1 - \mu^*} \right) \right|$$

and bounding each side of (33) yields  $\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) - \gamma_a(\varepsilon_a, \varepsilon) \leq C(\nu_a, \mu^*) \|\tilde{\nu} - \nu_a\|_1$ . In the two cases mentioned above, either because  $\lambda_a < 1$  or because all  $x$  in the support of  $\nu_a$  are such that  $x < 1$ , the quantity  $C(\nu_a, \mu^*)$  is finite. The proof of (31) is concluded by noting that when  $\varepsilon$  and  $\varepsilon_a$  satisfy the condition in (31), we have

$$\gamma_a(\varepsilon_a, \varepsilon) = \frac{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)}{1 + \varepsilon_a} + \frac{\varepsilon}{(1 - \mu^*)} \leq \frac{1 + \varepsilon_a/2}{1 + \varepsilon_a} \mathcal{K}_{\text{inf}}(\nu_a, \mu^*),$$

and thus,

$$\mathcal{K}_{\text{inf}}(\nu_a, \mu^*) - \gamma_a(\varepsilon_a, \varepsilon) \geq \frac{\varepsilon_a}{2(1 + \varepsilon_a)} \mathcal{K}_{\text{inf}}(\nu_a, \mu^*) \geq \frac{\varepsilon_a}{2} \mathcal{K}_{\text{inf}}(\nu_a, \mu^*).$$

**B.3.1. On the level sets of  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$ .** In the proof of Fact 2, we used the following lemma, which we now prove.

**LEMMA 8.** *For all  $\mu^* \in (0, 1)$ , the function  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$  is convex and continuous over  $\mathfrak{M}_1(\text{Supp}(\nu_a))$ . In particular, the sets  $\{\mathcal{K}_{\text{inf}}(\cdot, \mu^*) \leq \gamma\} \cap \mathfrak{M}_1(\text{Supp}(\nu_a))$  are closed convex subsets of  $\mathfrak{M}_1(\text{Supp}(\nu_a))$ , for all  $\gamma > 0$ .*

**PROOF.** We first show that  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$  is a convex function. Fix  $\alpha \in [0, 1]$ , two distributions  $\nu_1, \nu_2 \in \mathcal{F}$ , and consider two other distributions  $\nu'_1, \nu'_2 \in \mathcal{F}$  with  $\mathbb{E}(\nu'_1) > \mu^*$  and  $\mathbb{E}(\nu'_2) > \mu^*$ .

$$\begin{aligned} \mathcal{K}_{\text{inf}}(\alpha\nu_1 + (1 - \alpha)\nu_2, \mu^*) &\leq \text{KL}(\alpha\nu_1 + (1 - \alpha)\nu_2, \alpha\nu'_1 + (1 - \alpha)\nu'_2) \\ &\leq \alpha \text{KL}(\nu_1, \nu'_1) + (1 - \alpha) \text{KL}(\nu_2, \nu'_2), \end{aligned}$$

where the first inequality is by definition of  $\mathcal{K}_{\text{inf}}$  and the fact that  $\alpha\nu'_1 + (1 - \alpha)\nu'_2$  still has an expectation larger than  $\mu^*$ ; and where the second inequality follows from the joint convexity of the Kullback-Leibler divergence (see, e.g., Cover and Thomas 1991, Theorem 2.7.2). By taking infima over the possible  $\nu'_1$  and  $\nu'_2$ , we get

$$\mathcal{K}_{\text{inf}}(\alpha\nu_1 + (1 - \alpha)\nu_2, \mu^*) \leq \alpha \mathcal{K}_{\text{inf}}(\nu_1, \mu^*) + (1 - \alpha) \mathcal{K}_{\text{inf}}(\nu_2, \mu^*),$$

showing that the mapping  $\nu \in \mathcal{F} \mapsto \mathcal{K}_{\text{inf}}(\nu, \mu^*)$  is convex.

We now turn to the continuity of  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$ . We first show that it is bounded; indeed,  $\mathcal{K}_{\text{inf}}(\nu, \mu^*) = 0$  for all  $\nu \in \mathfrak{M}_1(\text{Supp}(\nu_a))$  with  $\mathbb{E}(\nu) \geq \mu^*$ , while for all  $\nu \in \mathfrak{M}_1(\text{Supp}(\nu_a))$  with  $\mathbb{E}(\nu) < \mu^*$ ,

$$\begin{aligned} \mathcal{K}_{\text{inf}}(\nu, \mu^*) &\leq \text{KL}\left(\nu, \frac{1 - \mu^*}{1 - \mathbb{E}(\nu)} \nu + \frac{\mu^* - \mathbb{E}(\nu)}{1 - \mathbb{E}(\nu)} \delta_1\right) \\ &\leq \sum_{x \in \text{Supp}(\nu)} \nu(\{x\}) \log \frac{\nu(\{x\})}{(1 - \mu^*) / (1 - \mathbb{E}(\nu)) \nu(\{x\})} \\ &\leq \log \frac{1 - \mathbb{E}(\nu)}{1 - \mu^*} \leq \log \frac{1}{1 - \mu^*}, \end{aligned}$$

where the first inequality holds by convexity of the Kullback-Leibler divergence. The function  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$  is therefore a bounded and convex function defined over the simplex  $\mathfrak{M}_1(\text{Supp}(\nu_a))$ ; consequently, it is upper semi-continuous (see Rockafellar 1970, Theorem 10.2). It suffices to show that  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$  is lower semi-continuous over  $\mathfrak{M}_1(\text{Supp}(\nu_a))$ . Using the notation of Lemma 4, for all  $\nu \in \mathfrak{M}_1(\text{Supp}(\nu_a))$ , the mapping

$$\lambda \in [0, 1] \mapsto \mathbb{E}_\nu[h_{\lambda, \mu^*}(X)] = \sum_{x \in \text{Supp}(\nu_a)} \nu(\{x\}) \log \left(1 - \lambda \frac{x - \mu^*}{1 - \mu^*}\right)$$

is continuous (with  $-\infty$  as a possible value at  $\lambda = 1$ ). The result of Lemma 4 can thus be rewritten as indicating that for all  $\nu \in \mathfrak{M}_1(\text{Supp}(\nu_a))$ ,

$$\mathcal{K}_{\text{inf}}(\nu, \mu^*) = \sup_{\lambda \in [0, 1]} \mathbb{E}_\nu[h_{\lambda, \mu^*}(X)].$$

Now, for each  $\lambda \in [0, 1)$ , but not necessarily for  $\lambda = 1$ , the mapping  $\nu \in \mathfrak{M}_1(\text{Supp}(\nu_a)) \mapsto \mathbb{E}_\nu[h_{\lambda, \mu^*}(X)]$  is continuous. The supremum of continuous functions being lower semi-continuous, this concludes the proof of the lower semi-continuity of  $\mathcal{K}_{\text{inf}}(\cdot, \mu^*)$ .  $\square$

**B.4. Proof of Theorem 2.** Combining (10) together with the proofs of Fact 1 and of Fact 2—more precisely, with the upper bounds (25) and (32)—, we obtain that under the conditions  $\varepsilon < \mu^*$  and  $\varepsilon \leq (1 - \mu^*) / (2\varepsilon_a \mathcal{K}_{\text{inf}}(\nu_a, \mu^*))$ , the expected number of times the suboptimal arm  $a$  is pulled satisfies

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + 2 + \left(\sum_{t=K}^{T-1} e^{-f(t)}\right) \left(3e + 2 + \frac{4}{\varepsilon^2} + \frac{8e}{\varepsilon^4}\right) \\ (34) \quad &+ \frac{\varepsilon_a f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{1}{1 - \exp\left(-\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2 / M(\nu_a, \mu^*)\right)}. \end{aligned}$$

We further upper bound the right-hand side of (34) to simplify the form of the bound. First, as  $\varepsilon < \mu^* < 1$ ,

$$3e + 2 + \frac{4}{\varepsilon^2} + \frac{8e}{\varepsilon^4} \leq (3e + 2 + 4 + 8e) \frac{1}{\varepsilon^4} \leq \frac{36}{\varepsilon^4}.$$

Second, because of the choice  $f(t) = \log(t) + \log(\log(t))$  for  $t \geq 2$ ,

$$\begin{aligned} \sum_{t=K}^{T-1} e^{-f(t)} &\leq \sum_{t=2}^{T-1} \frac{1}{t \log(t)} \leq \frac{1}{2 \log(2)} + \int_2^{T-1} \frac{1}{t \log(t)} dt \\ &\leq \frac{1}{2 \log(2)} - \log(\log(2)) + \log(\log(T-1)) \leq 2 + \log(\log(T)), \end{aligned}$$

where we used the same arguments as in (17). Third, we have

$$\frac{1}{1 - \exp\left(-\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2 / M(\nu_a, \mu^*)\right)} \leq 2 + 2 \frac{M(\nu_a, \mu^*)}{\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2};$$

as a consequence of the bound,  $1/(1 - e^{-x}) \leq 2 + 2/x$ , valid for all  $x > 0$ . The latter is obtained by distinguishing two cases: if  $x \geq 1$ , then  $1 - e^{-x} \geq 1 - e^{-1} > 1/2$ ; for  $0 < x \leq 1$ , we have  $e^{-x} \leq 1 - x + x^2/2$  and thus

$$\frac{1}{1 - e^{-x}} \leq \frac{1}{x - x^2/2} = \frac{1}{x(1 - x/2)} \leq \frac{2}{x}.$$

Putting these three upper bounds together, (34) implies that

$$\begin{aligned} (35) \quad \mathbb{E}[N_a(T)] &\leq \frac{f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{36}{\varepsilon^4} \left(2 + \log(\log(T))\right) \\ &\quad + \frac{\varepsilon_a f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + 2 \frac{M(\nu_a, \mu^*)}{\varepsilon_a^2 \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2} + 4. \end{aligned}$$

This bound could be optimized over the admissible values of the analysis parameters  $\varepsilon$  and  $\varepsilon_a$ . For the sake of readability however we only provide convenient values that balance the (orders of magnitude of the) two main terms of the bound, that is, the second and third terms on the right-hand side of (35). These values are  $\varepsilon = \mu^* (\log(T))^{-1/5}$  and  $\varepsilon_a$  such that  $\varepsilon \leq (1 - \mu^*) / (2\varepsilon_a \mathcal{K}_{\text{inf}}(\nu_a, \mu^*))$ ; because  $T \geq 3$ , the condition  $\varepsilon < \mu^*$  is satisfied. Substituting these values, we get

$$\begin{aligned} \mathbb{E}[N_a(T)] &\leq \frac{f(T)}{\mathcal{K}_{\text{inf}}(\nu_a, \mu^*)} + \frac{36}{(\mu^*)^4} \left(2 + \log(\log(T))\right) (\log(T))^{4/5} \\ &\quad + \frac{2\mu^* f(T) (\log(T))^{-1/5}}{(1 - \mu^*) \mathcal{K}_{\text{inf}}(\nu_a, \mu^*)^2} + \frac{(1 - \mu^*)^2 M(\nu_a, \mu^*)}{2(\mu^*)^2} (\log(T))^{2/5} + 4. \end{aligned}$$

Replacing  $f(T)$  with its value  $\log(T) + \log(\log(T))$  and bounding the quantity  $\log(\log(T)) (\log(T))^{-1/5}$  by 1 concludes the proof.

## APPENDIX C: MISCELLANEOUS RESULTS

**C.1. Maximizing the expectation under KL constraint.** In this section, we provide an algorithm to compute

$$(36) \quad U(\nu, \gamma) \stackrel{\text{def}}{=} \sup \left\{ \mathbb{E}(\nu') : \nu' \in \mathcal{F} \text{ and } \text{KL}(\nu, \nu') \leq \gamma \right\},$$

for  $\gamma > 0$  and  $\nu \in \mathcal{F}$ .

*C.1.1. Reduction to a finite-dimensional convex program.* We first show that in (36)  $\mathcal{F}$  can be replaced by  $\mathfrak{M}_1(\text{Supp}(\nu) \cup \{1\})$  without altering the value of  $U(\nu, \gamma)$ . To do so, we prove below the equality

$$\begin{aligned} & \sup \left\{ \mathbb{E}(\nu') : \nu' \in \mathfrak{M}_1([0, 1]) \text{ and } \text{KL}(\nu, \nu') \leq \gamma \right\} \\ &= \sup \left\{ \mathbb{E}(\nu') : \nu' \in \mathfrak{M}_1(\text{Supp}(\nu) \cup \{1\}) \text{ and } \text{KL}(\nu, \nu') \leq \gamma \right\}, \end{aligned}$$

which implies, by a sandwich argument, that both of these quantities are also equal to  $U(\nu, \gamma)$ .

We first establish that  $\sup \left\{ \mathbb{E}(\nu') : \nu' \in \mathfrak{M}_1([0, 1]) \text{ and } \text{KL}(\nu, \nu') \leq \gamma \right\}$  is achieved for some  $\nu_\star \in \mathfrak{M}_1([0, 1])$ . We then prove that  $\nu_\star$  has a support included in  $\text{Supp}(\nu) \cup \{1\}$ .

We equip the set  $\mathfrak{M}_1([0, 1])$  with the vague topology, i.e., the minimal topology such that, for all continuous functions  $f : [0, 1] \rightarrow \mathbb{R}$ , the mappings  $M_f : \nu' \in \mathfrak{M}_1([0, 1]) \mapsto M_f(\nu') = \mathbb{E}_{\nu'}[f(X)]$  are continuous. Prokhorov's theorem indicates that  $\mathfrak{M}_1([0, 1])$  is then a metrizable and compact space. Since the mapping  $\nu' \in \mathfrak{M}_1([0, 1]) \mapsto \text{KL}(\nu, \nu')$  is lower semi-continuous (see, e.g., Chaganty and Karandikar 1996),  $\{\nu' \in \mathfrak{M}_1([0, 1]) : \text{KL}(\nu, \nu') \leq \gamma\}$  is a closed and thus compact subset of  $\mathfrak{M}_1([0, 1])$ . Its image by the continuous mapping  $M_{\text{Id}}$ , where  $\text{Id} : x \in [0, 1] \mapsto x$  is therefore a compact subset of  $[0, 1]$ . Thus, the supremum of  $\left\{ \mathbb{E}(\nu') : \nu' \in \mathfrak{M}_1([0, 1]) \text{ and } \text{KL}(\nu, \nu') \leq \gamma \right\}$  is achieved, at a distribution  $\nu_\star$ .

Consider now the Lebesgue decomposition of  $\nu_\star$ ,

$$\nu_\star = \lambda \nu_\star^{\text{AC}} + (1 - \lambda) \nu_\star^{\text{sing}},$$

where  $\lambda \in [0, 1]$  and where  $\nu_\star^{\text{AC}}$  is a probability measure that is absolutely continuous with respect to  $\nu$ , i.e., that has support included in  $\text{Supp}(\nu)$ , while  $\nu_\star^{\text{sing}}$  is a probability measure that is singular with respect to  $\nu$ , i.e.,

$\nu_\star^{\text{sing}}(\text{Supp}(\nu)) = 0$ . Defining the probability measure  $\tilde{\nu} = \lambda \nu_\star^{\text{AC}} + (1 - \lambda) \delta_1$  and using the short-hand notations  $\nu_x$ ,  $\tilde{\nu}_x$ ,  $\nu_{\star,x}$ , and  $\nu_{\star,x}^{\text{AC}}$  instead of  $\nu(\{x\})$ ,  $\tilde{\nu}(\{x\})$ ,  $\nu_\star(\{x\})$ , and  $\nu_\star^{\text{AC}}(\{x\})$ , respectively, we have

$$\begin{aligned} \text{KL}(\nu, \nu_\star) - \text{KL}(\nu, \tilde{\nu}) &= \sum_{x \in \text{Supp}(\nu)} \nu_x \log \frac{\tilde{\nu}_x}{\nu_{\star,x}} \\ &= \begin{cases} 0 & \text{if } 1 \notin \text{Supp}(\nu); \\ \nu_1 \log \frac{\lambda \nu_{\star,1}^{\text{AC}} + 1 - \lambda}{\lambda \nu_{\star,1}^{\text{AC}}} & \text{if } 1 \in \text{Supp}(\nu). \end{cases} \end{aligned}$$

In all cases,  $\text{KL}(\nu, \tilde{\nu}) \leq \text{KL}(\nu, \nu_\star) \leq \gamma$ . Therefore, in view of the maximality of  $\text{E}(\nu_\star)$  under the latter constraint,

$$0 \leq \text{E}(\nu_\star) - \text{E}(\tilde{\nu}) = (1 - \lambda) \int_{[0,1]} (x - 1) d\nu_\star^{\text{sing}}(x);$$

and thus, either  $\lambda = 1$  (and therefore,  $\nu_\star$  has a support included in the one of  $\nu$ ), or,  $\nu_\star^{\text{sing}} = \delta_1$ , which corresponds to the case where  $\nu_\star$  has support included in  $\text{Supp}(\nu) \cup \{1\}$ .

*C.1.2. Algorithm for computing  $U(\nu, \gamma)$ .* Because of the reformulation  $U(\nu, \gamma) = \sup \{ \text{E}(\nu') : \nu' \in \mathfrak{M}_1(\text{Supp}(\nu) \cup \{1\}) \text{ and } \text{KL}(\nu, \nu') \leq \gamma \}$ , where we recall that  $\nu$  has a finite support,  $U(\nu, \gamma)$  appears as the value of a convex program which we restate under the following simpler form.

Let  $n$  be a positive integer and fix  $n$  elements of  $[0, 1]$ , the larger of them being equal to 1, denoted by  $0 \leq x_1 < \dots < x_{n-1} < x_n = 1$ . Probability measures over this set  $\{x_1, \dots, x_n\}$  are identified with  $n$ -tuples  $(q_1, \dots, q_n)$  such that  $q_i \geq 0$  for all  $i$  and  $q_1 + \dots + q_n = 1$ . A probability distribution  $(p_1, \dots, p_n)$  is given and the optimization problem at hand is to

$$(37) \quad \text{maximize } \sum_{i=1}^n q_i x_i \text{ under the constraints } \begin{cases} \forall i \in \{1, \dots, n\}, q_i \geq 0; \\ q_1 + \dots + q_n = 1; \\ \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \leq \gamma. \end{cases}$$

In Nilim and El Ghaoui (2005), a similar problem arises in a different context, and a somewhat different solution than the one exposed below is proposed for the case when the  $p_i$  are all positive (see also Filippi, Cappé and Garivier 2010). However, note that in the case of the computation of  $U(\nu, \gamma)$ , the identification of  $\text{Supp}(\nu) \cup \{1\}$  to  $\{x_1, \dots, x_n\}$  is such that  $p_i > 0$  for  $1 \leq i \leq n - 1$  (a condition assumed to be satisfied in the rest of this

section); but it can happen that  $p_n = 0$ , this is actually the case if and only if  $1 \notin \text{Supp}(\nu)$ . The optimization problem is trivial if  $n = 2$  and  $p_n = 0$ ; we thus assume in the sequel that either  $n \geq 3$  or  $p_n > 0$ , and in both cases, two components at least of  $(p_1, \dots, p_n)$  are positive. The solution  $(q_1^*, \dots, q_n^*)$  of (37) may be computed numerically by the following algorithm.

---

**Algorithm 4:** Maximization of the expectation under KL constraint.

---

**Parameters:** A set  $\mathcal{X} = \{x_1, \dots, x_n\}$  with  $0 \leq x_1 < \dots < x_n = 1$ ; a probability distribution  $(p_1, \dots, p_n)$  on  $\mathcal{X}$  with  $p_i > 0$  for  $1 \leq i \leq n-1$ ; a level  $\gamma > 0$

**Definitions:** Let  $a = 1$  when  $p_n > 0$  and  $a = x_{n-1}$  when  $p_n = 0$ ; consider the mapping

$$(38) \quad g : (a, +\infty) \mapsto g(\ell) = \sum_{i=1}^n p_i \log(\ell - x_i) + \log \left( \sum_{i=1}^n \frac{p_i}{\ell - x_i} \right).$$

**if**  $p_n = 0$  **and**  $g(1) < \gamma$  **then**

    Let  $r = \exp(g(1) - \varepsilon)$

**for**  $i = 1$  **to**  $n-1$  **do**

        └

$$q_i^* = r \frac{p_i / (1 - x_i)}{\sum_{j=1}^{n-1} p_j / (1 - x_j)}$$

    Let  $q_n^* = 1 - r$

**else**

    Find the root  $\ell$  of the equation  $g(\ell) = \gamma$  **for**  $i = 1$  **to**  $n$  **do**

        └

$$q_i^* = \frac{p_i / (\ell - x_i)}{\sum_{j=1}^n p_j / (\ell - x_j)}$$

Lemma 9 entails that the equation  $g(\ell) = \gamma$  admits a unique solution that can be computed using any numerical method (Newton's search or even simple dichotomy) as  $g$  is a convex decreasing function from  $(a, \infty)$  onto  $(0, \infty)$ , where  $a = 1$  or  $a = x_{n-1}$  as specified in Algorithm 4. In order to upper bound or to provide an approximate value of the root of the equation, one can use the following Taylor series approximation of the function  $g$  (detailed calculations are omitted) as  $\ell \rightarrow +\infty$ :  $g(\ell) = \sigma^2(p)/(2\ell^2) + o(\ell^{-2})$  where  $\sigma^2(p) = \sum_{i=1}^n p_i(1 - p_i)x_i^2$ .

To derive Algorithm 4, observe that (37) is a linear program under convex constraints with Lagrangian

$$(39) \quad \mathcal{L}(q_1, \dots, q_n, \ell, \ell', \ell_1, \dots, \ell_n) = \sum_{i=1}^n q_i x_i - \ell' \left( \sum_{i=1}^n p_i \log \frac{p_i}{q_i} - \gamma \right) - \ell \left( \sum_{i=1}^n q_i - 1 \right) + \sum_{i=1}^n \ell_i q_i;$$



since we proved in Section C.1.1 the existence of a solution, we know that it is characterized by the Karush-Kuhn-Tucker conditions.

More precisely, the argument  $(q_1^*, \dots, q_n^*)$  solving (37) must first satisfy the KL constraint in (37), which implies that  $q_i^* > 0$  as soon as  $p_i > 0$ , that is, at least for all  $i \leq n-1$ . Also, from (39), there exist real numbers  $\ell, \ell', \ell_1, \dots, \ell_n$  with

$$(40) \quad \forall i = 1, \dots, n \text{ s.t. } p_i > 0, \quad 0 = x_i + \ell' \frac{p_i}{q_i^*} + \ell_i - \ell;$$

$$(41) \quad \begin{aligned} & \text{if } p_n = 0, & 0 = x_n + \ell_n - \ell; \\ & & 0 \leq \ell'; \end{aligned}$$

$$\forall i = 1, \dots, n, \quad 0 \leq \ell_i;$$

$$(42) \quad 0 = \ell' \left( \sum_{i=1}^n p_i \log \frac{p_i}{q_i^*} - \gamma \right);$$

$$(43) \quad \forall i = 1, \dots, n, \quad 0 = \ell_i q_i^*.$$

Note that (40) and subsequent equations show that  $\ell \geq 0$  as well. Also, for all  $i$  such that  $q_i > 0$ , that is, at least for  $i \leq n-1$ , we get from (43) that  $\ell_i = 0$ , which, after substitution in (40) and provided that  $\ell' > 0$ , leads to:

$$(44) \quad \text{for all } i \text{ with } p_i > 0 \text{ and } q_i > 0, \quad \ell > x_i \text{ and } q_i^* = \ell' \frac{p_i}{\ell - x_i}.$$

*First case:*  $p_n > 0$ . As (40) is valid in this case for at least two  $i$  with  $\ell_i = 0$ , we necessarily have  $\ell' \neq 0$ . Because (44) is then valid for all  $i \leq n$ , we have  $\ell > x_n = 1$  and by summation  $\ell' \sum_{i=1}^n p_i / (\ell - x_i) = \sum_{i=1}^n q_i^* = 1$  and thus  $\ell' = (\sum_{i=1}^n p_i / (\ell - x_i))^{-1}$ . Hence,

$$q_i^* = \frac{p_i / (\ell - x_i)}{\sum_{j=1}^n p_j / (\ell - x_j)}, \quad \text{for all } i \leq n.$$

The parameter  $\ell$  can be characterized as follows: substituting the above expression into (42) yields

$$\gamma = \sum_{i=1}^n p_i \log \left( (\ell - x_i) \sum_{i=1}^n \frac{p_i}{\ell - x_i} \right) = g(\ell),$$

where  $g$  is defined in (38).

*Second case:  $p_n = 0$  and  $g(1) \geq \gamma$ .* Lemma 9 below shows that  $g$  is in this case a continuous decreasing mapping from  $(x_{n-1}, +\infty)$  onto  $(0, +\infty)$ ; there thus exists a unique  $\ell \in [1, +\infty)$  such that  $g(\ell) = \gamma$ ; in particular,  $\ell > x_i$  for all  $i \leq n-1$ . It can be checked directly that the distribution

$$q_n^* = 0 \quad \text{and} \quad q_i^* = \frac{p_i/(\ell - x_i)}{\sum_{j=1}^{n-1} p_j/(\ell - x_j)} \quad \text{for } i \leq n-1,$$

as well as the Lagrange multipliers  $\ell' = (\sum_{i=1}^{n-1} p_i/(\ell - x_i))^{-1}$ ,  $\ell_i = 0$  for  $i \leq n-1$ , and  $\ell_n = \ell - x_n = \ell - 1$  satisfy the constraints in (37) as well as the conditions (40)–(43).

*Third case:  $p_n = 0$  and  $g(1) < \gamma$ .* First note that (40) and (41), together with the fact that  $\ell_1 = 0$  and  $q_1^* > 0$ , imply that  $\ell = x_n + \ell_n \geq x_n > x_1$  and  $\ell = x_1 + \ell' p_1 / q_1^*$ ; thus it must be that  $\ell' > 0$ . Thus, legitimately applying (44), we have  $\ell > x_{n-1}$  and  $q_i^* = \ell' p_i / (\ell - x_i)$  for  $i \leq n-1$ . Also, as  $g(1) < \gamma$ , Lemma 9 shows in this case that there exists a unique  $\zeta \in (x_{n-1}, 1)$  such that  $g(\zeta) = \gamma$ .

Now, we prove by contradiction that  $q_n^* > 0$ . Indeed, if  $q_n^* = 0$ , then the same calculations as in the first case would show that  $g(\ell) = \gamma$ , thus that  $\ell = \zeta < 1$  as  $g$  is decreasing. The contradiction would be that (41) leads to  $\ell \geq x_n = 1$  as noted above. Therefore, we have  $q_n^* > 0$  and thus  $\ell_n = 0$  by (43) and  $\ell = 1$  by (41). The solution  $q_i^*$  can be rewritten, for  $i \neq n-1$ , as  $q_i^* = \ell' p_i / (1 - x_i)$ , so that, after substitution in (42) and using  $p_n = 0$ , we get

$$\gamma = \sum_{i=1}^{n-1} p_i \log \frac{p_i}{q_i^*} = \sum_{i=1}^{n-1} p_i \log \frac{1 - x_i}{\ell'},$$

or, equivalently,

$$\log(\ell') = -\gamma + \sum_{i=1}^{n-1} p_i \log(1 - x_i) = -\gamma + g(1) - \log \left( \sum_{i=1}^{n-1} \frac{p_i}{1 - x_i} \right).$$

Hence,  $\ell' = \exp(g(1) - \gamma) / (\sum_{j=1}^{n-1} p_j / (1 - x_j))$ , from which we conclude that  $q_n^* = 1 - \sum_{j=1}^{n-1} q_j^* = 1 - \exp(g(1) - \gamma)$ .

**C.1.3. Properties of the function  $g$ .** We prove in this section the following lemma regarding the function  $g$ .

**LEMMA 9.** *Let  $a = 1$  if  $p_n > 0$  and  $a = x_{n-1}$  if  $p_n = 0$ . The function  $g$  defined in (38) is a convex (thus continuous) decreasing mapping from  $(a, \infty)$  onto  $(0, \infty)$ .*

PROOF. We will make repeated uses in this proof of random variables  $Z_\ell$  defined for  $\ell > a$  and taking values  $1/(\ell - x_i)$  each with probability  $p_i$ , for  $i \in \{1, \dots, n\}$ . Note that because of our assumptions on  $(p_1, \dots, p_n)$ , which entail in particular that it has at least two different positive components, the random variables  $Z_\ell$  are not almost-surely constant.

The derivative of  $g$  equals

$$g'(\ell) = \sum_{i=1}^n p_i \frac{1}{\ell - x_i} - \frac{\sum_{i=1}^n p_i \frac{1}{(\ell - x_i)^2}}{\sum_{i=1}^n p_i \frac{1}{\ell - x_i}} = \frac{(\mathbb{E}[Z_\ell])^2 - \mathbb{E}[Z_\ell^2]}{\mathbb{E}[Z_\ell]},$$

and as  $Z_\ell$  is not almost-surely constant, Jensen's inequality shows that  $g'(\ell) > 0$ ; hence  $g$  is decreasing.

We recall that  $p_{n-1} > 0$  so that the probability  $p_a$  put on  $a$  by  $(p_1, \dots, p_n)$  is always in  $(0, 1)$ . Using Taylor expansions it can then be checked that  $g(\ell) \equiv \log(p_a(\ell - a)^{p_a - 1})$  when  $\ell \rightarrow a$  and, hence, that it tends to  $+\infty$  in  $a$ . Likewise,  $g(\ell) = O(1/\ell)$  when  $\ell \rightarrow +\infty$  and thus decreases to zero.

We conclude the proof by showing the convexity of  $g$ ; to do so, we show that its second derivative is nonnegative.

$$\begin{aligned} g''(\ell) &= - \sum_{i=1}^n \frac{p_i}{(\ell - x_i)^2} \\ &\quad + \frac{2 \left( \sum_{i=1}^n \frac{p_i}{(\ell - x_i)^3} \right) \left( \sum_{i=1}^n \frac{p_i}{\ell - x_i} \right) - \left( \sum_{i=1}^n \frac{p_i}{(\ell - x_i)^2} \right)^2}{\left( \sum_{i=1}^n \frac{p_i}{\ell - x_i} \right)^2} \\ &= \frac{-\mathbb{E}[Z_\ell^2] (\mathbb{E}[Z_\ell])^2 + 2 \mathbb{E}[Z_\ell^3] \mathbb{E}[Z_\ell] - (\mathbb{E}[Z_\ell^2])^2}{(\mathbb{E}[Z_\ell])^2}. \end{aligned}$$

The Cauchy-Schwarz inequality ensures that on the one hand,  $(\mathbb{E}[Z_\ell])^2 \leq \mathbb{E}[Z_\ell^2]$  and thus that  $\mathbb{E}[Z_\ell^2] (\mathbb{E}[Z_\ell])^2 \leq (\mathbb{E}[Z_\ell^2])^2$ , and that on the other hand,  $(\mathbb{E}[Z_\ell^2])^2 = (\mathbb{E}[Z_\ell^{3/2} Z_\ell^{1/2}])^2 \leq \mathbb{E}[Z_\ell^3] \mathbb{E}[Z_\ell]$ . These two inequalities show that  $g''(\ell) \geq 0$ , as claimed.  $\square$

**C.2. Proof of Proposition 1.** We merely sketch the proof of Proposition 1, based on the proof of Lemma 6. The same arguments as the one

at the beginning of Section B.2, and in particular, the sandwich equality described in Section C.1.1, show that

$$\{U(\widehat{\nu}_n, \varepsilon) \leq \mathbb{E}(\nu_0)\} \subseteq \left\{ \mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mathbb{E}(\nu_0)) \geq \varepsilon \right\},$$

where  $\mathcal{K}_{\text{inf}}$  is defined as in (2) with the model  $\mathcal{D} = \mathcal{F}$ . So, the question is only whether the result of Lemma 6 holds without the assumption that the underlying distribution at hand is in  $\mathcal{F}$ . The answer is seen to be positive. Indeed, the mentioned proof relies first on a control of  $\mathcal{K}_{\text{inf}}(\widehat{\nu}_n, \mathbb{E}(\nu_0))$ , which is based on Lemma 4; the latter is applied therein to  $\widehat{\nu}_n$ , which has finite support, even if the underlying distribution is not discrete. As for the second part of the proof of Lemma 6, it consists only of the application of a union bound and of a Chernoff bounding: it thus holds true as well.

**C.3. Sanov's inequality.** We consider a sequence  $Z_1, Z_2, \dots$  of real random variables, independent and identically distributed according to a distribution  $\nu$  with finite support  $\mathcal{S}$ . For all integers  $n \geq 1$ , we denote the empirical distribution corresponding to the first  $n$  elements of the sequence by

$$\widehat{\nu}_n = \frac{1}{n} \sum_{k=1}^n \delta_{Z_k}.$$

The following lemma, used in Section B.3 for proving Theorem 2, is a straightforward consequence of Dembo and Zeitouni (1998, Exercise 2.2.38).

LEMMA 10. *Let  $\mathcal{C}$  be a closed and convex subset of  $\mathfrak{M}_1(\mathcal{S})$ . Then, for all  $n \geq 1$ ,*

$$\mathbb{P}\{\widehat{\nu}_n \in \mathcal{C}\} \leq \exp\left(-n \inf_{\kappa \in \mathcal{C}} \text{KL}(\kappa, \nu)\right).$$

**C.4. Deviation inequality.** In this section, we prove the following maximal inequality that is needed in Section A.1 to prove Theorem 1.

LEMMA 11. *Consider a sequence  $Z_1, Z_2, \dots$  of independent and identically distributed real random variables with common expectation  $\mu_0$  and denote by  $\bar{Z}_n = (1/n) \sum_{k=1}^n Z_k$  their empirical means. Assume that there exists an open interval  $(\lambda_1, \lambda_2)$  of  $\mathbb{R}$  containing 0 and a strictly convex, continuously differentiable function  $\phi : (\lambda_1, \lambda_2) \rightarrow \mathbb{R}$  such that*

$$\forall \lambda \in (\lambda_1, \lambda_2), \quad \log \mathbb{E}[e^{\lambda Z_1}] \leq \phi(\lambda).$$

Then for all  $\varepsilon > 1$ ,

$$\mathbb{P}\left(\bigcup_{k=1}^n \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad k \phi^*(\bar{Z}_k) \geq \varepsilon \right\}\right) \leq e[\varepsilon \log(n)] e^{-\varepsilon},$$

where  $\phi^* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  is the convex conjugate of  $\phi$  defined by

$$(45) \quad \forall z \in \mathbb{R}, \quad \phi^*(z) = \sup_{\lambda \in (\lambda_1, \lambda_2)} \{ \lambda z - \phi(\lambda) \}.$$

As explained below,  $\mu_0$ , the expectation of the  $Z_k$ , is the argument of the global minimum of  $\phi^*$ , with  $\phi^*(\mu_0) = 0$ ; deviations of the empirical averages  $\bar{Z}_k$  from the mean  $\mu_0$  are here considered in terms of deviations of  $\phi^*(\bar{Z}_k)$  from 0.

Note that the bound in the lemma holds actually for all  $\varepsilon > 0$  as soon as  $n \geq 3$ , as it is a trivial bound (larger than 1) for  $\varepsilon \leq 1$  and  $n \geq 3$ . Also, symmetric arguments show that under the same assumptions, a similar deviation bound holds true also for deviations to the right:

$$\mathbb{P}\left(\bigcup_{k=1}^n \left\{ \mu_0 < \bar{Z}_k \quad \text{and} \quad k \phi^*(\bar{Z}_k) \geq \varepsilon \right\}\right) \leq e[\varepsilon \log(n)] e^{-\varepsilon}.$$

In this article, however, we only need a control of the deviations to the left.

*Some properties of  $\phi^*$ .* We start by reviewing some useful properties of  $\phi^*$ . First note that  $\phi^*$  is nonnegative (as can be seen by taking  $\lambda = 0$  in its definition) and is strictly convex on  $\{\phi^* < +\infty\}$  (see Rockafellar 1970, Chapter 26). Denoting by  $\mu_0 = \phi'(0)$  the common expectation of the  $Z_k$ , we note that by Jensen's inequality,  $\phi(\lambda) \geq \lambda\mu_0$  and hence that  $\phi^*(\mu_0) = 0$ , showing that  $\mu_0$  is the argument of the global minimum of  $\phi^*$ . In particular, in view of its strict convexity,  $\phi^*$  is non-increasing on  $(-\infty, \mu_0)$ , and even continuous and decreasing on the interval  $(-\infty, \mu_0) \cap \{\phi^* < +\infty\}$ ; symmetric properties hold on  $(\mu_0, +\infty)$ .

We now underline the fact that for all  $z \leq \mu_0$ ,

$$(46) \quad \phi^*(z) = \sup \left\{ \lambda z - \phi(\lambda) : \lambda \in (\lambda_1, 0) \quad \text{and} \quad \lambda z - \phi(\lambda) > 0 \right\}.$$

Indeed, denote by

$$\psi_z : \lambda \in (\lambda_1, \lambda_2) \mapsto \lambda z - \phi(\lambda)$$

the function to maximize. It is strictly concave. If there exists  $\lambda_z \in (\lambda_1, \lambda_2)$  such that  $\psi'_z(\lambda_z) = z - \phi'(\lambda_z) = 0$ , then  $\phi^*(z) = \psi_z(\lambda_z)$ ; since  $\phi'(0) = \mu_0$  and  $\phi'$  is increasing, we get from  $\phi'(\lambda_z) = z < \mu_0$  that  $\lambda_z < 0$ . This proves (46)

in this case. The remaining case is when  $\psi'_z(\lambda) \neq 0$  for all  $\lambda \in (\lambda_1, \lambda_2)$ . By continuity and since  $\psi'_z(0) = z - \mu_0 < 0$ , this means that  $\psi'_z < 0$  on  $(\lambda_1, \lambda_2)$ , that is,  $\psi_z$  is decreasing on  $(\lambda_1, \lambda_2)$ . The defining supremum of  $\phi^*(z)$  thus corresponds to the limit, as  $\lambda \rightarrow \lambda_1$ , of  $\psi_z(\lambda)$ . Since  $\psi_z(0) = 0$ , all the values  $\lambda \in (\lambda_1, 0)$  are such that  $\psi_z(\lambda) > 0$  while values  $\lambda \in [0, \lambda_2)$  are such that  $\psi_z(\lambda) \leq 0$ ; this proves (46) in this case as well.

PROOF OF LEMMA 11. We start with a peeling argument and to that end, introduce  $n_0 = 0$  and  $n_m = \lfloor \gamma^m \rfloor$ , for some  $\gamma > 1$  that will be chosen at the end of the proof. We denote by  $M = \lceil (\log n)/(\log \gamma) \rceil$  an upper bound on the number of elements in the peeling. Then,  $n_M \geq n$  and

$$\begin{aligned} & \mathbb{P} \left( \bigcup_{k=1}^n \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad k \phi^*(\bar{Z}_k) \geq \varepsilon \right\} \right) \\ & \leq \sum_{m=1}^M \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad \phi^*(\bar{Z}_k) \geq \varepsilon/n_m \right\} \right). \end{aligned}$$

Let  $m \in \{1, \dots, M\}$ . Since  $\phi^*$  is decreasing and continuous on the interval  $(-\infty, \mu_0) \cap \{\phi^* < +\infty\}$ , either  $\phi^* < \varepsilon/n_m$  on this interval and the  $m$ -th probability in the sum above is null; or there exists a unique  $z_m \in (-\infty, \mu_0)$  such that  $\phi^*(z_m) = \varepsilon/n_m$ . In this case, using again  $\phi^*$  is non-increasing on  $(-\infty, \mu_0)$ , we have, for all  $\lambda < 0$

$$\begin{aligned} & \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad \phi^*(\bar{Z}_k) \geq \varepsilon/n_m \right\} \right) = \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \bar{Z}_k \leq z_m \right\} \right) \\ & = \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \exp(\lambda(Z_1 + \dots + Z_k) - k \lambda z_m) \geq 1 \right\} \right) \\ & = \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \exp(\lambda(Z_1 + \dots + Z_k) - k \phi(\lambda)) \geq e^{k(\lambda z_m - \phi(\lambda))} \right\} \right) \\ & \leq \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \exp(\lambda(Z_1 + \dots + Z_k) - k \phi(\lambda)) \geq e^{(n_{m-1}+1)(\lambda z_m - \phi(\lambda))} \right\} \right), \end{aligned}$$

where the last inequality was obtained under the additional assumption that the considered  $\lambda < 0$  is such that  $\lambda z_m - \phi(\lambda) > 0$ .

The fact that the  $Z_k$  are independent and identically distributed together with the definition of  $\phi$ , imply that for all  $\lambda \in (\lambda_1, \lambda_2)$ , the process  $(W_{\lambda,k})$

defined by  $W_{\lambda,0} = 1$  and  $W_{\lambda,k} = \exp(\lambda(Z_1 + \dots + Z_k) - k\phi(\lambda))$  for  $k \geq 1$  is a positive supermartingale. As a consequence, Doob's maximal inequality entails that

$$\begin{aligned} & \mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \exp \left( \lambda \sum_{i=1}^k Z_i - k\phi(\lambda) \right) \geq e^{(n_{m-1}+1)(\lambda z_m - \phi(\lambda))} \right\} \right) \\ & \leq \mathbb{P} \left\{ \max_{k \geq n_{m-1}+1} \left\{ \exp \left( \lambda \sum_{i=1}^k Z_i - k\phi(\lambda) \right) \right\} \geq e^{(n_{m-1}+1)(\lambda z_m - \phi(\lambda))} \right\} \\ & \leq e^{-(n_{m-1}+1)(\lambda z_m - \phi(\lambda))}. \end{aligned}$$

The above bound being valid for all  $\lambda > 0$  with  $\lambda z_m - \phi(\lambda) > 0$ , we finally get from (46) that

$$\mathbb{P} \left( \bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad \phi^*(\bar{Z}_k) \geq \varepsilon/n_m \right\} \right) \leq e^{-(n_{m-1}+1)\phi^*(z_m)}.$$

In view of the definition of  $z_m$ , the right-hand side may be bounded by  $\exp(-(n_{m-1}+1)\phi^*(z_m)) = \exp(-\varepsilon(n_{m-1}+1)/n_m) \leq \exp(-\varepsilon/\gamma)$ , where we used the fact that by definition,  $n_{m-1}+1 \geq \gamma^{m-1}$  and  $n_m \leq \gamma^m$  (for all  $m \geq 1$ ). Putting everything together, we have proved that, in all cases,

$$\mathbb{P} \left( \bigcup_{k=1}^n \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad k\phi^*(\bar{Z}_k) \geq \varepsilon \right\} \right) \leq M e^{-\varepsilon/\gamma} \leq \left\lceil \frac{\log n}{\log \gamma} \right\rceil e^{-\varepsilon/\gamma}.$$

Choosing  $\gamma = \varepsilon/(\varepsilon - 1)$ , which is legitimate for  $\varepsilon > 1$ , and applying the inequality  $\log(1+x) \geq x/(1+x)$  to  $x = 1/(\varepsilon - 1) > -1$ , one obtains

$$\begin{aligned} \mathbb{P} \left( \bigcup_{k=1}^n \left\{ \mu_0 > \bar{Z}_k \quad \text{and} \quad k\phi^*(\bar{Z}_k) \geq \varepsilon \right\} \right) & \leq \left\lceil \frac{\log(n)}{\log(\varepsilon/(\varepsilon - 1))} \right\rceil e^{-\varepsilon+1} \\ & \leq \lceil \varepsilon \log(n) \rceil e^{-\varepsilon+1}. \end{aligned}$$

□

## REFERENCES

- AUDIBERT, J.-Y. and BUBECK, S. (2010). Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research* **11** 2785–2836.
- AUDIBERT, J.-Y., MUNOS, R. and SZEPESVÁRI, C. (2009). Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science* **410** 1876–1902.

- AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning* **47** 235–256.
- BORWEIN, J. M. and LEWIS, A. S. (1991). Duality relationships for entropy-like minimization problem. *SIAM Journal on Computation and Optimization* **29(2)** 325–338.
- BURNETAS, A. N. and KATEHAKIS, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* **17(2)** 122–142.
- BURNETAS, A. N. and KATEHAKIS, M. N. (1997). Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research* 222–255.
- BURNETAS, A. N. and KATEHAKIS, M. N. (2003). Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Sciences* 53–82.
- CAPPÉ, O., GARIVIER, A. and KAUFMANN, E. (2012). py/maBandits: Matlab and Python packages for multi-armed bandits. <http://mloss.org/software/view/415/>.
- CHAGANTY, N. R. and KARANDIKAR, R. L. (1996). Some properties of the Kullback-Leibler number. *Sankhya: The Indian Journal of Statistics* **58(1)** 69–80.
- CHANG, F. and LAI, T. L. (1987). Optimal stopping and dynamic allocation. *Advances in Applied Probability* **19** 829–853.
- CHOW, Y. and TEICHER, H. (1988). *Probability Theory*. Springer.
- COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. John Wiley.
- CROUZEIX, J. P. (1977). A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming* **13** 364–365.
- DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, second ed. Springer.
- FILIPPI, S., CAPPÉ, O. and GARIVIER, A. (2010). Optimism in reinforcement learning and Kullback-Leibler divergence. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing*.
- GARIVIER, A. and CAPPÉ, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*.
- GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* **41** 148–177.
- GITTINS, J., GLAZEBROOK, K. and WEBER, R. (2011). *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.
- HARARI-KERMADEC, H. (2006). Vraisemblance empirique généralisée et estimation semi-paramétrique PhD thesis, Université Paris-Ouest.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13–30.
- HONDA, J. and TAKEMURA, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the 23rd Annual Conference on Learning Theory*.
- HONDA, J. and TAKEMURA, A. (2011). An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning* **85** 361–391.
- HONDA, J. and TAKEMURA, A. (2012). Finite-time regret bound of a bandit algorithm for the semi-bounded support model. arXiv:1202.2277.
- LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22.
- LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*. Springer.
- MAILLARD, O.-A., MUNOS, R. and STOLTZ, G. (2011). A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 23rd Annual Conference on Learning Theory*.
- MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in*



- Mathematics* **1896**. Springer, Berlin. Lectures of the 33rd Summer School on Probability Theory, Saint-Flour, France, July 6–23, 2003.
- NILIM, A. and EL GHAOU, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* **53** 780–798.
- OWEN, A. B. (2001). *Empirical Likelihood. Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.
- ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society* **58** 527–535.
- ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton University Press.
- 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.
- THOMPSON, W. R. (1935). On the theory of apportionment. *American Journal of Mathematics* **57** 450–456.
- VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundation and Trends in Machine Learning* **1** 1–305.
- WALD, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16** 117–186.
- WEBER, R. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability* **2** 1024–1033.
- WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* **42** 143–149.
- LTCI, TELECOM PARISTECH, CNRS      IMT, UNIVERSITÉ PAUL SABATIER  
 46 RUE BARRAULT, 75013 PARIS, FRANCE      118 ROUTE DE NARBONNE, 31000 TOULOUSE, FRANCE  
 E-MAIL: cappe@telecom-paristech.fr      E-MAIL: aurelien.garivier@math.univ-toulouse.fr
- UNIVERSITY OF LEOBEN      INRIA LILLE, SEQUEL PROJECT  
 8700 LEOBEN, AUSTRIA      40 AVENUE HALLEY, 59650 VILLENEUVE D’ASCQ, FRANCE  
 E-MAIL: odalricambrym.maillard@gmail.com      E-MAIL: remi.munos@inria.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  
 E-MAIL: gilles.stoltz@ens.fr